ALBUM-BASED OBJECT-CENTRIC EVENT RECOGNITION

Shen-Fu Tsai, Thomas S. Huang

University of Illinois at Urbana-Champaign
Coordinated Science Laboratory
1308 W. Main Street, Urbana, IL 61801-2307
stsai8, t-huang1@illinois.edu

Feng Tang

Hewlett-Packard Labs
Multimedia Interaction and Understanding Lab
3000 Hanover Street Palo Alto, CA 94304-1185
feng.tang@hp.com

ABSTRACT

We explore event recognition in the personal photo album setting where the task is not to identify event in individual photo but in the whole album. This setting arises from the way people organize their own collections and the fact that individual photo in these albums often fails to convey meaningful event semantic behind the the album. We work on this problem in a object-centric manner, i.e. we train detectors for objects relevant to the events in the holiday dataset we built, and then detect these holidays based on object detector outputs. The prior knowledge, i.e. what objects are relevant to the event, is obtained statistically from mass image collection web site and thus tends to be more accurate and less biased.

Index Terms— Event recognition, prior knowledge, personal photo album, object detection

1. INTRODUCTION

Recognizing event in multimedia has been an interesting problem in multimedia community. Indeed, compared with objects and scenes, event usually express richer semantics which indicates directly what is going on in the multimedia. This is especially true for personal photo album, as people tend to be interested in certain event that has occurred in the past, and hence they very often would like to retrieve their personal collection based on event. Therefore, an event recognition system with reliable performance can be a very useful and crucial component for automatic photo management application.

Event is a higher level concept than object and scene. In fact, it is based upon these lower level concepts. In [1, 2], the authors built a set of 200 object detectors and used the detector outputs as image feature. They further applied this image feature for scene classification. It is then natural to ask if we can better utilize the detectors with help of prior knowledge of relations between event and object. Also, conventional image classification aims to classify individual image into one of the predefined categories, while in practice in most personal photo collection, the semantics is conveyed not by individual photo, but through the entire album. In other words, the underlying events of many photos in an album usually cannot be determined when they are viewed singly. They need to be treated as a whole for the viewer to be able to deduct what is going on in the album. This leads to a very interesting research problem which to the best of our knowledge has rarely been considered in the literature: namely, how do we infer the semantics from a set of images? How do the clues we obtain from different images interact with each other to give a more accurate and more robust inference of event? We point out here that the album event recognition is a problem different from video event detection, because consecutive frames in the latter bear clear temporal relations, while photos of an album are not necessarily ordered by time stamps, not to mention the fixed time duration between consecutive frames of video. The work [3] relies on meta information such as time or GPS to improve the album event recognition accuracy, however, those meta data might be changed due to image format change or photo redistribution, or they may even be unavailable. The authors of [4] discover frequent visual features within an event and then rank them using PageRank algorithm. Our work differs from theirs in many ways, one being that we explicitly employ object detection which is at the middle semantic level, higher than the low level feature used by [4].

In this work, we build a special dataset which consists of personal albums of 10 popular holidays in the United States (Section 2.1, Section 2.2). We further pick tens of objects relevant to these holidays (Section 2.3) and trained a detector for each of them (Section 3). We combine the outputs of the detectors to form a feature vector for each album, and then train and test the event detector. (Section 4). In Section 5 and Section 6, we discuss future directions and summarize our work, respectively.

2. DATA COLLECTION

2.1. Selecting Holidays

Our holiday dataset consists of photo albums from important and popular holidays. To decide what holidays to work on, we started from holiday statistics obtained from Flickr

978-1-61284-350-6/11/$26.00

Table 1. Number of queried images on Flickr

Rank	Holiday	#image
1	Christmas	6864906
2	Halloween	3811883
3	Easter	2047065
4	Thanksgiving	1654006
5	Independence Day	402611
6	New Year's Eve	356713
7	Mardi Gras	214093
8	Memorial Day	186663
9	St. Patrick's Day	181822
10	Valentine's Day	128520
11	Labor Day	121884
12	Mother's Day	108961

(http://www.flickr.com). Flickr holds tremendous photos uploaded by countless users in the world, so the statistics of holidays gathered on it comes from sufficient number of samples.

We obtain the initial pool of holiday from a Wikipedia page [1] that lists about 40 popular holidays in the United States. To rank the importance and popularity of these holidays, we query them on Flickr and count the number of returned photos.

Table 1 shows the number of Flickr images. These numbers may have changed a little bit as they were observed in October 2010, but the overall ranking should remain the same, because we are counting images accumulated since Flickr launched in 2004. For our holiday dataset, we decided to pick the top ten: Christmas, Halloween, Easter, Thanksgiving, Independence Day, New Year's Eve, Mardi Gras, Memorial Day, St. Patrick's Day, and Valentine's Day. We exclude Labor Day and Mother's Day because they are often not visually recognizable to human, and we decided to set our problem at least doable for human.

2.2. Collecting Holiday Albums

Although we determine what holidays to work on based on Flickr statistics, we collect albums from Google's Picasa Web Albums (http://picasaweb.google.com) instead, because we noted that it has albums closer to what ordinary people would upload, which are of more interest to us.

To find albums that contain specific holidays, we found that checking album titles gives more precise results. Therefore for each holiday H, we query its images using API provided by Picasa Web Albums and get a set of photos I_H. Then for each photo in I_H, we put its author into the set A_H if it is not already in A_H. A_H is the set of authors we are interested in when it comes to holiday H. Next, we download all the albums of each author in A_H with titles that contain string H

[1] http: //en.wikipedia.org/wiki/Public_holidaysi_in_the_United_States

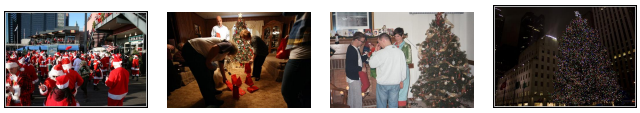

Fig. 1. Typical Christmas photos

Fig. 2. Typical Easter photos

or its equivalent form, e.g. all albums with title containing *Thanksgiving* or *Thanksgiving day*. However these albums do not always exhibit the corresponding holiday activity, so we manually examine each of them and discard the irrelevant albums. Again this is to set the problem at least doable for human. After manual inspection our dataset has about 50 albums for each holiday, 565 albums and 46609 photos in total. Figure 1 to Figure 10 show typical photos of these holidays which reflect the event semantic. However, it is also observed that not every single photo in these albums when viewed individually can be recognized as pertaining to the corresponding holiday. In fact, just like our own personal albums, many of these photos do not have much to do with the holiday, and only a portion of the images exhibit the holiday's property. Thus when we inspect each album with matching title, we either discard the whole album or keep it entirely, so that the resulting dataset is coherent with what personal album management application would face in real life.

2.3. Selecting Relevant Objects

As we will use an object-centric event recognition algorithm, we need to first select objects relevant to the holidays previously determined. Instead of manually assigning relevant objects based on our subjective knowledge, we prefer to decide what objects are relevant in a more systematic manner, e.g. similar to how we came up with the 10 popular holidays in the previous section. Therefore, again we count on Flickr for more unbiased statistics gathered from the masses.

To begin with, we need a initial pool of objects from which to choose the relevant objects. For distinction, we call this initial pool the *potentially* relevant objects. For each tag, through its API Flickr provides a list of *potentially* relevant tags, with the length of list ranging from 20 to 100. With this

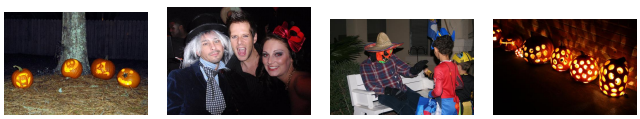

Fig. 3. Typical Halloween photos

Fig. 4. Typical Independence Day photos

Fig. 5. Typical Mardi Gras photos

Fig. 6. Typical Memorial Day photos

Fig. 7. Typical New Year's Eve photos

Fig. 8. Typical St. Patrick's Day photos

Fig. 9. Typical Thanksgiving photos

Fig. 10. Typical Valentine's Day photos

Table 2. Top 10 relevant tags to Easter

tag	relevance to Easter
egg OR eggs	0.0943680930259
hunt	0.0385255313823
bunnies OR bunny	0.034209284513
easteregg OR eastereggs	0.0165373738498
april	0.0123208604559
basket	0.0122615830513
weekend	0.00995692596803
happy	0.00879152543471
holiday OR holidays	0.00871898873185
church	0.00806227488627

we can obtain for each holiday H_i a set of *potentially* relevant tags R_{H_i}. We then obtain the set of *potentially* relevant objects $R = R_{H_1} \cup R_{H_2} \cup \ldots R_{H_{10}}$, the union of tags R_{H_i} *potentially* relevant to holiday H_i. In our case, there are totally 495 tags in R.

Next, for each holiday H_i we sort elements of R by

$$r(H_i, t) = \frac{|I(H_i) \cap I(t)|}{|I(H_i) \cup I(t)|} \quad (1)$$

, where t is a tag in R, and $I(x)$ denotes the set of queried images of string x on Flickr. Intuitively, this is the ratio of the number of the intersection of images associated with H_i and with t to the number of the union of them. Measurement $r(H_i, t)$ appropriately expresses co-occurrence relation between holiday H_i and object t. It penalizes the case where $|I(H_i) \cap I(t)| \approx |I(t)| \ll |I(H_i)|$, i.e. when the presence of object t almost implies the presence of holiday H_i, but very little can be said about the former when the latter is observed. Similarly the case where $|I(H_i) \cap I(t)| \approx |I(H_i)| \ll |I(t)|$ is also penalized. In the future we may also try using other measure of relevance to rank the tags with respect to holiday H_i.

As an example, Table 2 shows the top 10 tags for holiday Easter, sorted according to Equation 1. It is observed that the top 10 tags are indeed very related to Easter. However, we also noted that some of them are not objects, e.g. *hunt*, *april*, *weekend*, *happy*, and *holiday*. Hence out of the 495 tags in the initial set R, we manually compile a list of 305 non-object tags functioning as a stop word list. To control the number of relevant objects, after removing these non-object tags we pick up to 7 tags for each holiday such that each of their relevances to the holiday is at least 0.003. The resulting list consists of 53 relevant objects. Table 3 shows the final objects relevant to each of the holidays.

3. OBJECT DETECTORS

After deciding what objects to work on, we are ready to train the corresponding object detectors on which the object-

Table 3. Final relevant objects

holiday	objects
Christmas	christmas tree, decoration, gift, light, ornament, present, tree
Easter	basket, bunnies, chocolate, church, easter egg, egg, rabbit
Halloween	carving, costume, jackolantern, lantern, parade, pumpkin, zombie
Independence Day	bbq, explosion, firework, flag, parade, pyrotechnics sparkler
Mardi Gras	barkus, beads, costume, endymion, float, parade, zulu
Memorial Day	americanflag, bbq, cemetery, flag, memorialdayparade, parade, soldier
New Year's Eve	champagne, firework, sparkler
St. Patrick's Day	bagpipes, guinness, leprechaun, parade, saintpatricksdayparade, shamrock,stpatricksdayparade
Thanksgiving	cooking, dinner, food, parade, pie, pumpkin, turkey
Valentine's Day	candy, chocolate, cupcakes, cupid, heart, sweethearts

centric event recognition algorithm is to be built. The first step to training object detector is to collect sufficient training images with object bounding box. However, object bounding box is not readily available for all the relevant objects in our list, so after resorting to various sources including ImageNet[5], LabelMe[6], and Google Image Search, we modified our relevant object set to the 38-object list shown in Table 4. As can be seen, we skip some objects whose bounding box training images we have yet been able to collect from the web. We also put in some other objects which we think would be useful as well, although not directly relevant according to Flickr statistics. See the last row of Table 4.

We adopt the state-of-the-art object detector program provided by [7, 8] to build the 38 object detectors based on 100-200 training images per object. These general object detectors employ Histogram of Gradient (HOG) low level image features together with a part-based model to capture and allow the parts within an object. We train on half of the images and test against the other half to evaluate the performance of our object detectors. Table 5 shows the average precision (AP) performances of these detectors. We notice that the performance varies across different objects with their various visual variety and difficulty.

4. EVENT RECOGNITION

In this paper, our task is to recognize holiday, a particular form of event, from test albums. This is in contrast to conventional image classification that classifies individual image. As explained in Section 2.2, this is exactly the problem to be solved in real world because a set of photos can express the semantic more precisely.

Table 4. Relevant objects used in practice

holiday	objects
Christmas	christmas tree, gift
Easter	easter egg, basket, rabbit, church
Halloween	attire, pumpkin, jack-o-lantern
IndependenceDay	american flag, firework, crowd
MardiGras	mask, necklace, attire, feather boa, crowd
MemorialDay	american flag, uniform, military uniform, music band
NewYearsEve	champagne, firework, crowd
StPatricksDay	music band, crowd
Thanksgiving	food, dinner, turkey, pumpkin
ValentinesDay	heart, bouquet
not directly relevant	accordion, bassoon, child, cross, drum, euphonium, flag, french horn, light source, room light,shopping basket, soil, stage, table

It is actually not as trivial as one would imagine to extract a suitable feature for individual album, because the only features we want are those that mark the album off the others, and these features do not come from all images within the album, but only a portion of them. In other words, we do not want to extract anything from the photos that cannot be recognized when viewed individually. If we assume that these "neutral" images do not contain relevant objects and hence have low response of the corresponding detectors, it is then intuitive to consider the maximum response of every object detector on all locations of all photos within the album. Formally, we define our album feature as

$$\mathbf{x} = (x_1, x_2, \ldots, x_{38})^T \tag{2}$$

where x_i is the maximum response of object detector i against all images in the album. x_i can be intuitively interpreted as the maximum possibility that object i appears in the album, up to scaling and a monotonic transformation.

We test the proposed feature as follows. For each holiday, we treat it as a binary classification problem, and train on the training albums a set of Support Vector Machine (SVM) classifiers with different set of parameters. The best performance in terms of average precision (AP) and area under the ROC curve (AUC) among different parameters is reported. The experiment is carried out 10 times to obtain the average performance.

In Table 6, the results for different features are shown. We first normalize each dimension of $\mathbf{x}$ in Equation 2 to unit norm

Table 5. Average precision (AP) performance of object detectors

object	ap	object	ap	object	ap
accordion	0.333	drum	0.179	military uniform	0.005
american flag	0.679	easter egg	0.169	music band	0.199
attire	0.268	euphonium	0.338	necklace	0.658
basket	0.014	feather boa	0.003	pumpkin	0.209
bassoon	0.250	firework	0.592	rabbit	0.215
bouquet	0.375	flag	0.393	room light	0.102
champagne	0.272	food	0.320	shopping basket	0.002
child	0.069	french horn	0.514	soil	0.263
christmas tree	0.733	gift	0.243	stage	0.085
church	0.105	heart	0.380	table	0.423
cross	0.426	jack-o-lantern	0.604	turkey	0.433
crowd	0.490	light source	0.139	uniform	0.176
dinner	0.279	mask	0.381		

Table 6. Experimental results

Metric	AP			AUC		
Feature	all	ap	rel	all	ap	rel
Christmas	0.43	0.486	**0.67**	0.796	0.819	**0.885**
Easter	**0.34**	0.325	0.24	**0.825**	0.814	0.745
Halloween	0.55	**0.578**	0.52	0.821	**0.829**	0.783
Ind. Day	**0.52**	**0.520**	0.50	0.823	**0.826**	0.805
Mardi Gras	**0.55**	0.541	0.53	**0.895**	0.881	0.877
Memorial Day	**0.65**	0.637	0.32	**0.916**	0.901	0.801
New Year's Eve	**0.39**	0.376	0.33	**0.752**	0.744	0.732
St. Patrick's Day	**0.23**	0.199	**0.23**	**0.694**	0.662	0.652
Thanksgiving	0.76	**0.769**	0.71	0.964	**0.969**	0.936
Valentine's Day	**0.41**	0.404	0.33	**0.791**	0.787	0.725
Average	0.48	**0.484**	0.44	**0.828**	0.823	0.794

separately, where the variance of each dimension is estimated from training set. In Table 6, the second and 5th columns, denoted by "all", present the results with all dimensions of $\mathbf{x}$ (after normalization). The third and 6th columns, denoted by "ap", are results of the feature weighted by the estimated average precision performances of object detectors shown in Table 5, i.e. we heuristically shrink the output of poor object detectors and rely more on the good ones. In the 4th and 7th columns, denoted by "rel", we show results of the feature consisting of only relevant object detectors' output based on Table 4, e.g. when detecting Christmas only the dimensions corresponding to *christmas tree* and *gift are used*.

We observe that the average precision performance of the event detectors in Table 6 are in general higher than those of object detectors as shown in Table 5, though our event detectors is built solely upon object detectors. This may have something to do with the nature of our problem, i.e. album-based event recognition, as we do not need to recognize the event in every photo, but only have to determine whether it is present in the entire album, thus our proposed feature is more robust to detector output noise. Another possible reason is that the detection of each event is based on not only one object but several instead, so the information received about the event is actually richer than that of individual object.

We also noted that although using all dimensions of feature ("all") gives the best performance, for Christmas and St. Patrick's Day taking only relevant features ("rel") actually helps. For Christmas it might be that it is highly correlated with Christmas tree, whose object detector performs well enough with 0.733 average precision. On the other hand, as predicted the feature weighted by prior average precision performance of object detector ("ap") does work slightly better for some holidays. Overall, for all events there is a significant gap between our performance and random guess, whose average precision is roughly $(\#class)^{-1} = 0.1$ and area under the ROC curve is around 0.5.

5. FUTURE DIRECTIONS

There are several possible directions we would like to follow. First, for better performance we may want to build better object detectors. We may consider object detectors specialized for various objects, using different feature like color instead of the Histogram of Gradient (HOG) feature used by the current object general detector [8]. We can also increase the number of training images for the object detector. To obtain an even more realistic application, we would like to increase the number of holiday albums and the number of holidays or events.

In this paper we utilize album-based feature to do event classification. In this feature (from Equation 2), lots of information of individual image are discarded. Therefore we would like to have a more sophisticated model which essentially treat the task as multiple instance learning (MIL) prob-

lem [9]. In the setting of MIL, in addition to limited sample (photo) labels there are also bag (album) labels. Positive bag label indicates the existence of positive sample(s) within the bag though the actual indices of positive samples are unknown. Negative bag contains only negative samples. Clearly our problem of album-based event detection fits into the MIL setting, so it is indeed a direction worth going in order to establish a more accurate model for our album-based task.

We also want to model the relation between event and object. We want to build a probabilistic model that indicates given a hypothesized event E, what the chance of a certain object to appear is. The model is in fact more complicated than this, because aside from co-occurrence relation between event and object, the relation between objects has to be considered too. It is therefore an interesting topic as to how we can learn these relations from various sources and then impose them to our model.

6. SUMMARY

In this paper, we present our work on object-centric holiday recognition as a special form of event detection. We first determined what holidays to work on based on holiday popularity statistics gathered from Flickr. Then we apply prior knowledge of what objects are relevant to the holidays, again according to relevance measure based on Flickr statistics. After gathering holiday dataset and training images for object detectors, we trained a set of 38 object detectors and ran them against the holiday images. We came up with a simple album feature that takes the maximum responses of each object detector among all images in the album. Our experiments show that the object detector output and resulting album feature performs significantly better than random guess. For better performance and better utilization of prior knowledge, several possible future directions are discussed.

7. REFERENCES

[1] Eric P. Xing Li-Jia Li, Hao Su and Li Fei-Fei, "Object bank: A high-level image representation for scene classification & semantic feature sparsification," in *Neural Information Processing Systems (NIPS)*, Vancouver, Canada, December 2010.

[2] Yongwhan Lim Li-Jia Li, Hao Su and Li Fei-Fei, "Objects as attributes for scene classification," in *European Conference of Computer Vision (ECCV), International Workshop on Parts and Attributes*, Crete, Greece, September 2010.

[3] L. Cao, J. Luo, and T.S. Huang, "Annotating photo collections by label propagation according to multiple similarity cues," in *Proceeding of the 16th ACM international conference on Multimedia*. ACM, 2008, pp. 121–130.

[4] Naveed Imran, Jingen Liu, Jiebo Luo, and Mubarak Shah, "Event recognition from photo collections via pagerank," in *Proceedings of the 17th ACM international conference on Multimedia*, New York, NY, USA, 2009, MM '09, pp. 621–624, ACM.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *CVPR09*, 2009.

[6] Bryan C. Russell, Antonio Torralba, Kevin P. Murphy, and William T. Freeman, "Labelme: A database and web-based tool for image annotation," *Int. J. Comput. Vision*, vol. 77, pp. 157–173, May 2008.

[7] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester, "Discriminatively trained deformable part models, release 4," http://people.cs.uchicago.edu/∼pff/latent-release4/.

[8] Pedro F. Felzenszwalb, Ross B. Girshick, David McAllester, and Deva Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, pp. 1627–1645, 2010.

[9] Thomas G. Dietterich, Richard H. Lathrop, and Tomás Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artif. Intell.*, vol. 89, pp. 31–71, January 1997.