Image Annotation Within the Context of Personal Photo Collections Using Hierarchical Event and Scene Models

Liangliang Cao, Jiebo Luo, *Fellow, IEEE*, Henry Kautz, and Thomas S. Huang, *Life Fellow, IEEE*

***Abstract*—Most image annotation systems consider a single photo at a time and label photos individually. In this work, we focus on collections of personal photos and exploit the contextual information naturally implied by the associated GPS and time metadata. First, we employ a constrained clustering method to partition a photo collection into event-based subcollections, considering that the GPS records may be partly missing (a practical issue). We then use conditional random field (CRF) models to exploit the correlation between photos based on 1) time-location constraints and 2) the relationship between collection-level annotation (i.e., events) and image-level annotation (i.e., scenes). With the introduction of such a multilevel annotation hierarchy, our system addresses the problem of annotating consumer photo collections that requires a more hierarchical description of the customers' activities than do the simpler image annotation tasks. The efficacy of the proposed system is validated by extensive evaluation using a sizable geotagged personal photo collection database, which consists of over 100 photo collections and is manually labeled for 12 events and 12 scenes to create ground truth.**

***Index Terms*—Consumer photo collections, CRF, GPS, scene and event annotation.**

I. Introduction

In recent years, the flourishing of digital photos has presented a grand challenge to both computer vision and multimedia research communities: can a computer system produce satisfactory annotations automatically for personal photos? Furthermore, annotation of personal photos requires a higher level of descriptive annotation of people's activities. This is beyond the scope and capability of classic image retrieval systems [38], [30], [43], [37]: most of these systems were designed for simpler data such as the popular Corel image database and typically only provide simple image semantics (such as sunset, mountain, and lake), while the photos taken by personal cameras are much more complex and involve different people and various activities (such as beach time, birthday parties, wedding, and graduation).

To answer the question "what happened in the photo collection," we adopt the concept of *events* to describe the high level semantics applied to the entire collection. Although of high value to consumers, it is difficult to detect general events from a single image, due to the limitation in the content cues observable from a single image and the ambiguity in inferring high level semantics. In this scenario, an event label is selected to annotate a group of photos that form the event. In addition to the event labels, we are also interested in the environment where a photo was taken, e.g., was it indoors, in the city, or on the beach? Such information will be useful for organizing personal photos, and helpful for searching similarly themed photos from different users. To this end, we employ scene labels for each photo, which will not only make our annotation more descriptive but also help the customers organize and search photos more efficiently.

Although the goal of precise and detailed annotating is aggressive, we argue that it is possible to obtain descriptive annotation of events and scenes for consumer photo collections. One distinct characteristic of personal photos is that they are organized, or more accurately, stored in separate folders, in which the photos may be related to one another in some way. Another characteristic of consumer photos lies in the meta data recorded in the digital photo files. Such meta data includes the date and time when the photo is taken, and sometimes even the GPS location of the photo, all of which can be very useful to model the relationships between photos. While these two characteristics are largely neglected and unexploited in previous research, this paper will build a novel model which makes use of these characteristics together with the visual features and is able to effectively annotate the entire collection of related images instead of isolated images.

II. Related Work

Image retrieval has attracted much research interest since the late 1990s, with the goal to search in large databases for images similar to the query. In a retrieval process, the system assigns a score to every image in the database which indicates the similarity to the query image [38], [30], [37]. Such a similarity score facilitates the tasks of ranking and searching, but is limited for describing the content of images. In recent years, there has been a paradigm shift from query by visual similarity to semantic similarity, which would request more specific concept detection and annotation. A few recent studies involved image

Manuscript received April 14, 2008; revised October 13, 2008. Current version published January 16, 2009. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Mohan S. Kankanhalli.

L. Cao and T. S. Huang are with the Beckman Institute and Coordinated Science Lab, Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA (e-mail: cao4@ifp.uiuc.edu; huang@ifp.uiuc.edu).

J. Luo and H. Kautz are with the Kodak Research Laboratories Eastman, Kodak Company, Rochester, NY 14615 USA (e-mail: jiebo.luo@kodak.com; henry.kautz@kodak.com).

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

Digital Object Identifier 10.1109/TMM.2008.2009693

1520-9210/$25.00 © 2009 IEEE

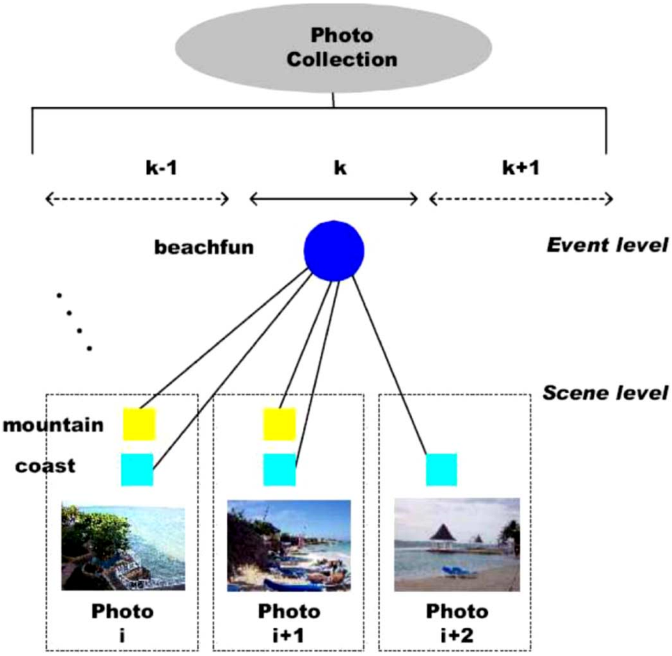


Fig. 1. Hierarchical annotation of photo collections.

annotation with multiple labels [5], [14], but nevertheless were limited to annotating individual photos as opposed to photo collections. As a result, these annotation systems typically focus on concepts that are related to objects and scenes, and rarely address what event an image corresponds to.

To understand better the image semantics, it will be beneficial to consider a collection of images instead of isolated images. Sivic *et al.*[34] showed that it is possible to recognize objects from image collections with the same categories of subjects. Simon *et al.* [33] tried to find the representative image from the collection of photos taken in the location. However, the image collections used in above work are "selected" versions that are organized by predefined themes and topics, and furthermore do not contain photos that may not belong to any predefined classes. We are more interested in understanding the "natural" personal collections which are related to natural events and subsequently loosely organized by the consumers into file folders. Annotation of such natural personal photo collections did not receive specific attention until recently. Wu *et al.* employed web images to learn concept templates in order to query personal photo collections [40] and they also proposed an active learning method for relevance feedback [41]. Note that Cooper *et al.* [9] and Loui *et al.* [23] also considered consumer photo collections, but they did not investigate the annotation problem as in this paper.

To provide a detailed description for photo collections, this paper tries to estimate both the scene and event categories. Scene recognition for single images has been studied in [10], [29], [19], which is part of the interests of this study. Although the images in previous work differ in some degree with the consumer photos (e.g., the database in [10], [29], [19] contains no people in the images), these techniques can be considered as the baseline of annotation for single photos upon which we can build our system. In contrast, event recognition has not received as much attention as scene classification because it clearly concerns higher level semantics, e.g., wedding and birthday, for which low-level visual features alone are found to be inadequate [22]. In previous work, event classification is limited to video analysis [12], [22], [45] or specific sports activities [3], [20]. However, with a collection of photos, it becomes possible to explore the semantic correlation among multiple photos.

This paper integrates both visual content and surrounding context, i.e., time and location for the annotation task, which has proven to be an effective way to bridge the semantic gap in multimedia understanding [24]. The correlation among photos of the same event is a form of useful context. Our approach to photo collection annotation employs both visual features and metadata including both time and GPS. As a new technology, GPS is mostly used in vehicle navigation and seldom used for photo classification [1]. Some researchers proposed to use continuous GPS traces to classify certain reoccurring human activities [21], [44]. In contrast, the GPS records associated with photo collections are discrete, sparse and sometimes missing. We combine such sparse GPS information together with visual features for photo annotation. Note that Naaman and colleagues have done extensive work related to GPS information [27], [26], [11], [16]. However, such work differs from ours in two aspects. First, we employ both time and GPS information as contextual data, and combine that with visual classifiers to recognize different concepts. In contrast, Naaman's work mainly used GPS and user tags. Note that our algorithm is designed to handle partially missing GPS information, which is a common problem for current GPS devices. Naaman's work did not consider that. Moreover, this paper and Naaman's work aim to solve different problems. Our work tries to annotate photos that do not yet have tags, while Naaman *et al.*'s work focused on image summarization and management by making use of existing tags.

III. Outline of Our Approach

Fig. 1 illustrates the annotation task fulfilled by this work. To provide a descriptive annotation for personal photos, we introduce two-level annotation for photo collections. In the upper level, we cluster photos into groups, and assign an event label to each group to denote the main activity common to all the photos in that group. The event label can either be one and only one of the predefined event classes, or "NULL," which indicates nothing of specific interest or categorizable. In the lower level, we assign each photo one or more scene class labels (a multi-label problem as opposed to a multiclass problem as with the event classes). In this paper, we will use the two-level model for the photo annotation task, which we believe will provide more specific descriptions of the photo collections.

Then the research question becomes: given a collection of personal photos, how can we generate more reliable annotations compared with using individual photos? Personal photos are taken in different places and at different times, describing different activities of different people. Indeed, these diverse factors make photo annotation a challenging task, especially for the exiting systems that rely only on visual information from individual images. Therefore, it is imperative to explore different sources of information associated with photo collections.

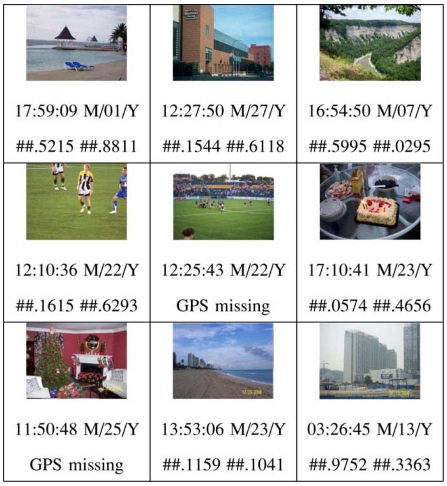


Fig. 2. Example of GPS-tagged consumer photos taken by different photographers at different time and locations. Below each photo, the first row shows the date and time when the photo was taken, and the second row shows the GPS tag. Note the month, year, and coordinates of GPS tag are removed to preserve privacy.

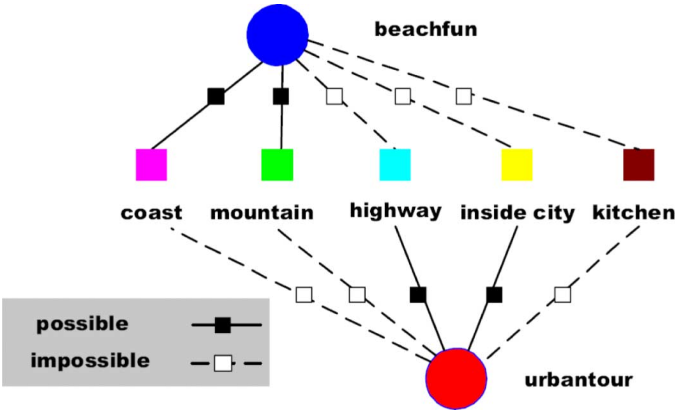


Fig. 3. Correlation between scene labels and event labels.

We first explore the correlation between scene labels. We estimate this type of correlation from camera metadata, a useful but often untapped source of information. Specifically, metadata includes timestamp and GPS tags. Every digital photo file records the date and time when the photo was taken (for example, JPEG file stores tags in the file header). An advanced camera can even record the location via a GPS receiver. However, due to the sensitivity limitation of the GPS receiver, GPS tags can be missing (especially for indoor photos). This paper will discuss how to make good use of such incomplete metadata information. Fig. 2 shows an example of using GPS and time tags to estimate the correlation between photos. The closer the GPS coordinates and the shorter the time intervals are, the stronger the correlation exists between the neighboring photos in their annotation labels.

Second and more importantly, we also consider the relations between scene labels and event labels. Fig. 3 shows examples of both possible (solid lines) and impossible (dashed lines) connection between scenes and events. The event "urbantour" can be linked to "highway" and "inside-city" scene labels, while it is unlikely to co-occur with "coast" or "kitchen." Our system will discover such relationships from the data, and demonstrate that combining such relationships should improve the annotation accuracy.

We build a unified model to account for the two types of correlation as illustrated in Figs. 2 and 3. This model first employs visual features to estimate the probability of isolate events and scenes, and then use metadata features to enforce the correlation between images and also labels at event and scene levels. Our work is developed on the basis of the discriminative model of Conditional Random Field (CRF) in [18], but is different because we introduce a hierarchical structure in order to infer semantics at both scene- and event-levels in a photo collection.

This paper is organized as follows. Section IV describes our dataset and the manual labeling needed for our experiments. Section V presents the basic model for scene annotation, which takes time and GPS information into account. Section VI considers partitioning photos into event clusters. Section VII takes into account the relationship between events and scenes and builds our complete annotation model. Experimental results are in Section VIII and we conclude this paper in Section IX.

IV. Dataset

We built a diverse geotagged photo dataset by camera handouts to different users. Each user took photos as usual and returned the camera with their photo collection. We received 103 photo collections of varying sizes (from four to 249 photos). These collections include extensive photo content. Some examples of the dataset are shown in Fig. 2.

Each photo has a time tag, and over half of the images have GPS tags. Both the time duration and the location range vary across different collections. The time duration can be less than one hour, or several days. Similarly, the GPS movement can be as far as several hundred miles (e.g., road trips) or have negligible change (e.g., one's backyard).

The dataset is labeled by a combination of the photographers (whenever possible) and researchers. We are interested in both indoor and outdoor activities and social events, which are categorized into 12 events. Note that the 12 events include a null category for "none of the above," which means our method can also handle the collections that are not of high interest. This is an important feature for a practical system. Consequently, each photo can be categorized into one and only one of these events. To make the labeling process consistent, we clarify the definitions of the event labels in Table I.

We also labeled each image with the scene labels using the class definitions from [31]: coast, open-country, forest, mountain, inside-city, suburb, highway, livingroom, bedroom, office, and kitchen. Here inside-city includes the original inside-city, plus street and tall-building, since our annotation task does not need to distinguish these three. Again, we also add a null scene class to handle the unspecified cases. Note that a photo can belong to more than one scene class, e.g., a beach photo may also contain mountain, leading to a multilabel problem [6].

V. Scene-Level Modeling

To model the correlation between the labels, we employ a conditional random field (CRF) model. CRF is a probabilistic model first presented by Lafferty *et al.*for the task of segmenting and labeling of sequential data [18]. Sutton *et al.* presented a

TABLE I
DEFINITIONS OF THE 12 EVENTS

Event name	*Detailed definition*
BeachFun	Containing people playing on the beach.
Ballgames	Containing players and the playing field, with or without balls. The field can be baseball, soccer, or football.
Skiing	Containing both snow and skier;on a slope as opposed to a backyard. Not at night.
Graduation	At least one subject in academic cap or gown.
Wedding	Bride must be present. Better with groom.
BirthdayParty	There should be cake or balloon or birthday hat. Can be indoor or outdoor.
Christmas	Christmas decoration, e.g., Christmas tree.
UrbanTour	Large portion of the photo should be buildings, (tall or many) and pavement. Not much green.
YardPark	Containing either grass or trees. May see short building. No sports field nor pavement. It should not be close-up of plants/grass/flowers.
FamilyTime	In the family or living room, with more than 2 people. Sofa or rug must appear, with some furniture.
Dining	Containing a table and dishes, with more than 2 people.
Null Event	None of above.

dynamic CRF to interpret natural language with long range dependencies [35]. Kumar and Hebert proposed discriminative random fields to model the spatial dependencies between image regions [17]. Although the problems studied in [18], [35], [17] are different from this work, their work suggested that CRF provides a powerful tool to model the correlation in sequential data.

Different from generative models such as the Hidden Markov Model, CRF models the conditional likelihoods instead of joint distributions, relaxing the assumption on distributions. Moreover, the feature function in CRF is more flexible than that in HMM, which makes it easier to take more features and factors into account. Let us first address how to model the correlation between scene labels, using time and GPS tags, and we will generalize the model for event annotation in Section VII.

When a photographer takes pictures, the surrounding scene is fairly stable even though he may look in different directions and at different objects. The less the time and location change, the less likely the scene labels of pictures can change from one to another. For example, if one took a photo of a "coast" scene at one time, it is unlikely that the next picture taken within five minutes would be "inside-city." For this reason, there are correlations between the scene labels of the photos that are taken within a short time interval and close location range.

Before introducing our event-scene model, we first define a number of notations. In a photo collection, the photos are represented by $\mathbf{x} = \{\mathbf{x}_i\}, i = 1, 2, \ldots, N$. The time tags and GPS tags are denoted as $\mathbf{t} = \{t_i\}$ and $\mathbf{p} = \{p_i\}$, where $p_i = \mathrm{NULL}$ when the GPS is missing.

We use s_i^k to denote labeling status of the ith photo for scene class k, with $1 \leq k \leq 11$. Here $s_i^k = 1$ means the scene label is true for $\mathbf{x}_i$, while $s_i^k = 0$ means that the scene label is null. Note that if $s_i^k = 0$ for all $1 \leq k \leq 11$, it means that $\mathbf{x}_i$ is not labeled as any of the known scene labels.

Given the time and GPS, We model the correlation using the conditional probability of the kth scene as

$$P(s^k|\mathbf{x}, \mathbf{t}, \mathbf{p}) = \frac{1}{Z_s} \exp \left(\sum_{i=1}^{N} \boldsymbol{\beta}^k \cdot \mathbf{f}_s^k \left(\mathbf{x}_i, s_i^k\right) + \sum_{i=1}^{N-1} \boldsymbol{\lambda}^k \cdot \mathbf{r}_s^k(t_i, t_{i+1}, p_i, p_{i+1}) \right)$$

where "$\cdot$" denotes the inner product between two vectors.

The log-likelihood function for scene k is given by

$$L_s^k = -\log Z_s + \sum_{i=1}^{N} \boldsymbol{\beta}^k \cdot \mathbf{f}_s^k \left(\mathbf{x}_i, s_i^k\right) + \sum_{i=1}^{N-1} \boldsymbol{\lambda}^k \cdot \mathbf{r}_s^k(t_i, t_{i+1}, p_i, p_{i+1}) \quad (1)$$

where $\mathbf{f}_s^k$ stands for the feature function of individual photos for class k, and $\mathbf{r}_s^k(t_i, t_{i+1}, p_i, p_{i+1})$ models the correlation between consecutive photos in the collection i and $i+1$. Z_s stands for a normalization constant. $\boldsymbol{\lambda}^k$ and $\boldsymbol{\beta}^k$ are the parameter vectors that are learned from the training data. L_s^k acts as the objective function in both the training and testing stages. For training, we learn the parameters $(\boldsymbol{\lambda}^k, \boldsymbol{\beta}^k)$ which maximizes L_s^k. For testing, given a photo collection $\{\mathbf{x}_i\}, \{p_i\}, \{t_i\}$, the labeling $\{s_i^k\}$ that maximize (1) will infer the most possible labels.

To obtain the feature function $\mathbf{f}_s^k$ for single photos, we employ the statistical features from [30] and [25]. An SVM classifier is trained for the public scene dataset [19]. The feature function is

$$\mathbf{f}_s^k(\mathbf{x}_i, s_i^k) = \begin{cases} (1, h^k(\mathbf{x}_i))^T, & \text{if } s_i^k > 0 \\ (-1, 1 - h^k(\mathbf{x}_i))^T, & \text{if } s_i^k = 0 \end{cases} \quad (2)$$

where $h^k(\mathbf{x}_i)$ is a sigmoid function used to shape the SVM score, $0 \leq h^k(\mathbf{x}_i) \leq 1$. Since there is no need to re-scale the sigmoid function, we can simply take $\boldsymbol{\beta}^k = (\hat{\beta}^k, 1)^T$, so

$$\boldsymbol{\beta}^k \cdot \mathbf{f}_s^k = \begin{cases} h^k(\mathbf{x}_i) + \hat{\beta}^k, & \text{if } s_i^k > 0 \\ 1 - h^k(\mathbf{x}_i) - \hat{\beta}^k, & \text{if } s_i^k = 0 \end{cases} \quad (3)$$

Given that the larger the differences in time and location are, the less correlation exists between consecutive labels. Moreover, when the consecutive labels are different, the correlation function should contribute little to the overall log-likelihood. With these observations, we define the correlation feature function as

$$\begin{aligned} \mathbf{r}_s^k(i,j) &= \mathbf{r}_s^k(t_i, t_j, p_i, p_j) \\ &= \begin{cases} \left(\frac{1}{1+\exp(dt_{ij})}, \frac{1}{1+\exp(dp_{ij})}\right)^T, & \text{if } s_i^k = s_j^k \\ \mathbf{0}^T, & \text{otherwise} \end{cases} \end{aligned} \quad (4)$$

where dt and dp denote changes in time and location, respectively. In this study, the time change is quantized in intervals of a quarter hour, while the location change is measured in units of a minute of arc on the earth sphere.

Note that the correlation function defined in (4) is able to handle the situation of partially missing GPS. If p_i or p_j is NULL, we treat $dp_{ij} = \infty$ and thus $(1)/(1 + \exp(dp_{ij})) = 0$, which means that the GPS tags impose no correlations on the overall log-likelihood function.

Although (1) considers the correlation in both time and GPS, it is not yet complete since no event labels are involved in this model; neither is the correlation between scenes and events. In what follows, we will add event annotation into the framework, and improve (1) to obtain the complete annotation model.

VI. Event-Level Modeling

In the setting of this paper, our event annotation involves two tasks: grouping photo into event clusters, and classifying each cluster into different event categories. Since it is not reliable to infer the event from single images without considering contextual information, we assign an event label to each cluster instead of single photos. In other words, all photos in the same cluster share the same event label. In this section, we first utilize the time and GPS tags to perform clustering, then validate the clustering accuracy using several criteria, and at the end we present the feature function for event classification.

A. Event Clustering by Time and Position

Our clustering algorithm is based on both time and GPS features. We ignore visual features because the users tend to change their subjects of interests when taking photos. Consequently, the visual features often vary dramatically even at the same event. The time tag is a useful feature for event clustering [23], and a small time interval may also suggest that the photos were taken in the same place. However, it cannot reliably tell whether people stayed in the same place for a long time or they already moved to another location. We next propose a reliable joint clustering method that makes good use of both time and GPS information and is also tolerant to missing GPS data.

Our clustering algorithm works as follows: first, we find baseline clusters from time only using the Mean-Shift algorithm [7]. Mean-Shift does not require us to specify the number of clusters. Since every photo contains a time tag, the baseline clusters can always be obtained from the entire collection. Next, for those samples with both time and GPS tags, we compute the target clustering with the GPS information added. We iteratively search from the baseline clusters for a sample that is not in but close to a sample already in. We add this sample to the same cluster containing its closest neighbors. This iteration will be performed until all the photos are added to. This algorithm is very fast and can obtain real time clustering results even for folders with more than one hundred of images. The details of the clustering algorithm are described as follows.

Procedure of our clustering algorithm

```
Input: Collection of photos. Each photo has a time
stamp, but only some of photos have GPS stamps.

Procedure:
  1: Obtain baseline clusters (subcollections) $C^t$ by
     clustering all the photos using time;
  2: Initialize the target clusters $C$ by clustering only the
     photos with both time and GPS information;
  3: Check whether there are new clusters $C_k^t \subseteq C^t$, such
     that $C_k^t \cap C = \Phi$. Add $C_k^t$ into $C$ as new clusters;
  4: Repeat the following until $C$ contains all the photos:
        4.1. select one example $\mathbf{x}_i$ such that
        $\mathbf{x}_i \in C^t, \mathbf{x}_i \notin C$. Then find the
        corresponding example $\mathbf{x}_i^c \in C$ with
        $\mathbf{x}_i^c = \arg\min_{\mathbf{x} \in C} \mathrm{dist}_T(\mathbf{x}, \mathbf{x}^i)$. Here $\mathrm{dist}_T$ is the
        Euclidean distance between the time tags.
        4.2. add $\mathbf{x}_i$ into $C$ with the cluster label the same
        as $\mathbf{x}_i^c$.

Output: $C$ as the final photo subcollections.
```

B. Clustering Evaluation

We evaluate our clustering algorithm on a small portion of our dataset for which the photographers kindly provide ground truth of event clusters. Since it is impractical to ask all the users to mark all the clusters, we only evaluated our algorithm on 17 photo collections (1394 photos in total).

There are many metrics for measuring the clustering accuracy. In this paper, we utilize two popular ones together with a new one that fits our requirements. The first criterion is Probabilistic Rand Index (PRI) [28], which counts the fraction of pairs of samples whose labels are consistent between the computed cluster and the ground truth, normalized by averaging across all the clusters in the ground truth. The second one is Local Consistency Error (LCE) [36], which is defined as the sum of the number of samples that belong to one cluster C_1 but not C_2, divided by the size of C_1. Here C_1 and C_2 denote the cluster from the ground truth and clustering method, respectively.

PRI and LCE use local or global normalization factors, respectively. However, in this study, we have different preferences on different types of errors: over-partition carries lower cost than under-partition because it is more difficult to assign the correct event label when two different events are inadvertently merged. Neither PRI nor LCE accounts for cost. Therefore, we propose a new metric called Partitioning Error Cost (PEC).

Suppose the computed clustering is $\{c_1, c_2, \ldots, c_n\}$ and the ground truth is $\{g_1, g_2, \ldots, g_m\}$. For each cluster c_i, we compute its contribution to the overall error:

$$\mathrm{err}_i = \begin{cases} 0, & \mathrm{if} \exists c_i = g_j \\ |c_i| w_1 N_Q, & \mathrm{elseif}\ \exists Q s.t. c_i = \bigcup_{j \in Q} g_j, \\ |c_i| w_2 N_P, & \mathrm{elseif}\ \exists P s.t. g_j = \bigcup_{j \in P} c_j, \\ |c_i|, & \mathrm{otherwise} \end{cases},$$

where $|c_i|$ is the number of samples in $c_i, Q \subseteq \{1, 2, \ldots, m\}, P \subseteq \{1, 2, \ldots, n\}$, and N_Q and N_P are the number of g_j and c_i in the union, respectively. And w_1 and w_2 are empirically set as 0.1 and 0.2, respectively, which penalizes under-partition more than over partition. Finally, we sum up the error cost and normalize it by the total number of samples:

$$\mathrm{err} = \frac{1}{N} \sum_i \mathrm{err}_i.$$

Our clustering algorithm is evaluated by these three metrics. Since there is no algorithm that can handle the missing GPS data, we compare our algorithm with the date-time clustering algorithm [23] and temporal similarity-based photo clustering algorithm [9], which are the state of art clustering algorithms using time only. To make a fair comparison, we choose the unsupervised algorithm in [9] instead of the supervised ones. Note that [27] present another clustering algorism based on GPS only. However, this algorithm is not applicable for our scenarios since GPS tags of many photos are missed. To make a more informative comparison, we also compare the simple algorithm that applies Mean-Shift to time only. Table II summarizes the evaluation results. It is clear that our method obtain the lowest error by all three metrics. Fig. 4 shows the clustering errors measured by PEC for all 17 photo collections. Our clustering algorithm outperforms the other two methods for virtually every folder.

Fig. 5 shows example results of our clustering algorithms on four photo collections. The clustering results by using time and GPS (soild red lines) are superior to those using time only (dashed blue lines), when compared to the ground truth of

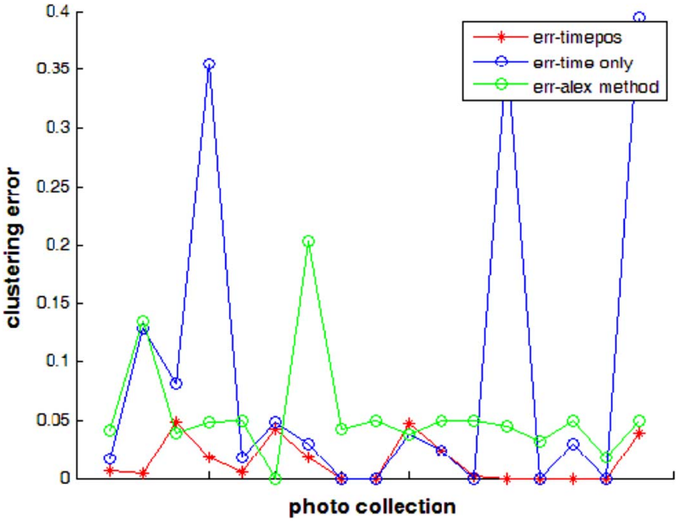


Fig. 4. Comparison of different clustering algorithms. The horizontal axis shows different image folders, and the vertical axis denotes the clustering errors measured by PEC.

TABLE II
EVALUATION OF THE ACCURACY OF CLUSTERING ALGORITHMS

Measures	Our Method	Time-only	Method in [23]	Method in [9]
PRI [7]	0.030420	0.057404	0.097914	0.188685
LCE [28]	0.000660	0.007702	0.001209	0.019309
PEC	0.015055	0.089888	0.055476	0.049860

events. More accurate clustering lay more reliable foundation for the subsequent event recognition. In general, using time only often leads to under-segmentation of clusters and would cause adverse chain reaction under the event-scene model because all the photos in each event share the same event class label.

C. Event Annotation Based on Computed Clusters

After obtaining the event clusters, we impose a feature function on each cluster. Following the standard practice in video concept detection [4], [2], we developed an SVM classifier for the 12 event classes (using software courtesy of University of Central Florida). We separately collected 200 photos for each class, and randomly select 70% of these images for training the multiclass SVM classifier [13]. The remaining 30% of the photos are used for validation.

Given an event subcollection C, our feature function for event e is

$$\mathbf{f}_g^e(C) = \left(1, \sum_{\mathbf{x}_i \in C} \frac{1}{1+\exp(-g^e(\mathbf{x}_i))}\right)^T \tag{5}$$

where $g^e(\mathbf{x}_i)$ is the SVM score of photo $\mathbf{x}_i$ for event e. For our annotation work, $1 \le e \le 11$ stand for the 11 classes of events, and for the null event $g^0 = 0$.

VII. Joint Annotation of Events and Scenes

Some researchers in the field of video annotation observed that it is beneficial to explore the relationship between different labels [8], [42]. These results suggest us to consider the label dependency between event and scene labels for images. As shown

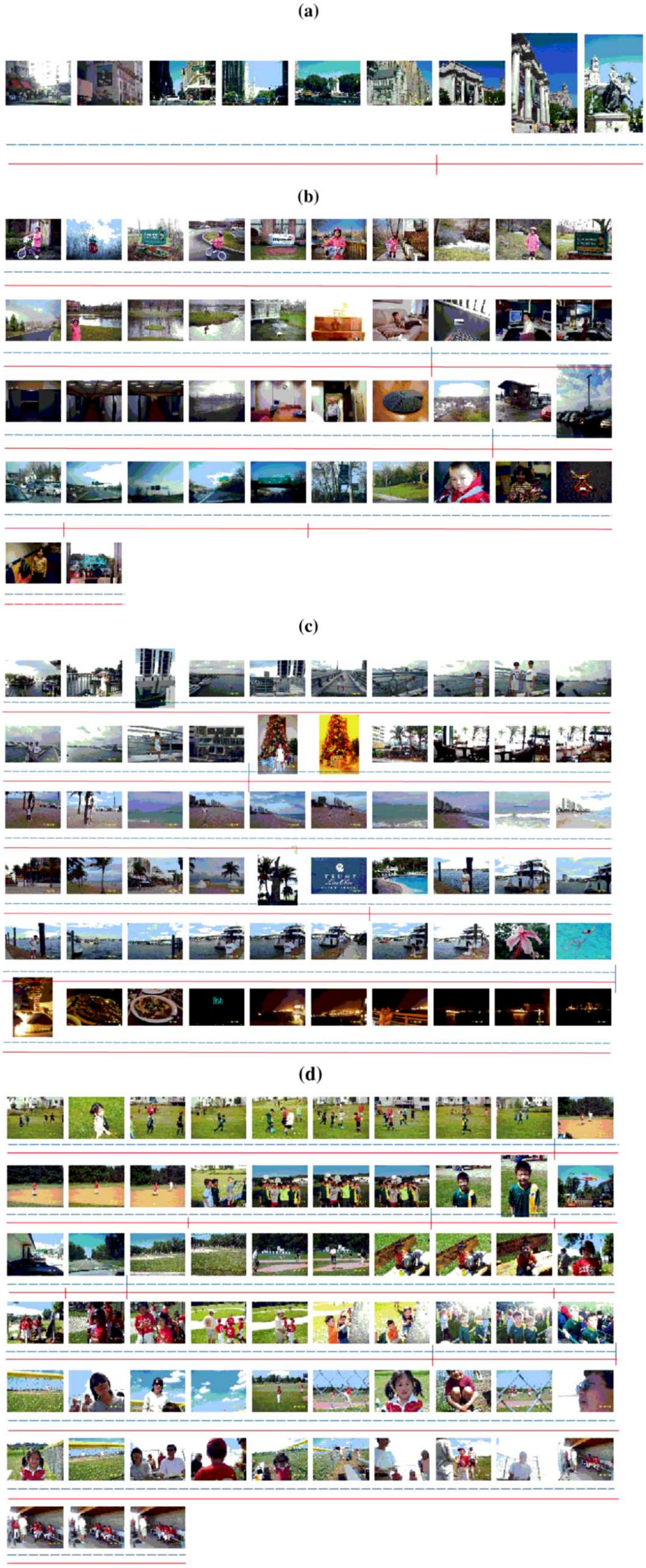


Fig. 5. Event clustering results on four photo collections (a)–(d). The dashed blue lines denote clustering results using time only. The solid red lines represent the results using time+GPS information. In both cases, different clusters are separated by a vertical bar (event boundary).

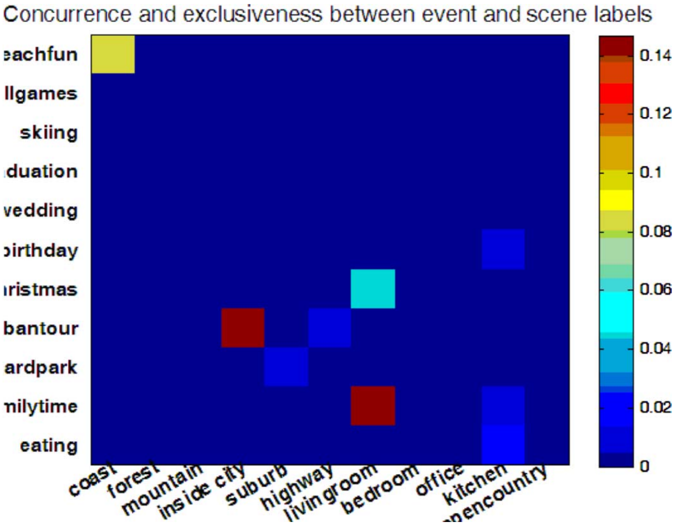


Fig. 6. Correlation between event and scene labels.

in Fig. 6, the event and scene labels are strongly correlated. Some of them are often concurrent, e.g., the beach-time event and the coast scene. More importantly, we find that some labels are mutually exclusive (negative correlation), for example, the yardpark event and the inside-city scene.

To model these two types of correlation, we employ the function

$$r_c(s^k,e)=\begin{cases}\delta(s^k=1), & \text{if } s^k \text{ and } e \text{ are concurrent}\\ -\delta(s^k=1), & \text{if } s^k \text{ and} e \text{ are exclusive}\\ 0, & \text{otherwise}\end{cases} \tag{6}$$

where exclusive labels mean that the two labels never appear together, and the concurrent labels means the the correlation between two labels are above a threshold (0.05 in this experiment). By searching through the training images, our algorithm can find concurrent and exclusive labels automatically. Fig. 6 shows these correlation pairs obtained from the training photos.

We have discussed the feature functions $f_g^e(C)$ and $r_c(S^k,e)$ for event annotation. Taking these functions into account, the log-likelihood function becomes

$$\begin{aligned} L = &-\log Z + \boldsymbol{\alpha}^e \cdot \mathbf{f}_g^e(\mathbf{x}) \\ &+ \sum_k \sum_i \boldsymbol{\beta}^k \cdot \mathbf{f}_s^k\left(\mathbf{x}_i, s_i^k\right) \\ &+ \sum_k \sum_i \boldsymbol{\lambda}^k \cdot \mathbf{r}_s^k(i,i+1) \\ &+ \sum_k \sum_i \mu^{k,e} r_c\left(s_i^k, e\right), \end{aligned} \tag{7}$$

where $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\mu}$ are parameters to be learned, $\mathbf{f}_s^k$ denotes the feature function of scene class k for each photo and $\mathbf{r}_s^k$ denotes for correlation function through time and GPS, as defined in Section 4.

Our complete log-likelihood function L is now more complex than the simple version in (1). The number of parameters is large, which makes it likely for the model to overfit. To reduce the overfitting, we add the following constraints to reduce the model complexity, and thus make it resistant to overfitting. We assume $\mu^{k,e} = \mu_1$ for $r_c(s_i^k,e) > 0$ and $\mu^{k,e} = \mu_2$ for $r_c(s_i^k,e) < 0$. Thus we only need two variables to represent the correlation of events and scenes. Finally, we add the constraint that $\boldsymbol{\alpha}^e = \mathbf{1}$ for all e. By observing (5) we can see that $\mathbf{f}_g^e(C)$ is properly normalized, so removing the parameter $\boldsymbol{\alpha}^e$ is reasonable.

After these simplifications, we can train the CRF models by minimizing L in (7). The training algorithm is summarized as follows. Given the log-likelihood in (7), we can compute the gradient ∇L. By denote the parameters as $\Phi = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\mu})$, we can expand the log-likelihood using Taylor series

$$L(\Phi + \Delta\Phi) = L(\Phi) + \nabla L(\Phi)^T \Delta\Phi + \frac{1}{2}\Delta\Phi^T B \Delta\Phi$$

which attains its extremum at

$$\nabla L(\Phi) + B\Delta\Phi = 0. \tag{8}$$

Note B is the Hessian matrix which can be efficiently updated by conjugate-gradient methods as in [32].

From (8) we can obtain the updating scheme

$$\Phi_{n+1} = \Phi_n + \Delta\Phi_n = \Phi_n - B^{-1}\nabla L(\Phi_n).$$

Since L is a convex function subject to Φ, this iterative approach will find the optimal parameters.

After learning the parameters, we can use our model to annotate the testing images. This process is accomplished by maximizing (7) subject to the label variable s_i^k and e_c with the trained parameters. The belief propagation method is employed for this task, which iteratively passes positive real vector valued messages between the variables until convergence. Compared with traditional mean field algorithm, belief propagation is more robust and more fast for the inference task [39].

VIII. Experimental Results and Discussions

From the geotagged dataset, we randomly select 50% of all the folders for training and the rest for testing. The testing results are compared with ground truth. Note that the ground truth of scene labels is for individual photos, while the ground truths of events are for photo subcollections. Although it takes days to train the basic concept detectors and CRF model, the testing process is very fast. The average annotation time for each image is less than one second.

First, we show the accuracy of scene labeling. Since scene-level annotation is a multilabel problem, we compute the precision and recall for each label, as shown in Fig. 7. From the figure, the recalls for most classes are satisfactory, while the precisions are lower. In other words, false alarms are the main errors in scene annotation. This demands more attention in future work.

At the event level, we compare our annotations with the real events at the subcollection level. We construct a confusion matrix for 11 events over subcollections, as shown in Fig. 8. Most classes are annotated successfully. Some event pairs may be confused because they share much visual similarity: wedding confused with graduation when graduates happen to wear white gown, and birthday confused with eating because both can show

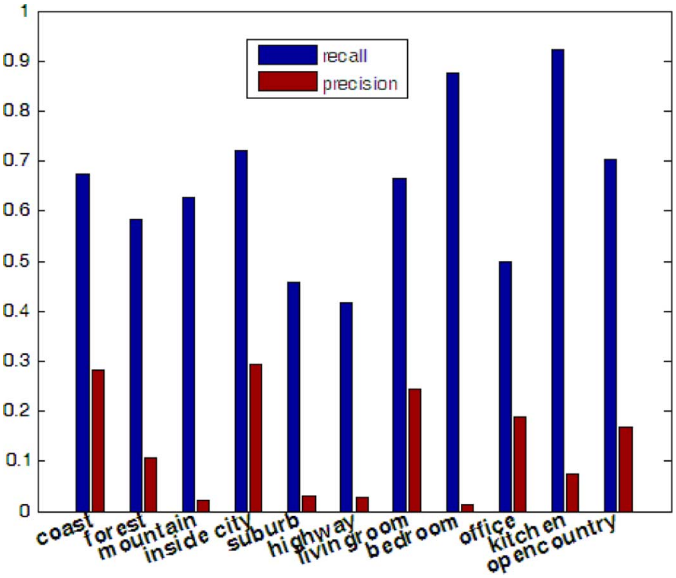
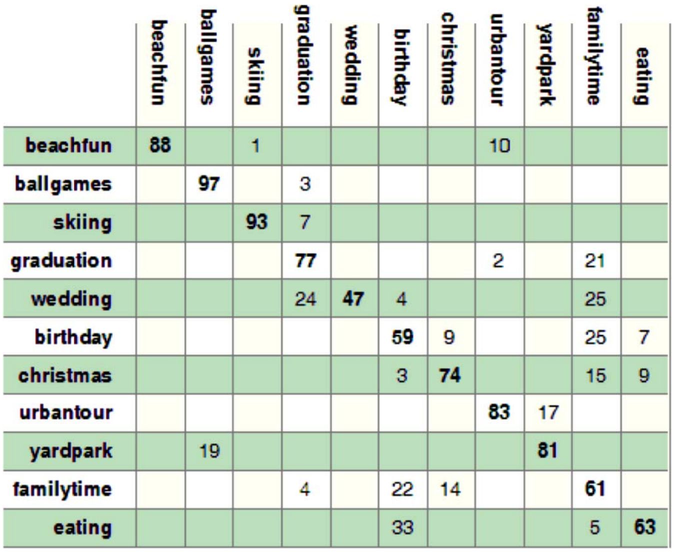


Fig. 7. Precision-recall for scene annotation.



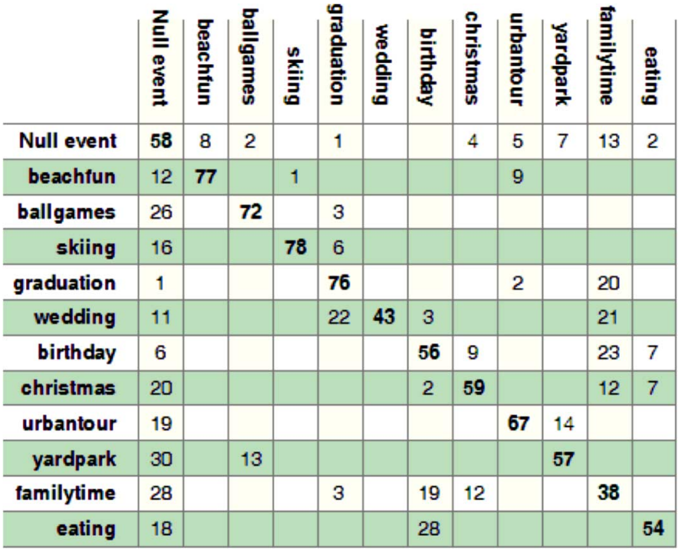
	beachfun	ballgames	skiing	graduation	wedding	birthday	christmas	urbantour	yardpark	familytime	eating
beachfun	**88**		1					10			
ballgames		**97**		3							
skiing			**93**	7							
graduation				**77**				2		21	
wedding				24	**47**	4				25	
birthday						**59**	9			25	7
christmas						3	**74**			15	9
urbantour								**83**	17		
yardpark		19							**81**		
familytime				4		22	14			**61**	
eating						33				5	**63**

Fig. 8. Confusion matrixes for the 11 events (74.8% average accuracy). Each column corresponds to ground-truth label of one event class. Each row corresponds to class labels predicted by the algorithm. All the numbers are percentage numbers.

food on the table (unless we can detect the birthday cake explicitly).

Event annotation becomes more difficult if we also consider the null event class. Fig. 9 shows the new confusion matrix for all the subcollections, including those of the null class. Unfortunately, some null-event subcollections are misclassified as one of the known events. However, we are pleased that such misclassification is limited and almost evenly distributed among all classes.

To test the benefit of our CRF model and the GPS information, we compare the annotation results by our model with GPS and time against those by using time information only, and those by individual detectors. To make a fair comparison, we consider only those collections with both GPS and time tags. Figs. 10 and 11 show the precision and recall for scene and event annotation, respectively. Figs. 12 and 13 compare the average precision (AP) for scene and event annotation, respectively. Our hierarchical event-scene model with time and GPS improves sig-



	Null event	beachfun	ballgames	skiing	graduation	wedding	birthday	christmas	urbantour	yardpark	familytime	eating
Null event	**58**	8	2		1			4	5	7	13	2
beachfun	12	**77**		1					9			
ballgames	26		**72**		3							
skiing	16			**78**	6							
graduation	1				**76**				2		20	
wedding	11				22	**43**	3				21	
birthday	6						**56**	9			23	7
christmas	20						2	**59**			12	7
urbantour	19								**67**	14		
yardpark	30		13							**57**		
familytime	28				3		19	12			**38**	
eating	18						28					**54**

Fig. 9. Confusion matrixes with the null event class (61.4%).

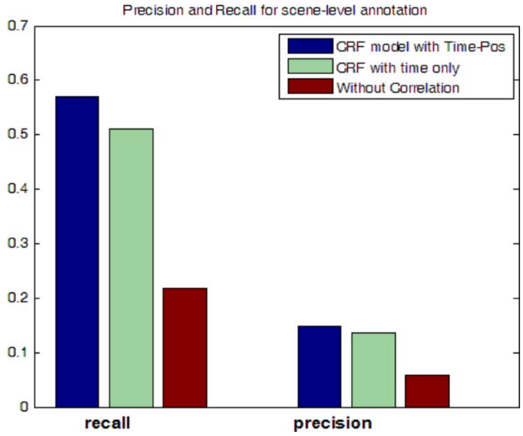


Fig. 10. Comparing scene-level annotation accuracy by our CRF model using both time and GPS, with the model using time only, and with the single detectors (without modeling correlations).

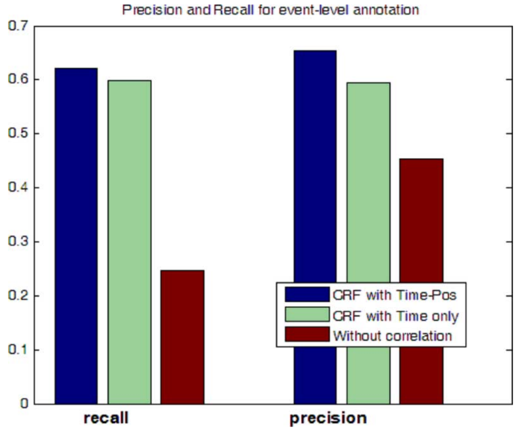


Fig. 11. Comparing event annotation by the proposed model using both time and GPS, with the model using time only, and with the individual detectors without modeling correlations.

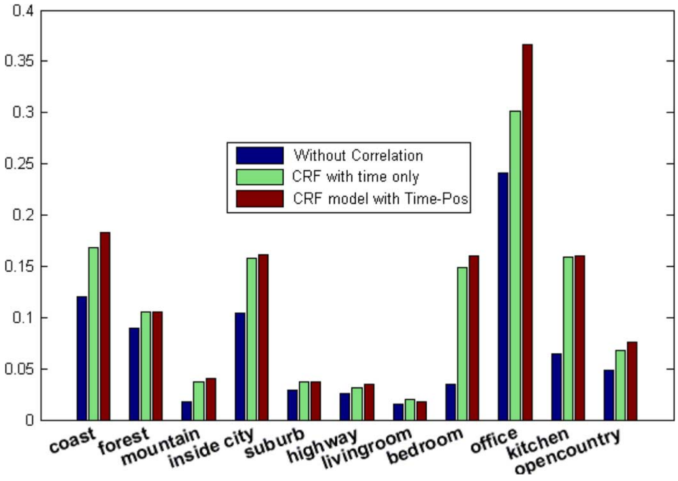


Fig. 12. Average precision (AP) for scene annotation.

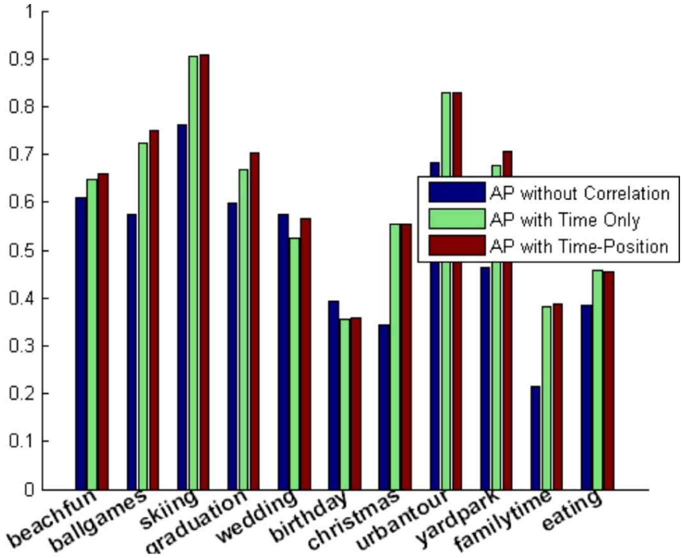


Fig. 13. AP for event annotation.

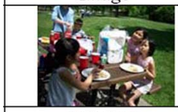
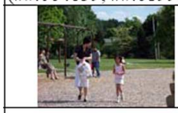
Image	*No CRF*	*CRF (with Time)*	*CRF (with Time+GPS)*
	Event: NULL	Event: **yardpark**	Event: **yardpark**
16:36:37, (##.064886,##.620047)	Scene: NULL	Scene: **suburb**	Scene: **suburb**
	Event: Null	Event: **yardpark**	Event: **yardpark**
16:37:32, (##.064628, ##.619994)	Scene: NULL	Scene: ***forest***, inside-city	Scene: ***forest***
	Event: NULL	Event: **yardpark**	Event: **yardpark**
17:33:20, (GPS missing)	Scene: NULL	Scene: ***forest***, **suburb**	Scene: ***forest***, **suburb**
	Event: yardpark	Event: ballgame	Event: **yardpark**
19:45:18, (##.064839, ##.619950)	Scene: forest	Scene: ***forest***, **suburb**	Scene: ***forest***, **suburb**
	Event: **yardparl**	Event: ballgame	Event: **yardpark**
20:14:14, (##.065428, ##.620514)	Scene: NULL	Scene: inside-city, **suburb**	Scene: inside-city, **suburb**

Fig. 14. Example results for one photo collection (a single event based on ground truth clustering). The event and scene labels in **bold** face are correct when compared to the ground truth. The scene labels in *italics* (e.g., forest for some photos) are also semantically correct even though the ground truth did not contain them.

nificantly both the precision and recall in both cases, in terms of the overall performance and those on the vast majority of individual classes (only wedding and birthday are slightly worse). Although the model with time only is not as competitive as the full model, it still performed much better than the isolated single detectors.

To illustrate the success of our hierarchical event-scene model, Figs. 14–16 contain annotation results on three photo collections at both event and scene levels. Due to the space limits, we can only show the most representative photos in each collection. Note that although each collection is a single event based on ground truth clustering, the actual event clustering by different algorithms may differ. The event and scene labels in bold face are correct when compared to the ground truth. In addition, the scene labels in italics are also semantically correct even though the ground truth did not contain them (typical for a multilabel problem such as scene classification). Clearly, the proposed hierarchical event-scene model provides better annotation at both event and scene levels than the isolated detectors, with the model using full time and GPS information as the best. Most notably, all the photos in the same collection share the same event label and more consistent scene labels, thanks to accurate event clustering using both time and GPS information and powerful interactions within the hierarchical model.

In general, GPS information is more available for outdoor photos although our experience is that wood-frame houses and rooms with large windows also allow reasonable GPS reception. Because our framework can also handle the collections in which the GPS information is missing for part of the photos, improved annotation can also be obtained for the indoor photos that are grouped together with the outdoor photos of the same events.

It is worthy noting that the goal of the event clustering algorithm is to produce semantically meaningful event grouping that is at the same time as close as possible to the group by the owner of the photos. In that sense, the clustering is intended to be useful on its own as an aid to photo management by the users, as well as provide a sound basis for the hierarchical annotation process. While ground truth clusters by the users would certainly be better for the annotation process, we have shown that the clustering algorithm proves to be adequate, precisely because the clustering algorithm (with the combination of timestamp and GPS) is able to group the photos in a similar fashion as the photo owners. Note that the users have no knowledge of the clustering algorithm when providing the ground truth clusters.

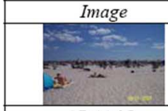
Image	*NO CRF*	*CRF (with time)*	*CRF (with time+GPS)*
	Event: NULL	Event: **beach-time**	Event: **beach-time**
17:44:05, (##.593733, ##.507444)	Scene: NULL	Scene: **coast**, suburb, open-country	Scene: **coast**, suburb, open-country
	Event: **beach-time**	Event: **beach-time**	Event: **beach-time**
17:46:07, (##.593086,##.506550)	Scene: **Coast**	Scene: **coast**, open-country	Scene: **coast**, open-country
	Event: NULL	Event: **beach-time**	Event: **beach-time**
17:54:55, (##.592978, ##.506458)	Scene: NULL	Scene: NULL	Scene: NULL
	Event: **beach-time**	Event: **beach-time**	Event: **beach-time**
18:00:51, (##.592944, ##.506422)	Scene: NULL	Scene: **coast**	Scene: **coast**
	Event: skiing	Event: **beach-time**	Event: **beach-time**
18:03:04, (GPS missing)	Scene: NULL	Scene: NULL	Scene: **coast**
	Event: NULL	Event: **beach-time**	Event: **beach-time**
18:22:03, (GPS missing)	Scene: **NULL**	Scene: **coast**	Scene: **coast**
	Event: familytime	Event: **beach-time**	Event: **beach-time**
18:32:52, (##.593022, ##.506550)	Scene: NULL	Scene: suburb	Scene: **coast**, suburb
	Event: NULL	Event: **beach-time**	Event: **beach-time**
18:52:23, (##.592992, ##.506556)	Scene: NULL	Scene: NULL	Scene: **coast**
	Event: familytime	Event: **beach-time**	Event: **beach-time**
19:12:38, (##.593003, ##.506533)	Scene: NULL	Scene: Inside-city, suburb	Scene: Inside-city, suburb

Fig. 15. Example results for one photo collection (a single event based on ground truth clustering).

IX. Conclusion and Future Work

We addresses the problem of annotating photo collections instead of single images. We built a sizable collection of geotagged personal photos, and defined a compact ontology of events and scenes suitable for consumers. We construct a CRF-based model that accounts for two types of correlations: 1) correlation by time and GPS tags and 2) correlation between scene- and event-level labels. Extensive experiments have shown that our hierarchical model significantly improves annotation in both precision and recall. Future directions include exploring (better) alternative baseline scene classifiers, integrating the physical place tags that can be derived from the GPS coordinates [15], expanding the scene-event ontology, and finding a solution to reduce the relative high level of confusion between certain events.

Image	*NO CRF*	*CRF (with time)*	*CRF (with GPS+Time)*
	Event: skiing	Event: **urban-tour**	Event: **urban-tour**
12:02:12, (##.276553, ##.738728)	Scene: NULL	Scene: **inside-city**, suburb	Scene: **inside-city**, suburb
	Event: yardpark	Event: **urban-tour**	Event: **urban-tour**
12:13:32, (##.276694, ##.740703)	Scene:	Scene: forest, **inside-city**, suburb	Scene: **inside-city**, suburb
	Event: NULL	Event: beach-time	Event: **urban-tour**
13:47:37 (GPS missing)	Scene: NULL	Scene: **inside-city**, suburb ***highway***, coast	Scene: **inside-city**, suburb, ***highway***, coast
	Event: NULL	Event: beach-time	Event: **urban-tour**
13:48:19, (##.264614, ##.750239)	Scene: NULL	Scene: suburb	Scene: suburb
	Event: beach-tir	Event: beach-time	Event: **urban-tour**
13:49:06, (##.264214, ##.747711)	Scene: NULL	Scene: **inside-city,** suburb ***highway***	Scene: **inside-city,** suburb, ***highway***
	Event: beach-tir	Event: **urban-tour**	Event: **urban-tour**
16:11:39, (##.655917, ##.612367)	Scene: NULL	Scene: **inside-city**, open-country	Scene: **inside-city**, open-country

Fig. 16. Example results for one photo collection (a single event based on ground truth clustering).

It is important to align with contemporary interests when expanding the annotation ontology. The popular tags published by Flickr include animals, birthday, architecture, art, baby, bird, camping, Christmas, car, cat, church, city, clouds, dog, family, flower, football, garden, girl, hiking, house, island, kids, lake, landscape, museum, ocean, party, people, portrait, river, rock, show, sky, snow, street, sunset, tree, urban, water, wedding, zoo, and so on. Many of them are related to scenes and events that should be considered in the future work.

References

[1] M. Agarwal and K. Konolige, "Real-time localization in outdoor environments using stereo vision and inexpensive GPS," in *Int. Conf. Pattern Recognition*, 2006.

[2] A. Amir, M. Berg, S.-F. Chang, W. Hsu, G. Iyengar, and C. Lin, "Ibm research trecvid-2003 video retrieval system," *NIST TRECVID*, 2003.

[3] J. Assfalg, M. Bertini, C. Colombo, and A. Bimbo, "Semantic annotation of sports videos," *IEEE Trans Multimedia*, 2002.

[4] Y. Aytar, O. Orhan, and M. Shah, "Improving semantic concept detection and retrieval using contextual estimates," in *Proc. ICME*, 2007.

[5] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. M. Blei, and M. I. Jordan, "Matching words and pictures," *J. Mach. Learn. Rese.*, vol. 3, pp. 1107–1135, 2003.

[6] M. Boutell, X. Shen, J. Luo, and C. Brown, "Learning multi-label semantic scene classification," *Pattern Recognition*, vol. 37, no. 9, pp. 1757–1771, 2004.

[7] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 603–619, 2002.

[8] M. Cooper, "Collective media annotation using undirected random field models," in *Int. Conf. Semantic Computing*, 2007, pp. 337–343.

[9] M. Cooper, J. Foote, A. Girgensohn, and L. Wilcox, "Temporal event clustering for digital photo collections," *ACM Trans. Multimedia Comput. Commun. Applicat.*, vol. 1, no. 3, pp. 269–288, 2005.

[10] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2005, pp. 524–531.

[11] A. Jaffe, M. Naaman, T. Tassa, and M. Davis, "Generating summaries and visualization for large collections of geo-referenced photographs," in *ACM Int. Workshop on Multimedia Information Retrieval*, 2006, pp. 89–98.

[12] W. Jiang, S.-F. Chang, and A. Loui, "Kernel sharing with joint boosting for multi-class concept detection," in *Proc. CVPR (Workshop on Semantic Learning in Multimedia)*, 2007.

[13] T. Joachims, "Making large scale SVM learning practical," in *Advances in Kernel Methods—Support Vector Learning*. Cambridge, MA: MIT Press, 1999.

[14] M. Johnson and R. Cipolla, "Improved image annotation and labeling through multi-label boosting," in *Brit. Machine Vision Conf.*, 2005.

[15] D. Joshi and J. Luo, "Inferring generic places based on visual content and bag of geotags," in *ACM Conf. Image and Video Retrieval*, 2008.

[16] L. Kennedy, M. Naaman, S. Ahern, R. Nair, and T. Rattenbury, "How flickr helps us make sense of the world: Context and content in community-contributed media collections," in *ACM Conf. Multimedia*, 2007, pp. 631–640.

[17] S. Kumar and M. Hebert, "Discriminative random fields," *Int. J. Comput. Vis.*, vol. 68, no. 2, pp. 179–201, 2006.

[18] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Int. Conf. Machine Learning*, 2001, pp. 282–289.

[19] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2005.

[20] L.-J. Li and L. Fei-Fei, "What, where and who? classifying event by scene and object recognition," in *IEEE Int. Conf. Computer Vision*, 2007.

[21] L. Liao, D. Fox, and H. Kautz, "Location-based activity recognition," in *Neural Inform. Process. Syst.*, 2005.

[22] J.-H. Lim, Q. Tian, and P. Mulhem, "Home photo content modeling for personalized event-based retrieval," *IEEE Multimedia*, vol. 10, no. 4, pp. 28–37, Oct.-Dec. 2008.

[23] A. Loui and A. Savakis, "Automated event clustering and quality screening of consumer pictures for digital albuming," *IEEE Trans. Multimedia*, vol. 5, no. 3, pp. 390–402, Jun. 2003.

[24] J. Luo, M. Boutell, and C. Brown, "Pictures are not taken in a vacuum—An overview of exploiting context for semantic scene content understanding," *IEEE Signal Process.*, vol. 23, pp. 101–114, 2006.

[25] W. Ma and H. J. Zhang, "Benchmarking of image features for content-based retrieval," in *Proc. Signals, Systems and Computers*, 1998.

[26] M. Naaman, "Leveraging Geo-Referenced Digital Photographs," Ph.D. dissertation, Stanford Univ., Stanford, CA, 2005.

[27] M. Naaman, Y. Song, A. Paepcke, and H. Garcia-Molina, "Automatic organization for digital photographs with geographic coordinates," in *Int. Conf. Digital Libraries*, 2004, vol. 7, pp. 53–62.

[28] C. Pantofaru and M. Hebert, A Comparison of Image Segmentation Algorithms Carnegie Mellon Univ., Pittsburgh, PA, Tech. Rep. CMU-RI-TR-05-40, 2005, pp. 383–394.

[29] P. Quelhas, F. Monay, J. Odobez, D. Gatica-Perez, T. Tuytelaars, and L. Van, "Gool. Modeling scenes with local descriptors and latent aspects," in *IEEE Int. Conf. Computer Vision*, 2005, vol. 1.

[30] Y. Rui, T. Huang, and S.-F. Chang, "Image retrieval: Current techniques, promising directions, and open issues," *J. Vis. Commun. Image Represent.*, vol. 10, pp. 39–62, 1999.

[31] B. Russell, A. Torralba, K. Murphy, and W. Freeman, "Labelme: A database and web-based tool for image annotation," *Int. J. Comput. Vis.*, 2007.

[32] F. Sha and F. Pereira, "Shallow parsing with conditional random fields," in *Proc. HLT-NAACL*, 2003, vol. 1, pp. 134–141.

[33] I. Simon, N. Snavely, and S. Seitz, "Scene summarization for online image collections," in *IEEE Int. Conf. Computer Vision*, 2007, pp. 1–8.

[34] J. Sivic, B. Russell, A. Efros, A. Zisserman, and W. Freeman, "Discovering object categories in image collections," in *IEEE Int. Conf. Computer Vision*, 2005, vol. 1, no. 65.

[35] C. A. Sutton, K. Rohanimanesh, and A. McCallum, "Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data," in *Int. Conf. Machine Learning*, 2004.

[36] D. Tal and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *IEEE Proc. Int. Conf. Computer Vision*, 2001.

[37] K. Tieu and P. Viola, "Boosting image retrieval," *Int. J. Comput. Vis.*, 2004.

[38] J. Wang, J. Li, and G. Wiederhold, "SIMPLIcity: Semantics-sensitive integrated matching for picture libraries," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 9, pp. 947–963, 1999.

[39] Y. Weiss, "Comparing the Mean Field Method and Belief Propagation for Approximate Inference in MRFs," in *Advanced Mean Field Methods*, D. Saad and M. Opper, Eds. Cambridge, MA: MIT Press, 2001.

[40] Y. Wu, J.-Y. Bouguet, A. Nefian, and I. Kozintsev, "Learning concept templates from web images to query personal image databases," in *IEEE Int. Conf. Multimedia and Expo*, 2007, pp. 1986–1989.

[41] Y. Wu, I. Kozintsev, J.-Y. Bouguet, and C. Dulong, "Sampling strategies for active learning in personal photo retrieval," in *IEEE Int. Conf. Multimedia and Expo*, 2006, pp. 529–532.

[42] R. Yan, M. Chen, and A. Hauptmann, "Mining relationship between video concepts using probabilistic graphical model," in *IEEE Int. Conf. Multimedia & Expo*, 2006.

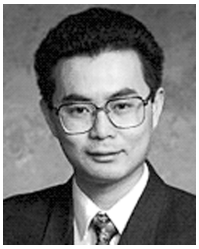
[43] A. Yavlinsky and D. Heesch, "An online system for gathering imagesimilarity judgments," in *ACM Proc. Multimedia*, 2007.

[44] J. Yuan, J. Luo, and Y. Wu, "Mining compositional features for boosting," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2008.

[45] L. Zelnik-Manor and M. Irani, "Event-based analysis of video," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2001.

Liangliang Cao received the B.E. degree from the University of Science and Technology of China in 2003, and the M.Phil. degree from the Chinese University of Hong Kong in 2005. He is now pursuing the Ph.D. degree in the Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign.

During the summer of 2007 and 2008, he was a research intern with Kodak Research Laboratories, Rochester, New York. He has worked on projects including 3–D object reconstruction and retrieval, image segmentation and recognition, consumer photo annotation and surveillance video analysis. His current research interests include computer vision, data mining and statistical learning.



Jiebo Luo (M'96–SM'99–F'08) received the B.S. degree from the University of Science and Technology of China in 1989 and the Ph.D. degree from the University of Rochester, Rochester, NY, in 1995, both in electrical engineering. He is a Senior Principal Scientist with the Kodak Research Laboratories, Rochester, NY. His research interests include signal and image processing, machine learning, computer vision, multimedia data mining, and computational photography. HE has authored over 130 technical papers and holds nearly 50 U.S. patents.

Dr. Luo has been involved in organizing numerous leading technical conferences sponsored by IEEE, ACM, and SPIE, including general chair of the 2008 *ACM International Conference on Content-based Image and Video Retrieval* (CIVR), area chair of the *2008 IEEE International Conference on Computer Vision and Pattern Recognition* (CVPR), program co-chair of the 2007 *SPIE International Symposium on Visual Communication and Image Processing* (VCIP), organizing committee member of the 2008 *ACM Multimedia Conference*, 2008/2006 *IEEE International Conference on Multimedia and Expo* (ICME), and 2002 *IEEE International Conference on Image Processing* (ICIP). Currently, he is on the editorial boards of the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (TPAMI), the IEEE TRANSACTIONS ON MULTIMEDIA (TMM), *Pattern Recognition* (PR), and the *Journal of Electronic Imaging*. He is the incoming Editor-in-Chief of the *Journal of Multimedia* (Academy Publisher). He is a guest editor for a number of influential special issues, including "Image Understanding for Digital Photos" (PR, 2005), "Real-World Image Annotation and Retrieval" (TPAMI, 2008), "Integration of Content and Context for Multimedia Management" (TMM, 2008), "Event Analysis in Video" (IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, 2009), and "Probabilistic Graphic Models in Computer Vision" (TPAMI, 2009). He is a Kodak Distinguished Inventor, a winner of the 2004 Eastman Innovation Award (Kodak's highest technology prize), a member of ACM, and a Fellow of SPIE.

Henry Kautz received the A.B. degree in mathematics from Cornell University, Ithaca, NY, the M.A. degree in creative writing from the The Johns Hopkins University, Baltimore, MD, the M.Sc. in computer science from the University of Toronto, Toronogto, ON, Canada, and the Ph.D. degree in computer science from the University of Rochester, Rochester, NY.

He had 13 years of industrial experience at AT&T Bell Laboratories, where he was head of the AI Principles Research Department. In 2000, he became a faculty member at the University of Washington, Seattle. From 2006 to 2007, he founded and directed the Intelligent Systems Center at Kodak Research Laboratories, Rochester, NY. He is currently Chair of the Department of Computer Science at the University of Rochester, where he directs the Laboratory for Assisted Cognition Environments. His research encompasses artificial intelligence, pervasive computing, and assistive technology. His contributions include efficient algorithms for logical and probabilistic reasoning, the planning as satisfiability framework, and methods for behavior recognition from sensor data. He is the author of more than 70 refereed publications.

Dr. Kautz is a Fellow of the American Association for the Advancement of Science, a Fellow of the Association for Advancement of Artificial Intelligence, and winner of the Computers & Thought Award from the International Joint Conference on Artificial Intelligence.

Thomas Huang (S'61–M'63–SM'76–F'79–LF'01) received the B.S. degree in electrical engineering from National Taiwan University, Taipei, Taiwan, and the M.S. and Sc.D. degrees in electrical engineering from the Massachusetts Institute of Technology (MIT), Cambridge.

He was on the faculty of the Department of Electrical Engineering at MIT from 1963 to 1973, and at the faculty of the School of Electrical Engineering and Director of its Laboratory for Information and Signal Processing at Purdue University, West Lafayette, IN, from 1973 to 1980. In 1980, he joined the University of Illinois at Urbana-Champaign, where he is now William L. Everitt Distinguished Professor of Electrical and Computer Engineering, Research Professor at the Coordinated Science Laboratory, Head of the Image Formation and Processing Group at the Beckman Institute for Advanced Science and Technology, and Co-chair of the Institute's major research theme: human-computer intelligent interaction. His professional interests lie in the broad area of information technology, especially the transmission and processing of multidimensional signals. He has published 20 books and over 500 papers in network theory, digital filtering, image processing, and computer vision.

Dr. Huang is a Member of the National Academy of Engineering, a Foreign Member of the Chinese Academies of Engineering and Science, and a Fellow of the International Association of Pattern Recognition and the Optical Society of America, and has received a Guggenheim Fellowship, an A. von Humboldt Foundation Senior U.S. Scientist Award, and a Fellowship from the Japan Association for the Promotion of Science. He received the IEEE Signal Processing Society's Technical Achievement Award in 1987, and the Society Award in 1991. He was awarded the IEEE Third Millennium Medal in 2000. Also in 2000, he received the Honda Lifetime Achievement Award for "contributions to motion analysis." In 2001, he received the IEEE Jack S. Kilby Medal. In 2002, he received the King-Sun Fu Prize, International Association of Pattern Recognition, and the Pan Wen-Yuan Outstanding Research Award. He is a Founding Editor of the *International Journal of Computer Vision, Graphics, and Image Processing* and Editor of the *Springer Series in Information Sciences*, published by Springer.