Annotating Photo Collections by Label Propagation According to Multiple Similarity Cues

Liangliang Cao
Beckman Institute and
Coordinated Science Lab
Department of ECE, UIUC
cao4@ifp.uiuc.edu

Jiebo Luo
Kodak Research Laboratories,
Eastman Kodak Company
Rochester, USA
jiebo.luo@kodak.com

Thomas S. Huang
Beckman Institute and
Coordinated Science Lab
Department of ECE, UIUC
huang@ifp.uiuc.edu

ABSTRACT

This paper considers the emerging problem of annotating personal photo collections that are taken by digital cameras and may have been subsequently organized by customers. Unlike the images from the web searching engine or commercial image banks (e.g. the Corel database), the photos in the same personal collection are related to each other in time, location, and content. Advanced technologies can record the GPS coordinates for each photo, and thus provide a richer source of context to model and enforce the correlation between the photos in the same collection. Recognizing the well-known limitations ("semantic gap") of visual recognition algorithms, we exploit the correlation between the photos to enhance the annotation performance. In our approach, high-confidence annotation labels are first obtained for certain photos and then propagated to the remaining photos in the same collection, according to time, location, and visual proximity (or similarity). A novel generative probabilistic model is employed, which outperforms the pervious linear propagation scheme. Experimental results have shown the advantages of the proposed annotation scheme.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval models; I.4 [**Image Processing and Computer Vision**]: Feature Measurement, Image Representation

General Terms

Algorithms, Experimentation

Keywords

Photo Collection, Label Propagation, Photo Similarities, Timestamp, GPS, SIFT, Color Histogram

1. INTRODUCTION

In the recent years, the popularity of digital cameras has lead to a flourish of personal digital photos. For example, Flickr [1] and Picassa Web Album [2] host millions of new personal photos uploaded every month. Compared with professional image banks such as Corel [3], these personal photos constitute an overwhelming source of images requiring efficient management. Recognizing and annotating these photos are of both high commercial potentials and broad research interests.

The difficulties in annotating personal photos lie in two aspects. First, such photos are of highly varying qualities, because they were taken by different people with different photography skills in different conditions. In contrast, the images in the Corel dataset were taken by professionals and thus share similarly well-controlled exposure conditions, Second, personal photos are far more complex in terms of semantic meaning. While Corel images are categorized in well-defined object and scene classes, personal photos contain unconstrained content and often are records of people, places, and events. All these factors pose greater changes for annotation, search and retrieval tasks.

One distinct but often overlooked feature of personal photos is that they are usually organized into collections or albums by time, location, and events. Since the users always move their photos from the camera to a computer, the photos are inevitably separated into file folders according to different dates. When the users want to share the photos with their friends, a natural and also informative way is to group the photos by location and date. The photos within the same file folder are often closely correlated to each other, since they were likely to be taken at the same time, place or event. This characteristic does not hold for generic image datasets. The motivation of this work is to utilize the folder organization to improve the annotation of diverse personal photos, and we differentiate our framework as annotation of photo collections.

In essence, photo collections provide rich information beyond the sum of individual photos. We assume that the photos in the same collection are taken by the same person using the camera under similar capture conditions. Under such an assumption, if two consecutive photos share similar visual features, it is likely that they describe the same scene or event. This is a powerful context that would not exist for general photos, which can describe different semantic content even if they contain similar color of texture features. In other words, the "semantic gap" in image similarity matching is inherently limited with the same photo collection. Moreover, computing the similarity among all possible

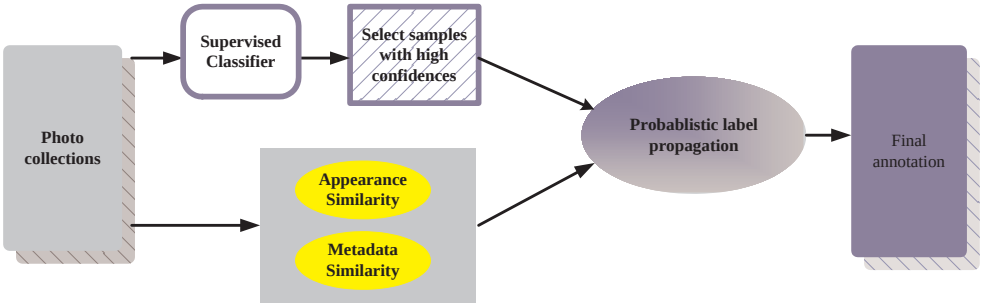


Figure 1: Overview of the proposed framework for annotating photo collections.

image pairs in a large database would be time consuming, while the computation for image pairs within a photo collection involves fewer photos that are already ordered in time and even location (when GPS information is available).

We can also model the photo similarity using metadata information such as timestamp and GPS tags. Every JPEG image file records the date and time when the photo was taken. 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). Since the photos in a collection are taken by the same camera, we can estimate whether labels of two photos are the same by the time and GPS information, either independent of or in conjunction with visual features. When the two photos are taken in a short time interval, it is unlikely that the scene or event labels change. Similarly, when two photos location does not change, the photos probably describe the same scene and event. Such metadata information was often overlooked in previous annotation work until recently [21]. Here we show that they are also useful for propagating labels in the same photo collection.

To test our method, we build a new database of photo collection by handing out cameras to different users over the period of 8 months. The database consists of 103 photo collections with different sizes (from 4 and 249 photos). To characterize the diverse semantics of personal photos, we labeled the database using an ontology of 12 events and 12 scenes. 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 mutually-exclusive 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 same class definitions from [4]: coast, open-country, forest, mountain, inside-city, suburb, highway, livingroom, bedroom, office, and kitchen. Note that here inside-city includes the three original classes of inside-city, street and tall-building, since our annotation task does not need to distinguish these three classes that are visually and semantically similar. Again, we also added a null scene class to handle the unspecified cases.

2. RELATED WORK

There has been a rich amount of research on image annotation. The most common approaches treat annotation as a supervised classification problem. These methods usually employ low level visual features [5] [10]. However, due to the diversity and complexity of visual content, it is difficult to obtain a satisfactory classifier for general images. To overcome the limitations of low level features, it is crucial to employ other sources of information for annotation.

One popular way to introduce useful new information in image retrieval is relevance feedback [6]. Relevance feedback requires the user to label some negative and positive examples from the initial retrieval results, and then the retrieval engine is updated in response to the use feedback. The updating algorithm can use simple re-weighting or query movement algorithms [6], or employ more sophisticate models [7] [8] [9] [25]. Relevance feedback is a powerful technique, but it relies on the user input which can become unpleasant if feedback is needed repeatedly.

This paper presents another way to enhance and extend the ability of supervised classifiers. Instead of employing the user interaction, we explore the pairwise similarity within a photo collection to develop an automatic annotation approach. Although there is no perfect classifier that can perfectly classify all the images, we can expect the classifier to be more accurate for those samples for which the classifier has high confidence. This was known as "rejection" and proved a good practice for SVM classifiers [24]. We treat those labels with high confidence as the initial "seed" labels, and propagate these labels to the remaining images in the same photo collection. We employ a probabilistic algorithm to accomplish the propagate task. The goal is to show that such a reject-and-propagate approach can significantly improve the recognition accuracy of photo annotation over running supervised classifiers over the images and accepting the individual results.

The overview of the proposed framework is shown in Fig. 1. The idea of our approach is partly motivated by the recent progress in semi-supervised learning algorithms [11] [12] [13]. This line of work designs a similarity matrix among data samples, and propagates initial labels to all the remaining samples using the similarity matrix. The propagation rules are usually iterative, for example, the iterative method in [13] is:

$$Y(t+1) = \alpha S \cdot Y(t) + (1-\alpha)Y^0 \quad (1)$$

where α is a parameter controlling the propagation rate,

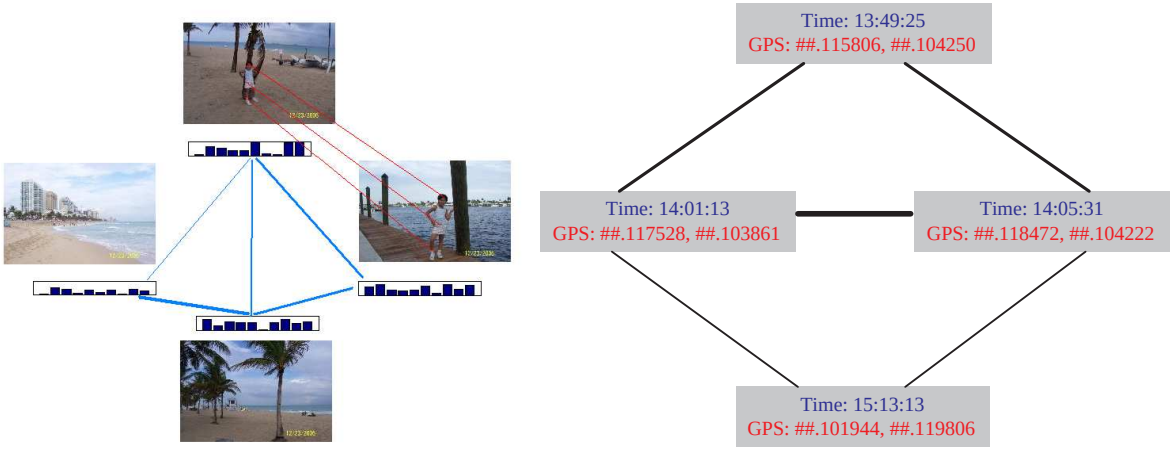


Figure 2: Modeling photo similarity using visual appearance features and metadata features. Left - similarity via appearance features: the red lines connect the similar SIFT patches corresponding to the same object, and the blue lines indicate high correlation between the color histograms. The thickness of the lines represents the degree of similarity. Right - similarity via metadata features: the blue text shows the timestamp tag of each image, while the red text shows the GPS tag. Note the actual coordinates of the GPS tags are removed to preserve privacy. Again the thickness of the lines denotes the degree of similarity.

and S is the Laplacian matrix computed from the similarity matrix. In each iteration, we obtain a score that estimates how likely samples are assigned with the given class, and Y^0 is the initial labels.

The approach of Eq.(1) is linear and simple to implement, and it has attracted much interest in image or video ranking [14] [15] [16] [17]. However, we chose not to employ Eq. (1) in our annotation task due to several reasons. First, using the updating process of Eq.(1), the initial labels may change in the propagation process. In other words, a sample with a positive initial label may be labeled as "negative" after the iterations. Second, there is no good theory to select the parameter α. One must try different values and select one empirically according to the experimental results. When there are multiple cues present to indicate image similarity, there would be more parameters to tune [18]. Finally, a simple linear model is not sufficient for modeling the probabilistic relationship in our label propagation task. Our goal is to design a probabilistic propagation model which does not requires numerous parameters of fusion weights and will not change the initial labels.

This paper is organized as follows: Section 3 describes how to model the pair-wise proximity or similarity between two images. We introduce our probabilistic propagation framework in Section 4, for the scenarios of both multi-label (scene) and multi-class (event) recognition. Section 5 compares the performance of our method with single classifiers, and presents examples of the different annotation approaches. The results validate the success of our propagation framework. We conclude this paper in Section 6.

3. SIMILARITIES WITHIN PHOTO COLLECTIONS

Most existing work [14] [15] [16] typically model the similarity between two images using low-level visual features. Due to the well-known gap between high-level semantics and low-level features, many images with different semantic content may share similar visual features, which suggest that it is beneficial to employ other sources of features to model the photo similarity. To model the photo correlation within the same collection, we employ both low level color features and scale invariant structure features (SIFT), together with the metadata features such as time and location. These metadata features are well suited for personal photo annotations, but not so for analyzing single photos. For example, for photos with close timestamps in the same personal photo collection, we can expect the photos to be semantically related to each other. However, if the two photos are taken by different people, most likely they are uncorrelated even if they were taken in the same time.

We employ two types of photo features to model pair-wise similarities between consecutive images. The first type is appearance features, including low level color features and SIFT features. The second type corresponds to metadata features, e.g., time and GPS. Fig.2 shows examples of these two types of similarity features, which will be discussed in detail below.

There are many forms of low level visual features, such as color, texture, and shape features. We do not use shape features in this study because the shape of the primary photo subject, people, changes with different actions and therefore is not invariant. We compute the color histogram in the LAB space for each photo, and use the correlation between two color histograms to model visual similarity.

Inspired by the recent advance in object recognition, we employ the SIFT features [19] together with the low level color features to model the visual similarity. SIFT is well suited for matching the same object in different images, and has shown effectiveness in image alignment and panoramic reconstruction [20]. Within the same photo collection, we

Table 1: 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.

expect that neighboring photos contain a common subject. Note that our task is more restricted than general object recognition, which requires a codebook or vocabulary obtained by extensive training processes. In contrast, our matching is much faster. Given two photos, we consider them as two sets of SIFT features. For each SIFT feature, we find two matching SIFT features in the other image, i.e. those with the highest and the second highest correlation. If the ratio of two correlation values is above a threshold (1.2 in our experiments), we decide that we find one pair of correspondent SIFT features. The more correspondent SIFT features are found, the more similar the two photos are.

We also employ metadata to model the similarity between two photos in a collection. We consider two kinds of metadata features, time and GPS. By the time features, the similarity between two photos is measured by the interval between the moments when the photos were taken. By the GPS features, the similarity is measured by the distance between the locations where the photos were taken. Such metadata information provides us with useful information for photo annotation. For example, if the user took photos near the beach, it is unlikely that he could move to inside the city within 5 minutes. Moreover, if the GPS tags show that the user moved only a few meters, the possibility that the user moved from mountain to indoors is extremely low. In short, if two consecutive photos are close in time and location, they tend to share the same labels.

4. PROBABILISTIC LABEL PROPAGATION

For the annotation task, we build a generative model for both modeling the image similarities and propagating the labels. The reason for developing a probabilistic model is threefold. First, it is nontrivial to combine diverse evidences measured by different means and represented by different metrics. For example, color similarities are represented by histogram correlations, and the subject similarity based on SIFT features is represented by integer numbers. Similarities by time and location are measured by minutes and meters, respectively. A probabilistic evidence fusion framework would allow all the information to be integrated in common terms of probabilities. Second, probabilistic models are capable of handling incomplete information gracefully. This properties are crucial especially for location features, since GPS tags sometimes can be missing due to the sensitivity limitation of the GPS receiver. Last but not the least, a probabilistic model can fully characterize the interacting effects from both positive and negative evidences, and estimate the true probability of each sample. Negative evidence is a unique feature of our framework, as now it becomes possible to propagate the fact that one image is not in a particular class to its neighbors. This is also useful in practice because the concept classifiers can provide both positive (that the image is of class A) and negative (that the image is not of class B). It is also possible for a user to provide both positive and negative initial labels, similar to relevance feedback where both positive and negative feedback are valuable. In contrast, the linear propagation method in [13] does not estimate the true probability and only utilizes positive evidence in the propagation.

4.1 Label Initialization

Following the standard practice in concept detection [22] [26] [27] [28], we developed a suite of SVM classifiers for both event and scene classes. Our scene classifiers are trained on the widely used 13-class UIUC dataset [4], which contains photos that are significantly different from the typical consumer photos used in our experiments. On the other hand, there are no public datasets to train our event classifiers, so we separately collected 200 photos for each event class by ourselves, and randomly selected 70% of these images to train the SVM classifiers [23]. Although such classifiers cannot classify every photo correctly, we can select those photos with high confidence scores and treat the labels generated by the SVM classifiers as the initialization for label propagation. Because we intend to utilize both positive and negative evidences, the labels with scores below the threshold of -1.0 are selected as negative initial evidence, and the labels with scores above the threshold of 0.2 are selected as positive initial evidence. These selected photos are also referred to as "seed" images.

4.2 Modeling Similarity using Multiple Features

Given two photos i and j, we denote the label variables as y_i and y_j. To model the similarity between photo i and

j, we consider photo features x_i, x_j, and their similarity is measured by $d_{ij} = Similarity(x_i, x_j)$.

To measure whether two images are correlated or not, we introduce a new variable for modeling the correlation between image i and j, which is defined as

$$s_{ij} = \begin{cases} 1 & if\ y_i = y_j \\ 0 & if\ y_i \neq y_j \end{cases}. \quad (2)$$

Note that here we do not model the photo label y directly. Instead, we use the appearance and metadata features to model s_{ij}, which characterizes whether the two photo labels are similar. This is a significant difference between our method and previous supervised methods [10] [21].

Now we can model the probability of image correlation by $P(s_{ij}|d_{ij})$. Using the Bayesian formula, we have

$$P(s_{ij} = \delta|d_{ij}) = \frac{P(d_{ij}|s_{ij} = \delta)P(s_{ij} = \delta)}{\sum_{\delta_1=\{0,1\}} P(d_{ij}|s_{ij} = \delta_1)P(s_{ij} = \delta_1)}, \quad (3)$$

where δ is Kronecker's delta function.

The probabilistic formulation of Eq. (3) can be easily learned from the data. Another benefit of Eq. (3) is that it provides a good framework to introduce multiple features. When each image is associated with multiple visual and metadata features, we denote them by $x_i = \{x_i^k\}$ and $x_j = \{x_j^k\}$, where $1 \leq k \leq K$ denotes the feature type. Now the similarity d_{ij} is represented by $d_{ij} = d_{ij}^k$, and each d_{ij}^k measures the similarity between x_i^k and x_j^k. Now we can model the conditional similarity as

$$P(d_{i,j}|s_{ij}) = \prod_{k=1}^{K} P(d_{ij}^k|s_{i,j}, d_{ij}^1, ..., d_{ij}^{k-1}) \quad (4)$$

Out of the four features, the metadata features (time and location) are independent with respect to the visual features. The visual features, SIFT and Color Histogram, describe the different aspect of photo content. Our previous study [29] has shown that we can treat such visual features as conditionally independent. So we can simplify $P(d_{i,j}|s_{ij})$ as

$$P(d_{i,j}|s_{ij}) = \prod_{k} P(d_{ij}^k|s_{i,j}). \quad (5)$$

By combining Eqs. (3) and (5) we determine the correlation probability $P(s_{ij}|d_{i,j})$.

Our probabilistic model can handle the partially missing GPS without difficulty. Suppose one feature k^0 is missing, then we can treat Eq.(3) as

$$P(s_{ij} = \delta|d_{ij}) = \frac{\prod_{k\neq k^0} P(d_{ij}^k|s_{ij} = \delta)P(s_{ij} = \delta)}{\sum_{\delta_1=\{0,1\}} \prod_{k\neq k^0} P(d_{ij}^k|s_{ij} = \delta_1)P(s_{ij} = \delta_1)}$$

5. LABEL PROPAGATION MODEL

Next we discuss our label propagation model. To make the representation simpler to follow, we begin our discussion with a two-class problem. For each task, we aim to infer the label y for each image, where $y_i = 1$ means an image should be assigned to the label, and $y_i = 0$ means not. The probability of image labels satisfies the constraint

$$P(y_i = 1) + P(y_i = 0) = 1.$$

Using the initialization method in Section 4.1, we obtain a set L of labeled images, where $P(y_i = 1) = 1$ or $P(y_i = 0) = 1$ if $i \in L$. The other images belong to the set of unlabeled images U, where $P(y_i = 1) = P(y_i = 0) = 0.5$ for $i \in U$.

Based on the discussion in Section 4.2, we can estimate the probability of label propagation using the correlation probability $P(s_{ij}|d_{ij})$

$$P(y_i \rightarrow y_j) = \lambda_i \cdot P(s_{ij} = 1|d_{ij}), \quad (6)$$

where λ_i is a normalization constant satisfying

$$\lambda_i = 1/\sum_{k\neq i} P(s_{ik} = 1|d_{ik})$$

In our scheme, each unlabeled photo $j \in U$ updates its probability by considering label probability of the other photos which are similar by any measure. There are two possible labels, $y = 0$ or $y = 1$, and we can compute them separately.

$$P_j^+ = \sum_{i\neq j} P(y_i = 1)P(y_i \rightarrow y_j)$$
$$P_j^- = \sum_{i\neq j} P(y_i = 0)P(y_i \rightarrow y_j) \quad (7)$$

Note that the updated probability does not satisfy the constraint of $P(y_i = 1) + P(y_i = 0) = 1$. We need to normalize them after each updating stage.

$$P(y_j = 1) \leftarrow \frac{P_j^+}{P_j^- + P_j^+}$$
$$P(y_j = 0) \leftarrow \frac{P_j^-}{P_j^- + P_j^+} \quad (8)$$

Since we have high confidence in the labeled set L, we only update the probability for $j \in U$. In each iteration, we update the probability for every unlabeled photo using (7) and (8). This procedure continues until it converges or reaches a maximum number of iterations (100 in our experiment).

Our propagation algorithm is summarized as follows:

Procedure of our algorithm

Input: Pairwise image similarity d_{ij}. Initialized photo set L with the labels $y_i = 1$ or $y_i = 0$, for $i \in L$.

Output: The estimated labels of photos in the unlabeled set U.

Procedure:

1. Estimate the correlation probability $P(s_{ij}|d_{i,j})$ according to eqs. (3) and (5).
2. Obtain propagation probability $P(y_i \rightarrow y_j)$ by normalizing $P(s_{ij}|d_{ij})$ using eq. (6)
3. Initialize $P(y_i = 1) = 1$ or $P(y_i = 0) = 1$ if $i \in L$. Initialize $P(y_j = 1) = P(y_j = 0) = 0.5$ for $j \in U$.

4. For each unlabeled photo $j \in U$, update $P(y_j)$ using eqs. (7) and (8).
5. Repeat step 4 until it converges or reaches a maximum number of iterations.
6. Assign $y_j = 1$ if $P(y_j = 1) > 0.5$. Otherwise let $y_j = 0$.

Our approach is different from the previous graph-based ranking algorithms [13] [14] [15] [16] in two aspects. On one hand, we adopt a full Bayesian framework which gracefully handles the feature fusion and missing GPS problems without the trouble of tuning extra parameters. On the other hand, our model provides explicit probabilistic estimations, while the linear propagation method in [13] only provides ranking scores without probabilistic meaning and may change the initial labels in the propagation process.

Our approach can be easily generalized to a multi-label problem by treating it as multiple two-class problems. If we do not allow more than one label for each image, we simply select the one with the largest probability of $P(y_j = 1)$.

6. EXPERIMENTAL RESULTS

We test our algorithm using the photo dataset described earlier. The dataset contains two types of annotation for each photo, events and scenes. The definition of events and scenes are described in Section 1. Note that we allow the labels of "NULL event" and "NULL scene".

From the dataset, we randomly selected 50% of the folders to learn the distributions of photo similarity for both event and scene correlation. Four types of features are employed for modeling photo similarities: SIFT, color histogram, time, and GPS (may be missing for some photos).

Fig. 3 compares the performances of our probabilistic propagation framework with the single SVM classifier over event recognition. It can be seen that our propagation model significantly outperforms the baseline of SVM classifiers. For the task of event annotation, our model obtains better recalls for all the 11 events. Moreover, our precision rates are similar to the baseline SVM classifiers, except for graduation, Christmas, family time and eating events where the propagation achieved considerably higher precision. With the help of our probabilistic model, the average of precisions for 11 events increases from 0.381 to 0.411, while the average of recalls increases from 0.331 to 0.506. In other words, our probabilistic propagation significantly improves the recall of event annotation by 52.9% and the precision by 7.9%.

Fig. 4 shows the confusion matrix for 11 events, with and without the Null event. We can see that the introduction of the Null event makes our annotation task more difficult. If we know that each photo belongs to one and only one of the 11 events, the event recognition error is significantly smaller. The main errors occur when distinguishing the event pairs that share much visual similarity, for example, wedding is often confused with graduation when the graduates happen to wear white gown, and birthday can be confused with Christmas because both allow children dressing up and colorful decorations are popular in both cases.

The task of scene annotation is a more challenging problem than event annotation, due to a cross-domain issue We used the standard scene dataset in [4] to train scene classifiers but it is somewhat different from the typical customer photos. The standard scene dataset hardly contains human subjects, which is one of the most important figures in personal photos. Scene photos in [4] are often taken from the perspectives of long or moderate distance, while the customer photos are often taken within 5-7 feet. Consequently, the SVM classifier itself does not perform scene annotation as well as in event annotation. However, Fig. 5 shows that our probabilistic propagation approach still significantly improves both precision and recall. With the help of probabilistic propagation, the average of precisions for 11 scenes increases from 0.097 to 0.106, while average of the recalls rise from 0.206 to 0.358. In other words, our model improves the recall of scene annotation by 73.8% and the precision by 9.3%.

To better understand the effects of different features for modeling photo similarity, we also conduct experiments which employ each of the four types of features for the propagation task. Table 2 shows the results of these experiments. Of the four similarity features we used for label propagation, the color histogram is the most useful for event annotation while the time feature obtains the best scene annotation results. This shows that the scene labels are more coherent in time than events. In both cases, the fusion of all feature is better than or as good as the best single feature.

To show the benefit of our propagation algorithm, we compare our method with the linear propagation algorithm employed in [13] [14] [15]. Because the previous non-probabilistic models cannot directly handle multiple similarity metrics, we use our Bayesian model in Eq.(6) to generate a combined similarity matrix and feed it into the two propagation algorithms. We compare our propagation method in Eqs. (7) and (8) with the linear propagation method using different parameters, and the results are shown in Table 3, where *E-Prec.* and *E-Recall* stand for the precision and recall for event annotation, and *S-Prec.* and *S-Recall* stand for the precision and recall for scene annotation. We can see that our probabilistic propagation method outperforms linear propagation method no matter what parameters are selected. This shows that our model is more effective than the linear propagation and does not bear the burden of tuning the parameters.

Fig. 6 compares the annotation results by the baseline SVM algorithm and our propagation method. We select three different collections and show the annotation of 10 photos (subsampled for display purpose only) in each collection. It can be seen that our propagation method outperforms the baseline method for all these three collections and provides better descriptions of both scenes and events.

7. CONCLUSION

In this work, we consider the image annotation problem within the context of personal photo collections. Rather than using trained classifiers to label each of the photos directly, we propose to use a reject-and-propagate approach where only the photos with high confidence scores are assigned labels initially and label propagation is used to assign labels to the photos rejected for their low classification

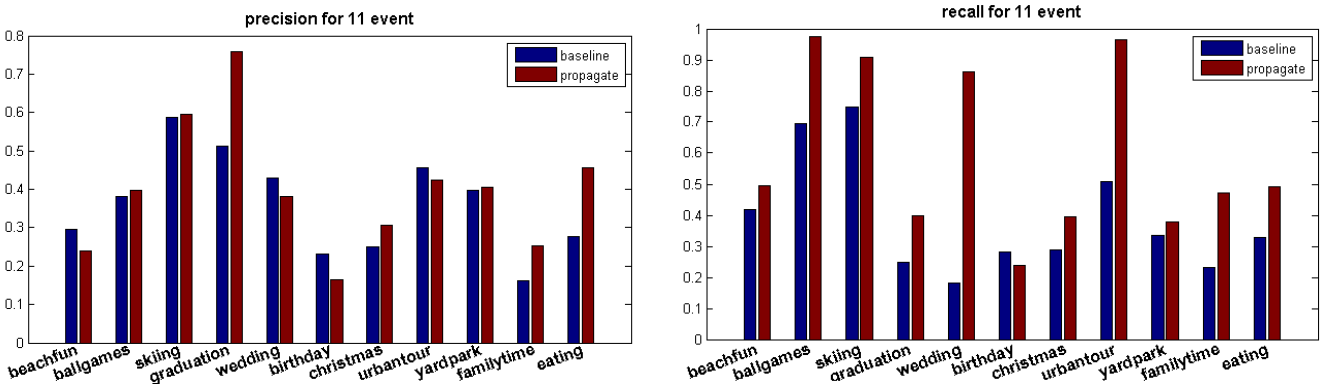


Figure 3: Our propagation model significantly improves the event annotation. The blue bars show the precision or recall rates using the baseline SVM classifiers. The red bars show the precision or recall rates using our label propagation framework.

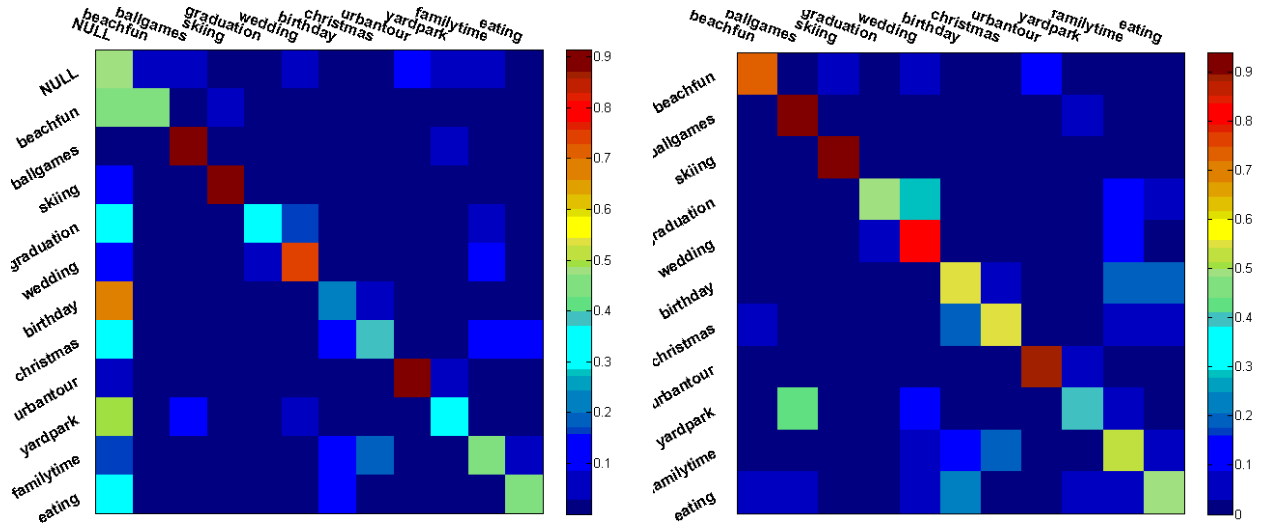


Figure 4: Left: Confusion matrix when considering null events. Right: Confusion matrix without considering null events.

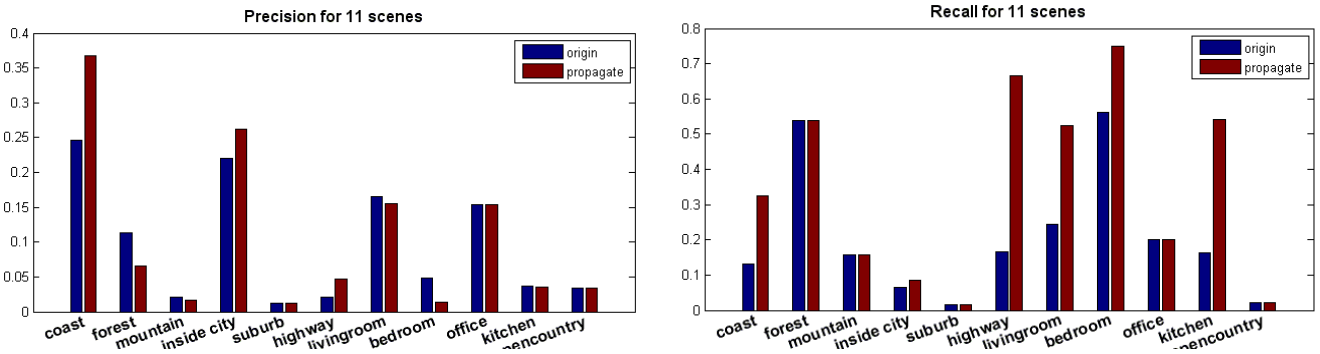


Figure 5: Our propagation model significantly improves the scene annotation. The blue bars show the precision or recall rates using the baseline SVM classifiers. The red bars show the precision or recall rates using our label propagation framework.

Table 2: Comparing propagation based on individual similarity features and fusion of multiple features.

Feature Name	*Event prec.*	*Event recall*	*Scene prec.*	*Scene recall*
Baseline (SVM)	0.381	0.331	0.097	0.206
SIFT	0.404	0.482	0.098	0.285
Color Histogram	*0.405*	*0.518*	0.102	0.321
Time	0.404	0.477	*0.106*	*0.350*
GPS	0.447	0.372	0.103	0.233
Fusion of 4 features	**0.411**	**0.506**	**0.106**	**0.358**

scores. This is a way to address the well-known limitations of current visual recognition algorithms, by exploiting the correlation between the photos to enhance the overall annotation performance. The label propagation is guided by similarity metrics in terms of time, location, and visual appearance. A novel probabilistic model is employed, which outperforms the linear propagation scheme. This label propagation framework also lends itself to semi-automatic annotation where an operator only needs to provide the initial high-confidence labels on a subset of the photos and the algorithm propagates the manual labels to the remainder of the photo collection.

Table 3: Comparing propagation methods

	E-Prec.	*E-Recall*	*S-Prec.*	*S-Recall*
Bayes propagation	**0.411**	**0.506**	**0.106**	**0.358**
Linear ($\alpha = 0.7$)	0.387	0.516	0.104	0.274
Linear ($\alpha = 0.5$)	0.377	0.506	0.100	0.305
Linear ($\alpha = 0.3$)	0.377	0.504	0.100	0.306

8. REFERENCES

[1] Flickr, http://www.flickr.com

[2] Picasa Web Album, http://picasaweb.google.com

[3] Corel stock photo database, http://www.corel.com

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

[5] A. K. Jain and A. Vailaya, Image retrieval using color and shape, Pattern Recognition, 29 (8), 1233-1244, 1996.

[6] Y. Rui, T.S. Huang, S Mehrotra, Content-based image retrieval with relevance feedback in MARS, ICIP, 1997.

[7] A. Blum, T. Mitchell, Combining labeled and unlabeled data with co-training, COLT, 1998.

[8] S. Tong, E. Chang, Support vector machine active learning for image retrieval, ACM Conf. Multimedia, 2001.

[9] S. Hoi, M.R. Lyu, A semi-supervised active learning framework for image retrieval, CVPR, 2005.

[10] G. Carneiro, A. B. Chan, P. J. Moreno, N. Vasconcelos, Supervised Learning of Semantic Classes for Image Annotation and Retrieval, IEEE Trans. PAMI, 2007.

[11] X. Zhu, Z. Ghahramani, and J. Lafferty, Semi-supervised learning using Gaussian fields and harmonic functions, ICML, 2003.

[12] F. Wang, C. Zhang, H.C. Shen, and J. Wang, Semi-Supervised Classification Using Linear Neighborhood Propagation, CVPR, 2006.

[13] D. Zhou , O. Bousquet, T.N. Lal, J. Weston and B. Scholkopf, Learning with Local and Global Consistency. NIPS, 2004.

[14] J. He, M. Li, H.-J. Zhang, H. Tong, C. Zhang , Manifold-ranking based image retrieval, ACM Conf. Multimedia, 2004.

[15] J. Liu, M Li, W.Y. Ma, Q. Liu, H. Lu, An adaptive graph model for automatic image annotation, ACM workshop on Multimedia Information Retrieval, 2006.

[16] H. Tong, J. He, M. Li, C Zhang, W-Y. Ma, Graph based multi-modality learning, ACM Conf. Multimedia, 2005.

[17] J. Liu, W. Lai, X-S Hua, Y Huang, and S. Li. Video Search Re-Ranking via Multi-Graph Propagation. ACM Conf. Multimedia, 2007.

[18] D. Zhou and C. Burges, Spectral Clustering and Transductive Learning with Multiple Views. ICML, 2007.

[19] D. Lowe, Distinctive Image Features from Scale-Invariant Keypoints, 60(2): 91-110, IJCV, 2004.

[20] Matthew Brown and David G. Lowe, Automatic panoramic image stitching using invariant features, IJCV, 74(1): 59-73, 2007.

[21] M. Boutell, and J. Luo. Beyond pixels: Exploiting camera metadata for photo classification. Pattern Recognition 38(6): 935-946, 2005.

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

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

[24] Y. Wang, H. Zhang. Detecting image orientation based on low-level visual content. Computer Vision and Image Understanding 93(3): 328-346, 2004.

[25] D. Tao, X. Tang, X. Li, and X. Wu. Asymmetric Bagging and Random Subspace for Support Vector Machines-based Relevance Feedback in Image Retrieval, IEEE Trans. PAMI, 28(7): 1088-1099, 2006.

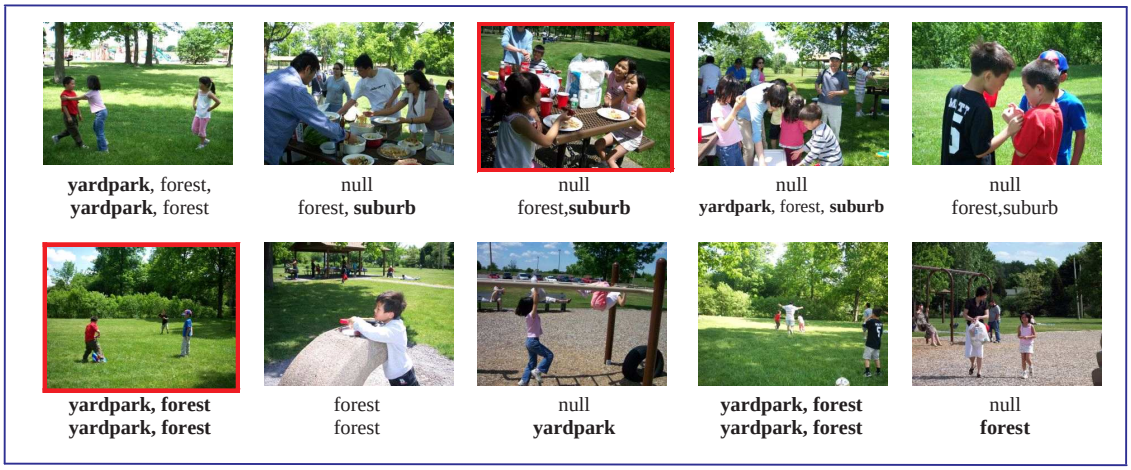
[26] P. Quelhas, et. al., Modeling Scenes with Local Descriptors and Latent Aspects, ICCV, 2005.

[27] S.F. Chang, et. al., Columbia Univ. TRECVID-2005 Video Search and High-Level Feature Extraction, Proc. TREC Video Retrieval Evaluation, 2005.

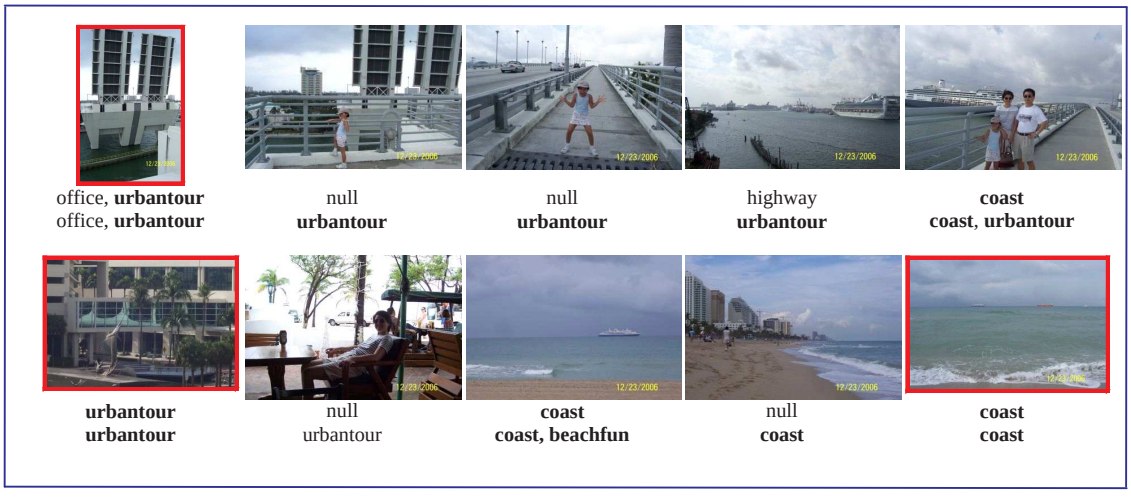
[28] M. Campbell, et. al., IBM Research TRECVID-2006 Video Retrieval System, Proc. TREC Video Retrieval Evaluation, 2006.

[29] M. Boutell and J. Luo, Beyond Pixels: Exploiting camera metadata for photo classification, Pattern Recognition 38(6): 935-946, 2005.

(a) Collection 1



(b) Collection 2



(c) Collection 3

Figure 6: Comparison of annotations using the baseline SVM classifier and our propagation approach. We show examples from three photo collections, each of which are bounded by a color rectangle. For each photo collection, rows 1 and 3 display photos, rows 2 and 4 show the annotation results by the baseline SVM, and rows 3 and 5 show the annotation by our propagation model. The images shown for each collection are in chronological order but not necessarily consecutive because of temporal subsampling (for display only). The "seed" images automatically selected by the rejection criteria are indicated by a red bounding box. The correct annotations are indicated by bold letters.