Annotating Collections of Photos Using Hierarchical Event and Scene Models

Liangliang Cao*, Jiebo Luo+, Henry Kautz+, Thomas S. Huang*

+Kodak Research Laboratories
Eastman Kodak Company
{jiebo.luo,henry.kautz}@kodak.com

*Dept. of Electrical and Computer Engineering
University of Illinois
{cao4,huang}@ifp.uiuc.edu

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 explore the associated GPS and time information for semantic annotation. First, we employ a constrained clustering method to partition a photo collection into event-based sub-collections, 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 multi-level 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 using a geotagged customer photo collection database, which consists of over 100 folders and is labeled for 12 events and 12 scenes.*

1. Introduction

In recent years, the flourishing of digital photos has presented a grand challenge to the computer vision research community: can a computer vision system produce satisfactory annotations automatically for personal photos? 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. This characteristic is largely neglected and unexploited in previous research. On the other hand, photo annotation requires more descriptive annotation of consumer activities and this is beyond the scope and capability of traditional image annotation and retrieval systems [1] –[4]. While the concern of traditional systems is the content of an isolated image, we believe that the task of photo annotation should draw more attention to what happened in the entire collection of related images.

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. In previous work, event classification is limited to video analysis [5] [6] [7] or specific sports activities [8] [9]. 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. However, with a collection of photos, it becomes possible to explore the semantic correlation among multiple photos. In this scenario, an event label is selected to annotate a group of photos that form the event.

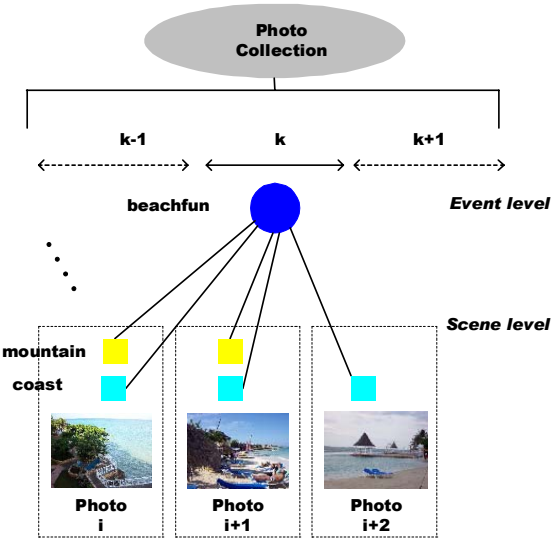

Fig. 1: Hierarchical annotation of photo collections.

In addition to the event labels, we are also interested in 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 make our annotation more descriptive. Since a photo can belong to more than one scene class, e.g., a beach photo may also contain mountain, this task is a multi-label problem [10].

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

978-1-4244-2243-2/08/$25.00 ©2008 IEEE

common to all the photos in that group. In the lower level, we assign each photo one or more *scene* labels. To our best knowledge, this two-level representation of photo collections has not been reported for image annotation.

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 factors make photo annotation a challenging task. To improve the annotation accuracy, we explore different sources of information associated with photo collections.

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.

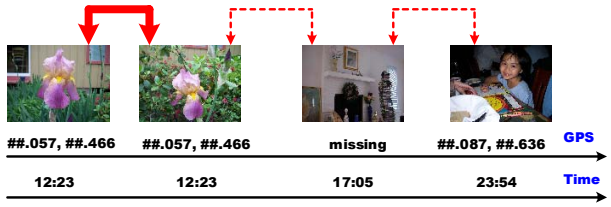

Fig. 2: Modeling correlation in a photo collection using time and GPS tags. The thickness of the red lines indicates the strength of the correlations between images. Note the actual coordinates of the GPS tags are removed to preserve privacy.

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, on the basis of the discriminative model of Conditional Random Field (CRF).

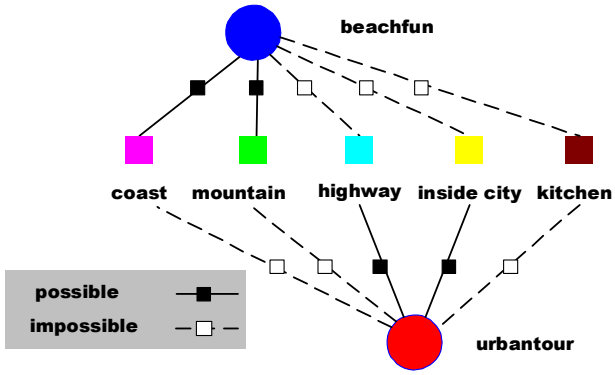

Fig.3: Correlation between scene labels and event labels.

To test our method, we built a database of personal photos. This dataset consists of more than 100 collections (folders) contributed by different users, and is labeled for 12 events and 12 scenes. While every photo is tagged by a timestamp, a majority of the images also possess *genuine* capture-time GPS tags (In contrast, most geotagged photos in Flickr were tagged manually after the fact). To the best of our knowledge, there is no other geotagged customer photo collection dataset of this nature and scale.

This paper is organized as follows. Section 2 reviews the related work and emphasizes our novelties. Section 3 describes our dataset and the manual labeling needed for our experiments. Section 4 presents the basic model for scene annotation, which takes time and GPS information into account. Section 5 considers partitioning photos into event clusters. Section 6 takes into account the relationship between events and scenes and builds our complete annotation model. Experimental results are in Section 7 and we conclude this paper in Section 8.

2. Related Work

Recently, much research work has been done in single image classification. A large percentage of the work is on general object recognition [11] [12] [13], which is different from our research problem. Another major research theme is scene recognition [14] [15] [16], which is part of the interests of this study. These techniques can be considered as the baseline of annotation for single photos and any of them would serve as a good baseline upon which we can build our system. In contrast, our work focuses on annotating an entire collection instead of merely one image, and modeling the correlation among images, and between events and scenes. 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 [6].A few studies involved image annotation with multiple labels [17] [18] but nevertheless were limited to annotating single photos as opposed to photo collections.

GPS information was used to classify certain reoccurring human activity in [19]. However, it relies on continuous GPS traces and does not use any visual information.

Our major contributions are three-fold. First, we are the first to consider the problem of personal photo annotation at the collection level. Second, we developed effective methods to utilize time and GPS information for photo annotation even when some of the GPS records are missing. Finally, we explore the relationship between event and scene labels in order to produce a detailed description of photo collections and the photos within them.

3. 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 4 to 249 photos). These collections include extensive photo content. Some examples of the dataset are shown in Fig.4.

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).

Fig.4. Example of our dataset. 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.

The dataset is completely labeled by the researcher. 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 1.

We also labeled each image with the scene labels using the class definitions from [14]: 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.

Table 1: Definitions of the12 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 appear. Better with groom
Birthday	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

4. Scene-Level Modeling

To model the correlation between the labels, we employ a conditional random field (CRF) model. CRF is a probabilistic model presented by Lafferty, McCallum and Pereira [20]. 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 6.

When the 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 more unlikely 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 5 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.

We first define a number of notations. In a photo collection, the photos are represented by $\{x_i\}$, $i = 1,2,..., N$. The time tags and GPS tags are denoted as $\{t_i\}$ and $\{p_i\}$, where $p_i = NULL$ when the GPS is missing.

We use s_i^k to denote labeling status of the i th photo for scene class k, with $1 \le k \le 11$. Here $s_i^k = 1$ means the scene label is true for x_i, while $s_i^k = 0$ means that the scene label is null. Note that if $s_i^k = 0$ for all $1 \le k \le 11$, it means that x_i is not labeled as any of the known scene labels.

To model the correlation using time and GPS, we employ the conditional likelihood function for scene k:

$$L_s^k = -\log Z_s + \sum_i \beta^k f_s^k(x_i, s_i^k) + \sum_{i,j \in N_c} \lambda^k R_s^k(s_i^k, s_j^k, t_i, t_j, p_i, p_j) \quad (1)$$

where f_s stands for the feature function for individual photos, and R_s^k models the correlation between s_i^k and s_j^k. The subscript $i, j \in N_c$ indicates consecutive photos in the collection, i.e.., $j = i+1$. Z_s stands for a normalization constant. λ^k and β^k are the parameter vectors that are learned from the training data.

Given that the larger the differences are 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 likelihood. With these observations, we define the correlation feature function as

$$R_s^k = \begin{cases} \left(\frac{1}{1+\exp(dt_{ij})}, \frac{1}{1+\exp(dp_{ij})}\right)^T, & \text{if } s_i^k = s_j^k \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

where we use a sigmoid function to model the correlation, with *dt* and *dp* denoting changes in time and location, respectively.

Note that the correlation function defined in (2) is able to handle the situation of missed GPS. If p_i or p_j is *NULL*, we treat $dp_{ij} = \infty$ and thus $\frac{1}{1+\exp(dp_{ij})} = 0$, which means that the GPS tags impose no correlations on the overall likelihood function.

L_s^k acts as the objective function in both the training and testing stages. For training, we learn the parameters (λ, β) which maximizes L_s^k. For testing, given a photo collection $\{x_i\}, \{t_i\}, \{p_i\}$, the labeling $\{s_i^k\}$ that maximizes (1) will infer the most possible labels.

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.

5. 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. 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.

5.1. 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 [21]; however, it cannot tell whether people stay in the same place for a long time or they already move 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 [22]. 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 C with the GPS information added. We iteratively search from the baseline clusters for a sample that is not in C but close to a sample already in C. We add this sample to the same cluster containing its closest neighbors. This iteration will be performed until all the photos are added to C. The only exception is that a substantial cluster that was formed by time only and does not overlap with C is added to C as a new cluster. The details of the clustering algorithm are described below.

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

1. Obtain baseline clusters (sub-collections) 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 $x_t \in C^t$, such that $x_t \notin C$. Also select another example $x_t \in C^t$ satisfying $x_n = \arg\min_x dist(x_t, x)$. Here $dist$ is the Euclidean distance between the time tags.
   4.2 add x_t into C with the cluster label the same as x_n.

Output: C as the final photo sub-collections.

5.2. Clustering evaluation

To evaluate our clustering algorithm, we benchmark against the ground truth set by the photographer who took the photos. Since it is impossible to ask all the users to mark 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) [23], 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) [24], 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 denotes 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).

The computed clustering is $C = \{c_1, c_2, c_3, ..., c_n\}$ and the ground truth is $G = \{g_1, g_2, g_3, ..., g_m\}$. For each cluster c_i, we compute its contribution to the overall error:

$$err_i = 0, \text{ if } \exists c_i = g_j$$
$$err_i = |c_i| \omega_1 N_\Omega, \text{ elseif } \exists c_i = \bigcup_{j \in \Omega} g_j$$
$$err_i = |c_i| \omega_2 N_P, \text{ elseif } \exists g_j = \bigcup_{i \in P} c_j$$
$$err_i = |c_i|, \text{ otherwise.}$$

where $|c_i|$ is the number of samples in $|c_i|$, and N_Ω and N_P are the number of g_j and c_i in the union, respectively. And ω_1 and ω_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:

$$err = \frac{1}{N} \sum_i err_i \quad (3)$$

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 [21], which is the state of art clustering algorithm using time only. To make a more informative comparison, we also compare the simple algorithm that applies Mean-Shift to time only. Table 2 summarizes the evaluation results. It is clear that our method obtain the lowest error by all three metrics. Figure 5 shows the clustering errors for all 18 photo collections. Our clustering algorithm outperforms the other two methods for virtually every folder. By adding the GPS information, we achieved better event clustering, which lays reliable groundwork for event recognition.

Table 2. Evaluation of the accuracy of clustering algorithms.

Measures	Our Method	Time-only	Method in [25]
PRI [23]	0.030420	0.057404	0.097914
LCE [24]	0.000660	0.007702	0.001209
PEC	0.015055	0.089888	0.055476

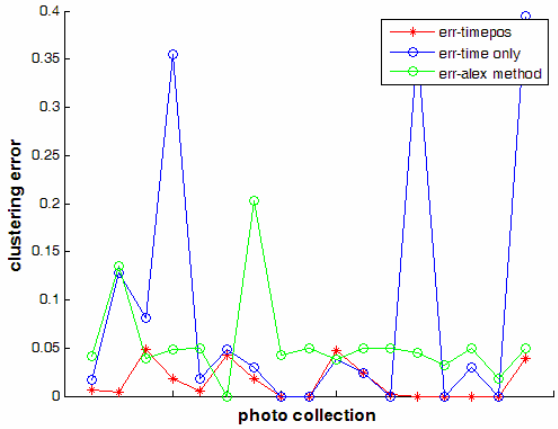

Figure 5: Comparison of different clustering algorithms. The horizontal axis shows different image folders, and the vertical axis denotes the clustering errors measured by PEC.

5.3. 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 [25] [26], we developed an SVM classifier for the 12 event classes. We *separately* collected 200 photos for each class, and randomly select 70% of these images for training the multi-class SVM classifier [27]. The remaining 30% of the photos are used for validation.

Given an event sub-collection C, our feature function for event e is

$$f_g^{\,e}(C) = (1, \sum_{x_i \in C} \frac{1}{1+\exp(-g^e(x_i))})^T \quad (3)$$

where $g^e(x_i)$ is the SVM score of photo x_i for each event. For our annotation work, $1 \le e \le 11$ stand for the 11 classes of events, and for the null event $g^0 \equiv 0$.

6. Joint Annotation of Events and Scenes

As shown in Fig. 6, the event and scene labels are strongly correlated. Some of them are often concurrent, e.g., the beachfun event and the coast scene. Others 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

$$R_c(s^k, e) = \begin{cases} \delta(s^k = 1), & \text{if } s^k \text{ and } e \text{ are concurrent} \\ -\delta(s^k = 1), & \text{if } s^k \text{ and } e \text{ are exclusive} \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

Note that formulation can handle both positive and negative correlation automatically. Fig. 6 shows these correlation pairs obtained from the training photos.

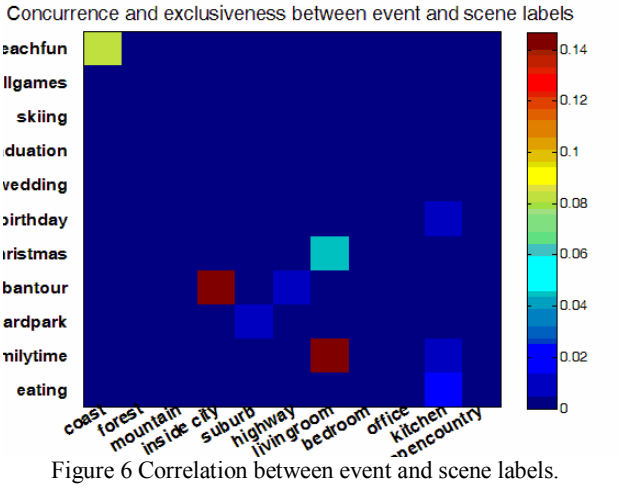

Figure 6 Correlation between event and scene labels.

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 likelihood function becomes

$$L = \alpha^e f_g^{\,e}(x) + \sum_k \sum_i \beta^k f_s^{\,k}(x_i, s_i^k) + \sum_k \sum_{i,j \in N_c} \lambda^k R_s^{\,k}(i,j) + \sum_k \sum_i \gamma^{k,e} R_c(s_i^{\,k}, e) + \log Z \quad (5)$$

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

Our complete 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 $\gamma^{k,e} = \gamma_1$ for $R_c(s_i^{\,k}, e) > 0$ and $\gamma^{k,e} = \gamma_2$ for $R_c(s_i^{\,k}, e) < 0$. Thus we only need two variables to represent the correlation of events and scenes.

To obtain the feature function f_s^k for single photos, we employ the statistical features from [2] and [28]. An SVM classifier is trained for the public scene dataset [14]. The feature function is

$$f_s^{\,k}(x_i, s_i^k) = \begin{cases} (1, \ h^k(x_i))^T, \text{if } s_i^k > 0 \\ (-1, \ 1 - h^k(x_i))^T, \ \text{if } s_i^k = 0 \end{cases} \quad (6)$$

where $h^k(x_i)$ is a sigmoid function used to shape the SVM score, $0 \le h^k(x_i) \le 1$. Then we let $\beta^k = (\hat{\beta}^k, 1)$, so

$$\beta^k f_s^{\,k} = \begin{cases} h^k(x_i) + \hat{\beta}^k, & \text{if } s_i^k > 0 \\ 1 - h^k(x_i) - \hat{\beta}^k, & \text{if } s_i^k = 0 \end{cases} \quad (7)$$

which is the simple form of the scene feature function.

Finally, we add the constraint that $\alpha^e = 1$ for all e. By observing (3) we can see that $f_g^{\,e}(C)$ is properly normalized, so removing the parameter α^e is reasonable.

After these simplifications, we can train the CRF models by minimizing L in (5). A conjugate-gradient method [29] is used to train the parameters. With the learned parameters, the hidden state s_i^k and e_c can be estimated by belief propagation.

7. Experimental Results

From the geotagged dataset, we randomly select 50% of 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 sub-collections.

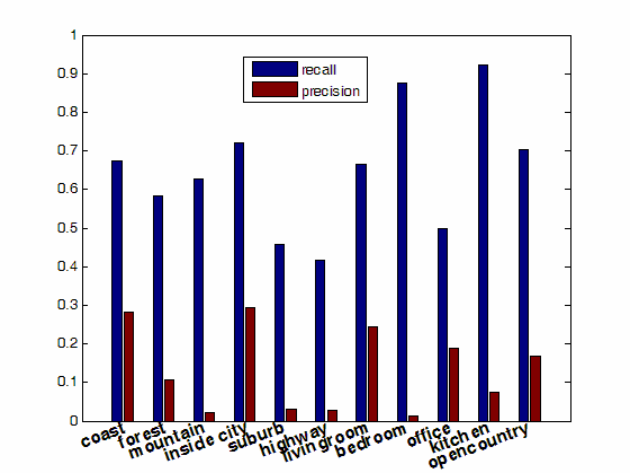

Figure 7 Precision-recall for scene annotation.

First, we show the accuracy of scene labeling. Since scene-level annotation is a multi-label 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 sub-collection level. We construct a confusion matrix for 11 events over sub-collections, 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 food on the table (unless we can detect the birthday cake explicitly).

	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**

Figure 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.

	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**

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

Event annotation becomes more difficult if we also consider the null event class. Fig. 9 shows the new confusion matrix for all the sub-collections, including those of the null class. Unfortunately, some null-event sub-collections 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. Fig. 10 and Fig. 11 show the precision and recall for scene and event annotation, respectively. They show that our hierarchical event-scene model with time and GPS improves significant both precision and recall in both cases. Although the model with time only is not as competitive as the full model, it is still much better than the single detectors.

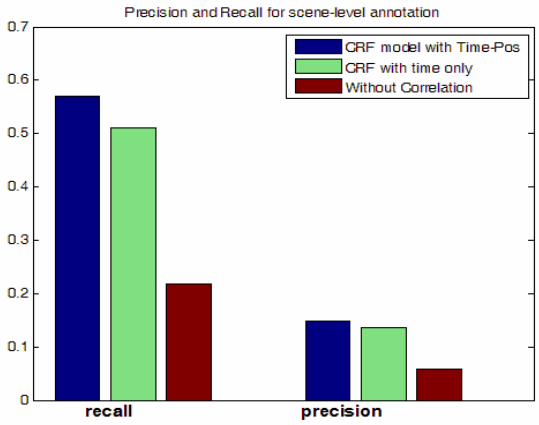

Figure 10 Comparing scene-level annotation accuracy by our CRF model using both time and GPS, with the model using time only, and with the single detectors (without modeling correlations).

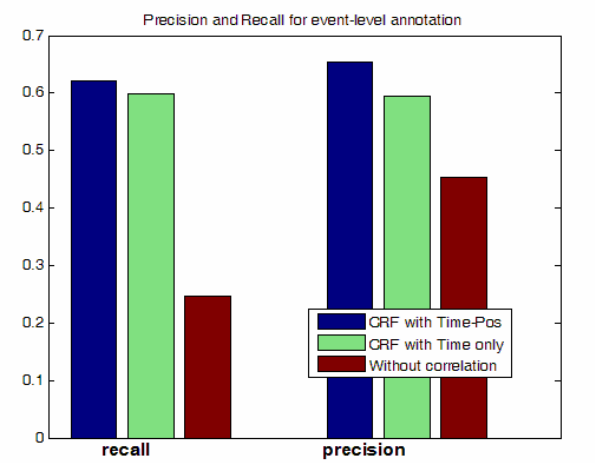

Figure 11 Comparing event annotation by the proposed model using both time and GPS, with the model using time only, and with the individual detectors without modeling correlations.

8. Conclusion and Future Work

This paper addresses the problem of annotating photo collections instead of single images. We built a medium-size collection of geotagged photos, and defined a compact ontology of events and scenes 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. The experiments show that our hierarchical model significantly improves annotation in both precision and recall.

Future directions include exploring (better) alternative baseline scene classifiers (e.g., [9][16]) and the physical location derived from the GPS coordinates, expanding the scene-event ontology, and finding a solution to reduce the relative high level of confusion between certain events.

References

[1] J.Z. Wang, J. Li, and G. Wiederhold. SIMPLIcity: Semantics-Sensitive Integrated Matching for Picture Libraries, *PAMI*, 23(9): 947-963, 2001.

[2] Y. Rui, T.S. Huang, and S.-F. Chang. Image retrieval: current techniques, promising directions, and open issues, *J. Visual Comm. and Image Representation*, 10:39–62, 1999.

[3] A. Yavlinsky and D. Heesch. An online system for gathering image similarity judgments. *Proc. ACM Multimedia* 2007.

[4] K Tieu and P. Viola. Boosting Image Retrieval, IJCV, 2004.

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

[6] J.-H. Lim, Q. Tian, and P. Mulhem. Home photo content modeling for personalized event-based retrieval. *IEEE Multimedia*, 10(4): 28 – 37, 2003.

[7] L. Zelnik-Manor and M. Irani. Event-based analysis of video, *CVPR*, 2001.

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

[9] L.-J. Li and L. Fei-Fei. What, where and who? Classifying event by scene and object recognition**.** *ICCV,* 2007.

[10] M. Boutell, X. Shen, J. Luo, and C. Brown. Learning multi-label semantic scene classification. *Pattern Recognition,* 37(9): 1757-1771, 2004.

[11] B. Russell, A. Torralba, K.P. Murphy, and W.T. Freeman. Labelme: A database and web-based tool for image annotation，*IJCV*，2007.

[12] M. Everingham, L. V. Gool, C. Williams, and A. Zisserman. The 2006 PASCAL visual object classes challenge. In *The PASCAL Visual Object Classes Challenge Workshop*, 2006.

[13] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. *CVPR Workshop on Generative-Model Based Vision*. 2004.

[14] L. Fei-Fei and P. Perona. A Bayesian hierarchical model for learning natural scene categories, *CVPR*, 2005.

[15] P. Quelhas. Modeling scenes with local descriptors and latent aspects, *ICCV*, 2005.

[16] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, *CVPR*, 2005.

[17] K. Barnard, P. Duygulu, N. Freitas, D. Forsyth, D. Blei, and M. Jordan. Matching Words and Pictures, *Journal of Machine Learning Research*, 3:1107-1135. 2003.

[18] M.A. Johnson and R. Cipolla. Improved image annotation and labeling through multi-label boosting. *BMVC*, 2005.

[19] L. Liao, D. Fox, and H. Kautz. Location-based activity recognition, *NIPS*, 2005.

[20] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *ICML*, 2001.

[21] A. Loui and A. Savakis. Automated event clustering and quality screening of consumer pictures for digital albuming, *IEEE Trans. Multimedia*, 5(3): 390-402, 2003.

[22] D. Comaniciu and P. Meer. Mean shift: a robust approach toward feature space analysis, *PAMI*, 2002.

[23] C. Pantofaru and M. Hebert. A comparison of image segmentation algorithms. *Technical Report CMU-RI-TR-05-40, Carnegie Mellon University*, 2005.

[24] D. Tal and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. *ICCV*, 2:416–423, 2001.

[25] Y. Aytar, O.B. Orhan, and M. Shah. Improving Semantic Concept Detection and Retrieval Using Contextual Estimates, *ICME*, 2007.

[26] A. Amir, M. Berg, S.-F. Chang, W. Hsu, G. Iyengar, and C.Y. Lin. IBM research TRECVID-2003 video retrieval system, *NIST TRECVID*, 2003.

[27] T. Joachims. Making large scale SVM learning practical, *Advances in Kernel Methods - Support Vector Learning*, MIT-Press, 1999.

[28] W.Y. Ma and H. J. Zhang. "Benchmarking of image features for content-based retrieval", *Proc. Signals, Systems & Computers*, 1: 253-257, 1998.

[29] F. Sha and F. Pereira. Shallow Parsing with Conditional Random Fields, *Proc. of HLT-NAACL*， pp. 213-220, 2003.