

Event Recognition from Photo Collections via PageRank

Naveed Imran

Computer Vision Lab
University of Central Florida
snaveedimran@gmail.com

Jingen Liu

Computer Vision Lab
University of Central Florida
liujg@cs.ucf.edu

Jiebo Luo

Kodak Research Labs
Rochester, NY, USA
jiebo.luo@kodak.com

Mubarak Shah

Computer Vision Lab
University of Central Florida
shah@cs.ucf.edu

ABSTRACT

We propose a method of mining most informative features for the event recognition from photo collections. Our goal is to classify different event categories based on the visual content of a group of photos that constitute the event. Such photo groups are typical in a personal photo collection of different events. Visual features are extracted from the images, yet the features from individual images are often noisy and not all of them represent the distinguishing characteristics of an event. We employ the PageRank technique to mine the most informative features from the images that belong to the same event. Subsequently, we classify different event categories using the multiple images of the same event because we argue that they are more informative about the content of an event rather than any single image. We compare our proposed approach with the standard bag of features method (BOF) and observe considerable improvements in recognition accuracy.

Categories and Subject Descriptors

H.2.8 [Database Management]: Database Applications – data mining, image databases

H.3.3 [Information Search and Retrieval]: Retrieval models

General Terms

Algorithms, Experimentation

Keywords

Event category recognition, PageRank, CBIR

1. INTRODUCTION

Automatic event recognition is an active research area due to its broad applications including browsing through a database of personal photo collections and content based image/video indexing & retrieval. Traditionally, in computer vision community, the event recognition problem has been associated with recognizing certain activities in video sequences e.g. opening a door, picking up an object etc. However, in daily life, significant occurrences such as a social gathering or a trip are also known as events.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

MM'09, October 19–24, 2009, Beijing, China.

Copyright 2009 ACM 978-1-60558-608-3/09/10...\$10.00.

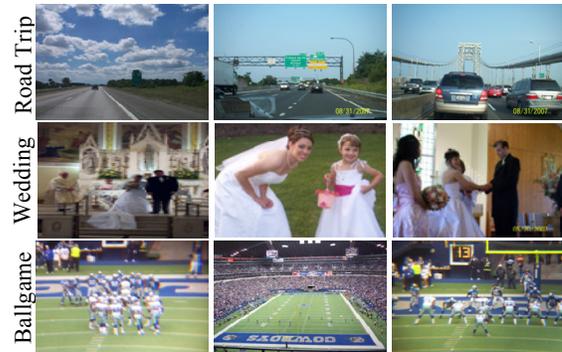


Figure 1: Image examples of three events selected from our 10-event-categories dataset of 3453 photos in total. Detailed description is in section 3.

People frequently take pictures to keep the memory of events in their lives and save those photos in the computer or web albums. Figure 1 exemplifies some photos from a typical photo album. The first row shows photos taken during a road trip while the photos taken at a wedding event are shown at second row. The last row photos are those taken at a sports event. These types of images are abundant and a personal collection may contain thousands of such photos. Typically, a user puts these pictures into one new folder upon returning from a fun trip or event. Automatic event category recognition of these photos is helpful for management of the ever increasing personal photo collections.

There is a considerable amount of research work in the area of event category recognition in which different approaches have been employed for varied objectives. We divide the research directions for event recognition into two categories: using motion features and using only static features. Irani et al. use motion features for the event based analysis of the video in [1]. They define events as long term temporal objects and propose some statistical measure for event clustering. Xu et al.'s work is related to the problem of visual event recognition in television broadcast news videos [2]. They use bag of features and kernel based method for measuring video clip similarity. In Ebadollahi et al.'s work, video clips containing different events are classified by employing temporal dynamics [3]. In TRECVID community, they usually use image features from key-frames for high level features task [2, 4]. Liu et al. exploit both motion features and static features for recognizing events from unconstrained YouTube videos [5]. Their dataset contains amateur videos of bicycling, horse riding and basketball playing, and so on. Another interesting work is by Jiang et al. who investigate automatic detection of semantic events in user's mixed image and video collections [6]. A seman-

tic event detection approach is used in this work to exploit an event-level Bag-of-Features (BOF) representation to model typical events. At the other end of the spectrum, Torralba et al. propose a holistic representation of the image to describe nature of the scene [7].

A photo collection stands as a feasible and appealing medium between the extremes of video clips and single photos for categorizing the nature of a scene. Luo et al. are among the first to use collections of photos instead of using a single photo for event level annotation [10]. They also explore the GPS and time information for semantic annotation. Another work by the same group deals with the event recognition using extracted features from multiple images in a personal photo collection [11]. In a similar spirit, Wu et al. use web images to learn concept templates in order to query personal photo collections [9]. Similarly, the goal of this study is to efficiently represent the visual characteristics of a group of photos for the purpose of automatic event category recognition. However, we are interested in extracting the most informative features from a group of images related to the same event for event recognition in personal photo collections. In this work, we define an event as a scene that may (often) or may not contain a human activity, yet depicts a type of personal photo album category. We define an instance of the event as a set of images related to one topic. For example, Christmas Day is the name of a folder of photos taken on this particular day, and the folder of images is an instance of the Christmas event category. A learned algorithm should be able to predict the category of the event (i.e. Christmas) when we feed an instance (i.e., a set of photos of this particular event) as a query.

Although the human vision system can recognize many types of events by just looking at one picture, the recognition ability is, however, improved when looking at a set of pictures belonging to the same event. One image may not be much informative about the nature of an event. For example, by just looking at first picture in Figure 2, we cannot say about which category it belongs to; however by looking at other pictures from the same instance of this event the category type ‘beach’ is easily revealed to us. Therefore, we advocate the use of multiple images to recognize the class of an instance of event.

The extraction of good features from photos is crucial to event recognition. The bag of features (BOF) model can be used to capture the statistical information of the visual features. The use of BOF method in computer vision is inspired by the success of bag of words (BOW) method in text document analysis. Here each visual feature corresponds to a word in the document, so an image, like a document, can be represented by a histogram of visual words. BOF has been widely employed in object, scene, and action recognition [12] due to its simplicity and good performance.

An event is characterized by some features which are consistent among all the photos in that instance. The key is to discover those consistent features. We think those as the most informative ones and choose to mine these features using the PageRank (PR) technique. PR is an analysis algorithm to analyze the interaction among the features, by assigning a ranking score to each feature as its relative significance in the feature network. For a given instance of an event, we build a large directed graph of features called *feature similarity graph* (FSG). Here, a vertex denotes a feature, and an edge represents a match with another feature. If a feature is consistently matched with many other features, we con-

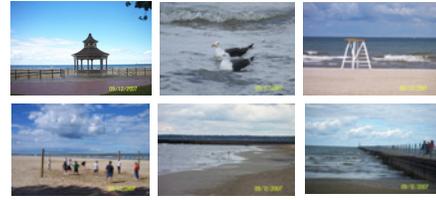


Figure 2: Some photos from beach category.

sider it more significant than others. In this way, PR ranks the more informative features higher by giving them higher scores.

The PR technique has been successfully utilized in Google search engine. Jing et al. apply PR to large scale image search problem [13]. The approach of [8] is learning models of object categories in an unsupervised manner by using PR. Recently, PR has been employed to tackle the problem of event recognition in extracting consistent features from unconstrained realistic videos without the need for tracking [5]. In our problem, some images belonging to the same event category may contain distracting background and foreground objects, thus the characteristic information of an event is mingled with distracting features. Inspired by the successes of [5, 8, 13], we propose to use PR technique to discover the relatively important features for a different problem domain that deals with collections of images.

2. OUR PROPOSED APPROACH

We present a method for event recognition based on integrating BOF and PR for mining most informative visual features. Figure 3 illustrates the flow diagram of our proposed approach. The visual features are extracted from multiple photos, and visual similarity graph is built by matching features across the photos for a given instance of an event. PageRank is used for computing the ranking score for all features to find the most representative vertices in feature network. The BOF method is generally a two-stage process, consisting of vocabulary construction followed by category model learning (e.g. with SVM). For vocabulary construction, we perform k-means clustering and classification is accomplished by a Support Vector Machine.

2.1 Visual Feature Extraction

The SIFT descriptor proposed by Lowe is popular descriptor in terms of scale invariant features for describing local image structures. SIFT features are invariant to image scale and rotation, and provide robust matching across a substantial range of affine distortion, change in 3-D viewpoint, addition of noise, and change in illumination [14]. These features have been successfully applied to object and scene recognition. For each image in an event, we use the SIFT descriptors as one of its features.

We observe that color is also an important feature to distinguish among different events. Therefore to increase the discriminative power, we also use color descriptors. Our color features are based on the mean values of RGB color channels for image patches. The combination of both descriptors has outperformed using only one feature type since they are potentially complementary.

To characterize the images, we choose the color and SIFT features to represent the image with an unordered set of descriptors. An image is partitioned into fixed number of overlapping patches and each patch is represented by a 12 dimensional RGB color descriptor and a 32 dimensional SIFT descriptor, similar to [12].

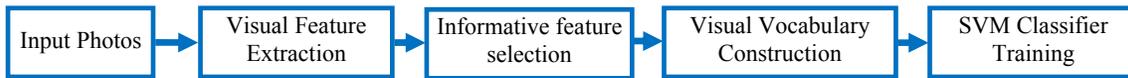


Figure 3: The flow diagram of the training phase of our proposed method.

2.2 Discriminative Feature Selection

The extracted features from the images are dense yet noisy. Only a subset of them characterizes the nature of an event. We are interested in finding only those discriminative features. The extraction process of most informative features consists of two major steps: Feature similarity graph construction by image matching and visual feature ranking by PageRank. We describe the procedure briefly as follows.

2.2.1 Construction of feature similarity graph

The feature similarity graph (FSG) is a directed graph $G = (V, E, W)$, where V is the vertex set (the visual feature set), E is the edge set, and W is the associated adjacency matrix with weight representing the degree of match between the linked features [5]. In order to discover the discriminative features of an event, image matching is performed on all the images within the same instance of an event. For constructing the FSG, we first randomly sample some of features from each image. For all pairs of images within one event, we retrieve n matched candidate feature pairs estimated by comparing the Euclidean distance between a pair of features represented by the SIFT and color descriptors. In this way, links in FSG are established by finding matches between features in different images.

There are different ways for performing this matching operation. One is applying spectral matching to each pair of images. This approach combines matching based on local data with geometric consistency constraints by finding a combination of correspondences that are globally most consistent, based on pair wise relations between features [15]. Another way of matching is simple and we have used it in our work. In this method, the best candidate match for each feature is found by identifying its nearest neighbor in the database of features from the training images. The nearest neighbor is defined as the data point with minimum Euclidean distance for the descriptor vector. A more effective measure is obtained by comparing the distance of the closest neighbor to that of the second neighbor. This measure performs well because correct matches need to have the closest neighbor significantly closer than the closest incorrect match to achieve reliable matching [14].

Next, a weighted adjacency matrix P ($n \times n$) is constructed where a node represents a pair of matched features (i, j) , and edge weights represent the Euclidean distance between features. We normalize this weighted adjacency matrix so that the distance measures correspond to the similarity measures between features. In this way, we obtain an $n \times n$ sparse matrix W .

2.2.2 Feature ranking via PageRank

Given the FSG with vertices and a set of weighted edges, we are interested in measuring the importance of each vertex. The intuition is that a vertex very similar to an “important informative vertex” should rank higher than others that are further away. The important informative vertices are features corresponding to the consistent characteristics of an event. We treat the FSG the same way as a graph of all linked web pages. Each vertex is similar to a web page and all the edge weights associated with a vertex can be considered as the votes cast by the linked vertices. For consistent features of an event, we should obtain consistent matches and

the distracting features should generally be outliers. To rank these features according to their importance, we apply PageRank.

PageRank computes a rank vector to estimate the importance of all the web pages on the Web by analyzing the hyperlinks connecting web documents [13]. In our case, suppose Pr is a $l \times n$ PageRank vector with each entry corresponding to the PR value of the feature, we can solve the problem using the following equation:

$$Pr = \alpha * Pr * W + (\alpha * Pr * b + 1 - \alpha) * v$$

where α is the scaling factor, b is an indicator vector identifying the vertices with zero out-degree, W is the weights matrix, and v is an $n \times 1$ transport vector with uniform probability distribution over the vertices. The initial PageRank value for each vertex is $1/n$. For each instance of an event, we compute its PageRank vector Pr . Based on the rank of Pr values, we select the top μ features as the informative ones. The threshold for selecting the top features is determined empirically. Depending upon the diversity of the characteristic features of an event, we may need to select more number of features as the informative ones. However if the typical features of an event are consistent, only a small percentage of the features will be sufficient to represent that event.

2.3 Event Representation

We represent an instance of the event as a super bag of features because there are multiple images. We can either select all of the features or some of them to generate the vocabulary. We randomly choose some of the features and perform K-means clustering to form the codebook. The size for the codebook is 500. In general, a larger visual vocabulary performs better, but over-specific visual words may eventually over-fit the data. Empirically, we choose this number as optimal between the over fitting and generalization extremes. We build separate codebooks for color and texture.

Next, we perform vector quantization and form the histogram of words. Thus each image is represented by a bag of words or by histograms of code words. By taking the mean of all the normalized histograms of collections of images within one event, we end up with a word histogram of size 1000. Each bin in the histogram indicates the occurrence frequency of the corresponding word. The complete vocabulary contains the integrated information of both color and texture. For selecting the informative features from the images, we can either mine the features from one image or images within one instance of an event. We choose the latter case, since some of the images can be outliers and may not represent the typical characteristics of an event. Once the most informative features are mined with PageRank, each event instance is represented with histogram of visual words.

3. Experiments and Discussion

A photo collection of 10 common types of event categories $C = \{\text{Christmas, backyard, ball game, beach, birthday, city walk, hiking, road trip, skiing, wedding}\}$ is used as the dataset. There are both indoor and outdoor events. The dataset is quite challenging because within some of the event categories, one instance may dramatically vary visually from the other instances. Each category in the dataset contains a variable number of event instances ranging from 7 to 10. One instance of the category (i.e. event) con-

Christmas	80.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Backyard	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Ballgame	0.0	0.0	50.0	0.0	0.0	0.0	0.0	50.0	0.0	0.0
Beach	0.0	0.0	14.3	71.4	14.3	0.0	0.0	0.0	0.0	0.0
Birthday	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0
City Walk	30.0	20.0	0.0	0.0	0.0	50.0	0.0	0.0	0.0	0.0
Hiking	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0
Road Trip	0.0	0.0	20.0	0.0	0.0	0.0	0.0	80.0	0.0	0.0
Skiing	0.0	0.0	0.0	0.0	0.0	0.0	11.1	0.0	88.9	0.0
Wedding	0.0	0.0	0.0	0.0	14.3	0.0	0.0	0.0	0.0	85.7
	Chris	Byard	Bgame	Beach	Birthday	Cityw	Hiking	Roadt	Skiing	Wed

Figure 4: Confusion table for the best run (81.2%) among the eight runs of the cross validation, which has an overall average accuracy of 78.1% using the PageRank technique.

tains a variable number of images ranging from 7 to 108. For example in the city category, photos taken from different cities are stored in 9 folders each of which represents an instance of the event. There are a total of 3453 photos in the dataset.

To verify the effectiveness of our method, we compare it with the standard BOF approach. In the first experiment, we randomly select 2000 features from each image. The FSG is constructed for all the images within the same folder and using PageRank, the top 20% features (400 features) are selected as the most informative ones. Then using the learned vocabulary, we represent each event as a bag of words. We choose SVM with histogram intersection kernel as classifier. In the testing phase, 7-fold cross validation scheme is used and the average recognition accuracy observed is 78.1%. The confusion table (for the best run in the cross validation) is shown in Figure 4. The largest confusion is between ball game and road trip events. This is consistent with our observation that both event types are outdoors and to a large extent contain green-colored fields and man-made structures.

On the other hand, to have a fair comparison with the standard BOF method, we randomly select 400 features from each image. Using the same learned vocabulary, we represent each event as a bag of words without using PageRank. The average accuracy is found to be 74.0%. Therefore an improvement of about 4.1% is achieved using the proposed method. Moreover, we can lower the matching threshold when performing image matching, thus retaining a higher number of matched features between pair of images. Repeating the steps of the first experiment, we achieved a further improved average accuracy of 79.4% compared to 78.1% in the first experiment. In general, further lowering the threshold can produce better performance but computation is more expensive. It may also be notable that although the recognition accuracy has been reported to 80.7% for this problem [11], yet our intent is to demonstrate the improvement in results using PageRank Method. We used a very few features from each image instead of using all of them and achieved comparable results.

In another experiment, to demonstrate that recognition is more robust with multiple photos than a single photo, we take a single photo to represent one instance of event and perform event classification on this new event representation. We run the experiments 8 times. As shown in Figure 5, these results confirm our conjecture that more robust event recognition can be achieved using multiple photos of an event (even without using PageRank).

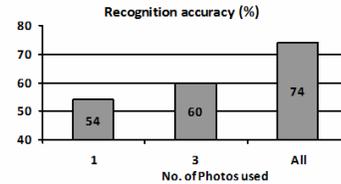


Figure 5: Average accuracy when different number of images selected from each event for recognition.

4. Conclusions

We present a novel approach to discover the most informative features for event recognition from photo collections. In order to acquire good features, we employ the PageRank technique in mining the informative visual features in combination with the bag of features method for event recognition. In particular, the consistent and thus distinguishing features of an event based on the intrinsic correlation among the images of the same event are utilized. The experiments verified that our approach is effective for event recognition, and the average accuracy is improved using proposed mining method.

5. REFERENCES

- [1] L. Zelnik-Manor and M. Irani. Event-based analysis of video, CVPR 2001.
- [2] D. Xu and S.-F. Chang. Video event recognition using kernel methods with multilevel temporal alignment, PAMI 2008.
- [3] S. Ebadollahi, L. Xie, S.-F. Chang, J. R. Smith. Visual event detection using multi-dimensional concept dynamics, ICME 2006.
- [4] J. Liu, Y. Zhai, et al. University of Central Florida at TRECVID 2006: High-Level Feature Extraction and Video Search, TRECVID 2006.
- [5] J. Liu, J. Luo, and M. Shah. Recognizing realistic actions from videos in the wild, CVPR 2009.
- [6] W. Jiang and A. C. Loui. Semantic event detection based on visual concept prediction, ICME 2008.
- [7] A. Oliva and A. Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope, IJCV 2001.
- [8] G. Kim, C. F. and M. Hebert. Unsupervised modeling of object categories using link analysis techniques, CVPR 2008.
- [9] Y. Wu, J.-Y. Bouguet, A. Nefian, and I. Kozintsev. Learning concept templates from web images to query personal image databases, ICME 2007.
- [10] L. Cao, J. Luo, H. Kautz and T. S. Huang. Annotating collections of photos using hierarchical event and scene models, CVPR 2008.
- [11] J. Yuan, J. Luo and Y. Wu. Mining compositional features, CVPR 2008.
- [12] J. Liu, Y. Yang and M. Shah. Learning Semantic Visual Vocabularies Using Diffusion Distance, CVPR 2009.
- [13] Y. Jing and S. Baluja. VisualRank: Applying PageRank to large-scale image search, PAMI 2008.
- [14] D. G. Lowe. Distinctive image features from scale invariant keypoints, IJCV 2004.
- [15] M. Leordeanu and M. Hebert. A spectral technique for correspondence problems using pair wise constraints, ICCV 2005.