Compositional Object Pattern: A New Model For Album Event Recognition*

Shen-Fu Tsai
Coordinated Science Laboratory
University of Illinois at Urbana-Champaign
stsai8@illinois.edu

Liangliang Cao
Multimedia Group
IBM Watson Research Center
liangliang.cao@us.ibm.com

Feng Tang
Hewlett-Packard Labs
1501 Page Mill Road, Palo Alto, CA 94304 USA
feng.tang@hp.com

Thomas S. Huang
Coordinated Science Laboratory
University of Illinois at Urbana-Champaign
t-huang1@illinois.edu

ABSTRACT

In this paper, we study the problem of recognizing events in personal photo albums. In consumer photo collections or online photo communities, photos are usually organized in albums according to their events. However, interpreting photo albums is more complicated than the traditional problem of understanding single photos, because albums generally exhibit much more varieties than single image. To solve this challenge, we propose a novel representation, called Compositional Object Pattern, which characterizes object level pattern conveying much richer semantic than low level visual feature. To interpret the rich semantics in albums, we mine frequent object patterns in the training set, and then rank them by their discriminating power. The album feature is then set as the frequencies of these frequent and discriminative patterns, called Compositional Object Pattern Frequency(COPF). We show with experimental result that our algorithm is capable of recognizing holidays with accuracy higher than the baseline method.

Categories and Subject Descriptors

I.4 [**Image Processing and Computer Vision**]: Feature Measurement, Image Representation

General Terms

Algorithms, Experimentation

*Area chair: Kiyoharu Aizawa

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, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
MM'11, November 28–December 1, 2011, Scottsdale, Arizona, USA.
Copyright 2011 ACM 978-1-4503-0616-4/11/11 ...$10.00.

Keywords

Photo Albums, Event Recognition, Compositional Object Pattern

1. INTRODUCTION

In vision and multimedia community, many studies have been dedicated to single image understanding, especially object recognition and scene classification. In contrast, there are less attention paid to album-level event classification, although most consumer photos are kept in the form of albums. Because people usually organize their photo albums corresponding to different events, it is crucial for consumer photo management systems to recognize the underlying event within each album. However, the task of album event recognition is harder than conventional single image understanding, as a collection of photos exhibits more variety than single image. From another viewpoint, event is a higher level concept than other subject like object and scene, and therefore to interpret the event in an album is a more challenging task. In particular, albums often consist of "typical" and "non-typical" photos. "Typical" photos are those when viewed individually the event can be recognized by human. Conventional image-based classification techniques will probably fail to recognize the non-typical photos and hence decrease the accuracy of album event recognition.

Since an event involves many high level concepts, it is difficult for low level feature to capture its rich semantic. However, most previous image understanding methods mostly use low level features for high level semantic analysis. The semantic gap between low-level features and high level concepts impose a serious problem for current event recognition methods. As suggested in [1], low-level feature based image classification is not able to achieve very high accuracy in event classification. The work [1] relies on meta information such as time or GPS to improve the recognition accuracy, however, those meta data might be changed due to image format change or photo redistribution. In this paper, we explore another way to bridge the semantic gaps, which employs high level object detection for high level semantic inference.

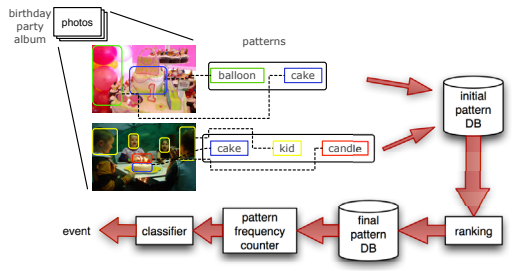

Figure 1: Overview of our proposed algorithm

Our work is partly motivated by [10], which employed an object detector bank to extract object semantic as the image feature for classification. The resulting image feature works well on "typical" images, however, could not be used for "non-typical" ones. In contrast, this paper explores middle level concepts in the album level, and proposed a new model called Compositional Object Patterns to mine the context information within a photo albums, which could reliably recognize the event robustly subject to the affects of "non-typical" images.

We propose to solve the problem of album event classification by mining relevant and discriminative object patterns in the albums. The large number of the mined frequent patterns are further ranked by their discriminating power, and the top ranked ones are the final patterns that are used to represent each album. We show that our algorithm discovers semantically meaningful object pattern, and yields satisfying performance on a photo album dataset from real life.

The rest of the paper is organized as follows. Section 2 reviews previous relevant works. Section 3 describes how itemset mining is used to explore object patterns. Section 4 presents the proposed pattern ranking followed by album representation. Section 5 and Section 6 present experimental results and conclusion, respectively.

2. RELATED WORKS

Collection-level analysis of cosumer photos has not drawn too much attention in the past. In [2, 3] the authors explored relationship between collection-level annotation and image-level annotation with help of GPS and time information associated with the photos, whereas in our work we are more interested in understanding image based on only visual content. Authors of [8] proposed to describe images and video frames by a set of scores of visual concepts. Then, unlike our pattern mining approach, they cluster the resulting image features directly to build event feature vocabulary, which may be affected by irrelevant visual feature and noisy output of visual concept detector, while the patterns we discover should be more robust to these two factors.

In recent years, there has been a couple of works on mining compositional feature. The authors of [5] pointed out that compositional feature, viewed as a combination of primitive features, is a non-linear transformation from the primitives and thus can have higher discriminative power. They propose Fisher score and information gain to measure relevance of the mined compositional feature. In [14], the authors mine frequent compositional patterns in the whole training set followed by multiclass Adaboost to select feature for image-based classification. Combined with decision tree classifier, [7] proposes to discover frequent yet discriminative pattern while growing the decision tree during training process. While these mentioned works assume general feature vector, [15] explicitly explores frequent quantized visual patterns with self-discovered distance metric used by clustering algorithm. Note that none of these works explicitly uses object detection. The feature closest to middle level concept detection is the PCA-SIFT[9] visual primitive used in [15].

3. ITEMSET MINING

3.1 Itemset Terminology

A *transaction database* $\mathcal{T}$ is a set of *transactions*, where each transaction T, associated with a unique transaction id, is a set of distinct *items*. Let $I = \{i_1, \ldots, i_{|I|}\}$ be the complete set of distinct items appearing in $\mathcal{T}$. An *itemset* D is a non-empty subset of I. A transaction T is said to *contain* itemset D if $D \subseteq T$. The number of transactions in $\mathcal{T}$ containing itemset D is the *support* of itemset D, denoted as $sup(D)$. Given a minimum support threshold min_sup, an itemset D is *frequent* if $sup(D) \geq min\_sup$. An itemset D is a *frequent closed itemset* if it is frequent and there does not exist any proper superset $D' \supset D$ such that $sup(D') = sup(D)$.

3.2 Proposed Object Itemset

3.2.1 Object Detection

For each image I, N object detectors are run against it, each producing a set of detections with some score. For each detector we keep only the maximum response (score) it generates, resulting in a N-dimensional feature vector $m(I) = (m_1(I), \ldots, m_N(I))$ where $m_n(I)$ is the maximum response of detector n within image I.

3.2.2 Itemizing Object Detection

To convert each continuous-valued $m_i(I)$ in $m(I)$, for each dimension(object) we apply Lloyd's algorithm[11] to compute least square quantization based on training set, assuming k quantization levels. Hence the quantized features is $q(I) = (q_1(I), \ldots, q_N(I))$, where $q_n(I) \in \{1, \ldots, k\}, \forall n$. With quantized detector responses $q(I)$, we then convert it to a transaction $T(I) = \{t(q_1(I)), \ldots, t(q_N(I))\}$, where function $t()$ transforms a quantized response from a specific detector to a unique item based on $t(q_n(I)) = k * (n-1) + q_n(I) \in \{1, \ldots, Nk\}$. Therefore, there are totally Nk distinct items, numbered from 1 to Nk. Each transaction has exactly N items.

3.2.3 Closed Itemset Mining

We apply mining algorithm $|\mathcal{E}|$ times, once for each event E_i from images in the training albums labeld as E_i, where $\mathcal{E} = \{E_1, \ldots, E_{|\mathcal{E}|}\}$ denotes the set of distinct events.

4. COMPOSITIONAL OBJECT PATTERN FREQUENCY (COPF) ALGORITHM

4.1 Pattern Ranking

For each event E_i, the set of patterns $\mathcal{D}_{E_i}$ thus mined are candidates for the final patterns used for recognizing E_i. To pick up the most discriminative patterns, we rank them

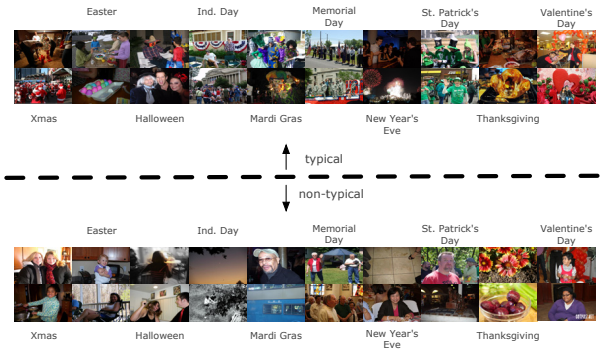

Figure 2: Sample typical and non-typical photos in the 10-Holiday dataset

according to their relevance to E_i. For each $D_j^{E_i} \in \mathcal{D}_{E_i}$ and for each album a_m in training set, we obtain the relative frequency f_j^m which is the percentage of photos containing itemset $D_j^{E_i}$ in album a_m, so $0 \leq f_j^m \leq 1$. Every album a_m is also associated with a binary variable $y_m^{E_i} \in \{0, 1\}$ depending on whether it is labeled E_i. Thus we can compute the average precision(AP) of detecting E_i based solely on f_j^m. We thus select the top patterns for each event E_i and form the final set of primitive patterns $\mathcal{P} = \{P_j\}_{j=1}^{|\mathcal{P}|}$ as the union of the top relevant patterns of all events.

4.2 Pattern Frequency And Classification

The final feature for album a_m is then computed as $f_m = (f_1^m, \ldots, f_{|\mathcal{P}|}^m)^T$ where each f_j^m is the relative frequency of photos containing P_j in album a_m. We call it Compositional Object Pattern Frequency (COPF). We train multiclass Support Vector Machine(SVM) classifier for event classification.

5. EXPERIMENTS

We collected a 10-Holiday dataset, came up with a list of 38 relevant objects semi-automatically, and trained the corresponding object detectors. For more details see [12]. Some typical and non-typical holiday photos are shown in Figure 2.

5.1 Implementation

5.1.1 Compostional Object Pattern Frequency(COPF)

The proposed Compositional Object Pattern Frequency (COPF) algorithm described in Section 4 is applied, where the number of quantization level is set to $k = 7$. When mining frequent closed itemsets, we use the program of CLOSET+ [13] provided by its authors, with threshold of support set to 30. For multiclass SVM classifier we adopt LIBSVM package provided at [4], using 5-fold cross validation on training set for parameter tuning.

5.1.2 Image-based Multiclass Adaboost (SAMME)

As a baseline model for comparison, we implement the algorithm in [14] as it is the most similar work to ours. The authors of this work proposed to mine frequent and discriminative compositional patterns from the whole training set in contrast of our mining frequent patterns with respect to each class seperately followed by pattern ranking. In their work, the mined patterns are then used by multiclass Adaboost[16] to find the most discriminative features for classification. When implementing their algorithm, we set k, the number of quantization level, to 7, same as COPF. However, the algorithm is only for image-based classification. So we attach the album label to all of its images for training, and when testing the album label is the majority vote of all its estimated image labels.

5.2 Results

5.2.1 Mined patterns

Before checking the classification accuracy, we first examine some top ranking object patterns. In Figure 3, we show some of the top patterns for each event. Here the number of quantization is set to $k = 7$, and each parenthesis encloses a corresponding level from 1 to k. By definition, objects that are absent in the pattern are "unimportant" or "irrelevant", i.e. whether they appear or not do not help distinguishing the event from others, just like the concept "person" does not help very much in recognizing "outdoor", as previously explained. We observe in the figure that most of the mined patterns are meaningful, in that they match our common knowledge and understanding of the holidays and objects. Moreover, these patterns are indeed sparse in objects, enabling our algorithm to get rid of detection noise, especially from the detectors of the irrelevant objects. Nonetheless, there is some noise in the patterns as well, mostly due to imperfect object detectors.

Holiday	top ranking patterns
Christmas	christmas tree(7) christmas tree(7) feather boa(3)
Easter	attire(3) firework(3) rabbit(6) flag(1) turkey(5)
Halloween	jack-o-lantern(7) pumpkin(7) uniform(3) food(3) jack-o-lantern(7) pumpkin(7)
Independence Day	child(1) easter egg(1) firework(7) child(1) firework(7) uniform(2)
Mardi Gras	attire(6) drum(4) light source(3) uniform(5) bassoon(2) champagne(3) easter egg(2) food(3) table(2)
Memorial Day	euphonium(3) jack-o-lantern(1) uniform(5) stage(6) uniform(6)
New Year's Eve	food(2) room light(4) room light(4) uniform(3)
St. Patrick's Day	child(3) cross(4) easter egg(3) french horn(3) military uniform(4) bouquet(4) food(5) rabbit(4) uniform(4)
Thanksgiving	christmas tree(1) easter egg(4) room light(2) attire(2) christmas tree(1) jack-o-lantern(2)
Valentine's Day	american flag(2) basket(3) heart(6) american flag(2) basket(3) christmas tree(1) heart(6)

Figure 3: Top ranking patterns for each holiday. Full response is 7.

5.2.2 Accuracy comparison

Figure 4 compares the accuracy performances of the baseline and proposed algorithms. Overall, COPF works significantly better with its average accuracy being 46.2%, which is 24.2% higher than SAMME. In fact, SAMME virtually breaks down with accuracy mostly under 30%. This is because it attempts to distinguish individual image before voting, while many images are common to more than one events, i.e. the overlapping between events is quite significant and it is impossible to classify these photos into one single event. SAMME works the best with New Year's Eve photo album because it is the only event composed of many night scene

pictures taken at night. It is also observed that, COPF does not work well on St. Patrick's Day, probably because its main feature, various green objects, is not well captured as the general object detectors we use are based on HOG feature[6].

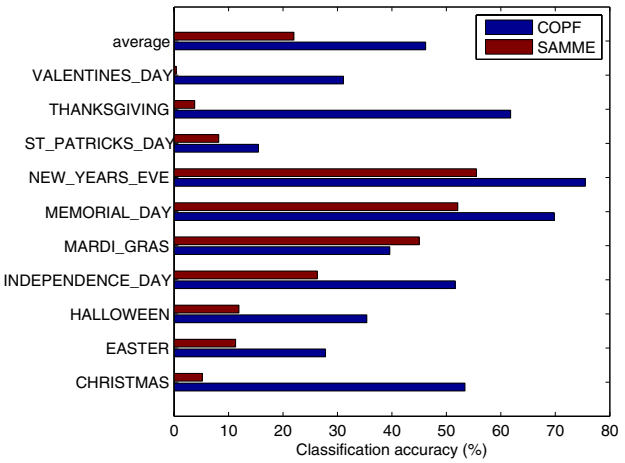

Figure 4: Classification accuracy comparison between the proposed algorithm COPF and the baseline algorithm SAMME

6. CONCLUSION AND FUTURE WORK

In this paper, we propose a novel model named Compositional Object Pattern to attack the problem of album event recognition. To accommodate the scenario of photo albums, we propose a novel object pattern mining algorithm which mines and ranks the discriminative object patterns for event classification. Based on mined patterns, albums are characterized by the relative frequencies of discriminative patterns. Experimental results show that our algorithm can distinguish events more accurately than baseline of the single image based classification. The object patterns discovered by our algorithm verify that object level semantic are meaningful and reliable.

In the future, we are interested in incorporating pattern itemset with the relative location of object detection, and we believe such context information will be useful in many visual applications.

7. ACKNOWLEDGEMENT

This research was supported in part by HP Innovation Research Program and the National Science Council, Taiwan, R.O.C. under contract NSC-095-SAF-I-564-035-TMS (to Shen-Fu Tsai).

8. REFERENCES

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

[2] L. Cao, J. Luo, H. S. Kautz, and T. S. Huang. Annotating collections of photos using hierarchical event and scene models. In *CVPR*. IEEE Computer Society, 2008.

[3] L. Cao, J. Yu, J. Luo, and T. S. Huang. Enhancing semantic and geographic annotation of web images via logistic canonical correlation regression. In W. Gao, Y. Rui, A. Hanjalic, C. Xu, E. G. Steinbach, A. El-Saddik, and M. X. Zhou, editors, *ACM Multimedia*, pages 125–134. ACM, 2009.

[4] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[5] H. Cheng, X. Yan, J. Han, and C.-W. Hsu. Discriminative frequent pattern analysis for effective classification. In *ICDE*, pages 716–725, 2007.

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pages 886–893, 2005.

[7] W. Fan, K. Zhang, H. Cheng, J. Gao, X. Yan, J. Han, P. Yu, and O. Verscheure. Direct mining of discriminative and essential frequent patterns via model-based search tree. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pages 230–238, New York, NY, USA, 2008. ACM.

[8] W. Jiang and A. C. Loui. Semantic event detection for consumer photo and video collections. In *ICME*, pages 313–316. IEEE, 2008.

[9] Y. Ke and R. Sukthankar. Pca-sift: a more distinctive representation for local image descriptors. *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2004 CVPR 2004*, 2:506–513, 2004.

[10] E. P. X. Li-Jia Li, Hao Su and L. 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.

[11] S. Lloyd. Least squares quantization in pcm. *Information Theory, IEEE Transactions on*, 28(2):129–137, Mar. 1982.

[12] S.-F. Tsai, F. Tang, and T. S. Huang. Album-based object-centric event recognition. *Workshop on Visual Content Identification and Search in 2011 IEEE International Conference on Multimedia and Expo (ICME 2011)*.

[13] J. Wang, J. Han, and J. Pei. Closet+: searching for the best strategies for mining frequent closed itemsets. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '03, pages 236–245, New York, NY, USA, 2003. ACM.

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

[15] J. Yuan, Y. Wu, and M. Yang. From frequent itemsets to semantically meaningful visual patterns. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '07, pages 864–873, New York, NY, USA, 2007. ACM.

[16] J. Zhu, H. Zou, S. Rosset, and T. Hastie. Multi-class adaboost. Technical report, 2005.