SEMANTIC EVENT DETECTION FOR CONSUMER PHOTO AND VIDEO COLLECTIONS

Wei Jiang

Columbia University, New York, NY

Alexander C. Loui

Eastman Kodak Company, Rochester, NY

ABSTRACT

The automatic detection of semantic events in users' image and video collections is an important technique for content management and retrieval. In this paper we propose a novel semantic event detection approach by considering an event-level Bag-of-Features (BOF) representation to model typical events. Based on this BOF representation, semantic events are detected in a concept space instead of the original low-level visual feature space. There are two advantages of our approach: we can avoid the sensitivity problem by decreasing the influence of difficult or erroneous images or videos in measuring the event-level similarity; also we can utilize the power of higher-level concept scores in describing semantic events. Experiments over a large real consumer database confirm the effectiveness of our approach.

Index Terms— Semantic event detection, concept detection

1. INTRODUCTION

Automatic albuming of consumer photos and videos has gained great interest in recent years [1]. One popular approach is to organize photos and videos according to *events* by chronological order and by visual similarities. In this paper we explore the important issue of semantic classification of the organized events from automatic albuming systems. That is, we try to tag semantic meanings (called *semantic events* in the rest of this paper) to the generated events by automatic albuming. This is a difficult problem in several aspects: first, we need to process photos and videos simultaneously which often both exist in real consumer collections; second, we need to accommodate the diverse semantic content of the consumer photos/videos, i.e., we aim at developing generic algorithms for detecting different semantic events instead of specific individual methods for detecting each semantic event; finally, our method needs to be robust to errors resulting from automatic albuming systems.

This paper presents a novel semantic event detection algorithm. We propose an event-level *Bag-Of-Features* (*BOF*) representation to model events, and based on this BOF representation semantic events are detected in a *concept space* instead of the original low-level feature space. The following highlights the motivations and advantages of our approach.

• We develop an event-level representation, where each event is modeled by a BOF feature vector based on which semantic event detectors are built directly. Compared with the naive approach where image-level feature vectors are used for training classifiers, our approach is more robust to the difficult images or mistakenly organized images within events. For example, Fig. 1 shows a "birthday" event, where some images (marked by red rectangular) are hard to detect. These difficult images usually make the decision boundary too complex to model. By adopting the event-level feature representation, we will be able to avoid the sensitivity problem by decreasing the influence of difficult or erroneous images/videos in measuring even-level similarities. As shown in Sec. 3, good detection performance can be obtained with a small number of support vectors for SVM classifiers, i.e., the classification problem is significantly simplified by the event-level representation.

• Complex semantic events are usually generated by the concurrences of elementary visual concepts. For example, "wedding" is a semantic event associated with people, park, etc., evolving with a certain pattern. In this paper, elementary concepts are first detected from images, and semantic event detectors are built in the concept space instead of in the original low-level feature space. Our algorithm benefits from such an approach in two aspects. First, visual concepts are higher-level and more intuitive descriptors than original low-level features. As seen in our experiments and previous works [2], concept scores are powerful to model semantic events. Second, our concept space is formed by 21 concept detectors [3] trained over Kodak's consumer video dataset [4]. These concept detectors play the important role of incorporating additional information from the previous video set to help detect semantic events in the current data collection.

The proposed semantic event detection approach is outlined in Fig. 2. Experiments over the Kodak's collection from real consumers confirm the effectiveness of our method. In the rest of this paper, we provide details of our algorithm, followed by the experiments and discussions.

2. SEMANTIC EVENT DETECTION

We start with some definitions and terminologies. Assume that we have a large collection of data, including photos and video clips from real consumers. The entire data set can be partitioned into a set of *macro-events*, and each macro-event is further partitioned into a set of *events*. The partition is based on the capture time of each photo/video and the color

Fig. 1. Example of difficult images in a "birthday" event.

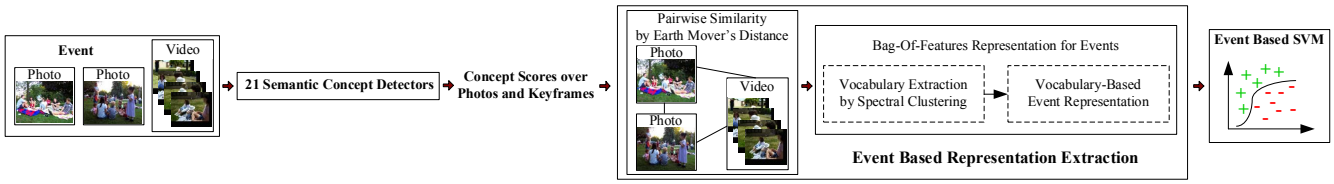

Fig. 2. Framework of the semantic event detection algorithm.

similarity between photos/videos, by using the previously developed event clustering algorithm [1]. Let E_t denote the t-th event which contains m_p^t photos and m_v^t videos, I_i^t and V_j^t denote the i-th photo and j-th video in E_t. Our target is to tag semantic meanings, i.e., *semantic events* S_E such as "wedding" and "birthday", to event E_t.

2.1. Concept Score Feature Construction

In this paper, we assume that the semantic events are generated by concurrent *visual concepts* like park and people. Let $C_1, \ldots, C_N$ denote N visual concepts. In [3] we have developed 21 ($N = 21$) SVM-based concept detectors using low-level color, texture, and edge visual features over Kodak's consumer video data set [4]. These concept detectors can be applied to generate 21 concept detection scores $p(C_1, I_i^t), \ldots, p(C_N, I_i^t)$ for each image I_i^t. These concept scores form a feature vector to represent image I_i^t in the concept space as: $\mathbf{f}(I_i^t) = [p(C_1, I_i^t), \ldots, p(C_N, I_i^t)]^T$.

Since the video clip from real consumers usually has diverse visual content from one long shot, each video V_j^t is partitioned into a set of segments $V_{j,1}^t, \ldots, V_{j,m_j}^t$, each with 5-sec length. Keyframes are then uniformly sampled from the video segments for every 0.5 second. Let $I_{j,k,l}^t$ be the l-th keyframe in the k-th segment. $I_{j,k,l}^t$ can also be represented by vector $\mathbf{f}(I_{j,k,l}^t)$ in the concept space.

Here we define that both photos and video segments are *data points*, represented by x. For example, event E_t contains $m^t = m_p^t + \tilde{m}_v^t$ data points in total, where $\tilde{m}_v^t$ is the entire number of video segments from the m_v^t video clips in E_t. In the next subsections, we will develop our semantic event detection algorithm based on these data points and the corresponding concept score features.

2.2. BOF Representation for Semantic Events

In this work, we extend the idea of Bag-Of-Features and construct a robust *concept vocabulary* for describing events.

The BOF representation has been proven effective for detecting generic concepts for images [5], [6]. In BOF, images are represented by a set of orderless local descriptors (SIFT features [6] or segmented regions [5]). Through clustering techniques, a middle-level *visual vocabulary* is constructed where each visual word is formed by a group of local descriptors. Each visual word is considered as a robust and denoised visual term for describing images.

Let S_E denote a semantic event, e.g. "wedding", and let $E_1, \ldots, E_M$ denote M events containing this semantic event. Each E_t is formed by m_p^t photos and $\tilde{m}_v^t$ video segments. Similar to the visual vocabulary, a concept vocabulary can be constructed by clustering these $\sum_{t=1}^{M} m^t$, (where $m^t = m_p^t + \tilde{m}_v^t$) data points into n concept words. Each concept word can be treated as a pattern of concept concurrences that is a common character for describing all the events containing S_E. Specifically, to accommodate both photo and video data points, the spectral clustering algorithm [7] is adopted to construct the concept vocabulary based on pairwise similarities measured by the *Earth Mover's Distance* (*EMD*) [8].

2.2.1. Pairwise similarity by EMD

We treat each data point as a set of images, i.e., one image for a photo and multiple images for a video segment. Then EMD [8] is used to measure the similarity between two data points (image sets). Note that there are many ways to compute the distance between two image sets, e.g. the maximum /minimum/mean distance between images in these two sets. These methods are easily influenced by noisy outlier images, while EMD provides a more robust distance metric. EMD finds a minimum weighted distance among all pairwise distances between two image sets subject to weight-normalization constraints, and allows partial match between data points and can alleviate the influence of outlier images.

The EMD between two data points is calculated as follows. Assume that there are n_1 and n_2 images in data points x_1 and x_2, respectively. The EMD between x_1 and x_2 is a linear combination of ground distance $d(I_p^1, I_q^2)$ weighted by flow $f(I_p^1, I_q^2)$ between any two images $I_p^1 \in x_i$, $I_q^2 \in x_2$.

$$D(x_1, x_2) = \frac{\sum_{p=1}^{n_1}\sum_{q=1}^{n_2} d(I_p^1, I_q^2) f(I_p^1, I_q^2)}{\sum_{p=1}^{n_1}\sum_{q=1}^{n_2} f(I_p^1, I_q^2)} \quad (1)$$

where an optimal flow matrix $f(I_p^1, I_q^2)$ is obtained from the following linear program:

$$\min_f \sum_{p=1}^{n_1}\sum_{q=1}^{n_2} d(I_p^1, I_q^2) f(I_p^1, I_q^2)$$

$$s.t.\ \forall\, 1 \le p \le n_1, 1 \le q \le n_2, f(I_p^1, I_q^2) \ge 0,$$

$$\sum_{p=1}^{n_1} f(I_p^1, I_q^2) \le w_q^2, \sum_{q=1}^{n_2} f(I_p^1, I_q^2) \le w_p^1,$$

$$\sum_{p=1}^{n_1}\sum_{q=1}^{n_2} f(I_p^1, I_q^2) = \min\left\{\sum_{p=1}^{n_1} w_p^1, \sum_{q=1}^{n_2} w_q^2\right\}$$

where w_p^1 and w_q^2 are weights of image I_p^1 and I_q^2 in data points x_1 and x_2, respectively. Here we take equal weights: $w_p^1 = 1/n_1$, $w_q^2 = 1/n_2$. The Euclidean distance over concept score features is used as the distance $d(I_p^1, I_q^2)$. From Eq(1), EMD finds the best matching image pairs in two data points. The weight normalization constraints ensure that each image has enough matches in the other set. When both x_1 and x_2 are photos, the EMD is just the Euclidean distance.

The pairwise EMD is then converted to the pairwise similarity by a Gaussian function: $S(x_1, x_2) = \exp(-D(x_1, x_2)/r)$, where r is the mean of all pairwise distances between all training data points.

2.2.2. Codebook construction and BOF representation

Spectral clustering is a technique for finding groups in data sets consisting of similarities between pairs of data points. Here we adopt the algorithm developed in [7], which can be described as follows. Given the similarity matrix $S(x_i, x_j)$:

• Get affine matrix $A_{ij} = S(x_i, x_j)$, if $i \neq j$, and $A_{ii} = 0$.

• Define diagonal matrix $D_{ii} = \sum_j A_{ij}$. Get $L = D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$.

• Find eigenvectors $\mathbf{u}_1, \ldots, \mathbf{u}_n$ of L corresponding to the n largest eigenvalues, and get $U = [\mathbf{u}_1, \ldots, \mathbf{u}_n]$, where n is determined by the energy ratio of eigenvalues to keep.

• Get matrix V from U by re-normalizing U's rows to have unit length.

• Treat each row in V as a point in $\mathbb{R}^n$ (the i-th row corresponding to the original i-th data point), and cluster all the points into n clusters via the K-means algorithm.

Each data cluster obtained by the spectral clustering algorithm is called a concept word, and all the clusters form a concept vocabulary to represent and detect semantic events. Let W_j^i denote the j-th word learned for semantic event S_{Ei}, $S(x, W_j^i)$ denotes the similarity of data x to word W_j^i calculated as the maximum similarity between x and the member data points in W_j^i: $S(x, W_j^i) = \max_{x_k \in W_j^i} S(x_k, x)$, where $S(x_k, x)$ is defined in the same way as in Sec. 2.2.1. For each data x, vector $[S(x, W_1^i), \ldots, S(x, W_n^i)]^T$ can be treated as a BOF feature vector for x. Assume that event E_t contains m^t data points, and based on above BOF feature vectors, event E_t can also be represented by a BOF feature vector $\mathbf{f}_{\text{bof}}(E_t)$ as: $\mathbf{f}_{\text{bof}}(E_t) = [\max_{x \in E_t} S(x, W_1^i), \ldots, \max_{x \in E_t} S(x, W_n^i)]^T$. Finally, using the BOF feature $\mathbf{f}_{\text{bof}}$ a binary one-vs.-all SVM classifier can be learned to detect semantic event S_{Ei}.

3. EXPERIMENTS

We evaluate our algorithm over 1972 consumer events from Kodak's consumer dataset, which are labeled to 10 different semantic events whose detailed definitions are shown in Table 1. A total of 1261 events are randomly selected for training, and the rest are used for testing. Please note that the training and testing data are partitioned at the macro-event level, i.e., events from the same macro-event will be treated together as training or testing data. This avoids the situation where similar events from the same macro-event are separated, which will simplify the classification problem.

We use *average precision* (*AP*) as the performance metric, which has been used as an official metric for video concept detection [9]. It calculates the average of precision values at different recall points on the precision-recall curve, and thus evaluates the effectiveness of a classifier in detecting a specific semantic event. When multiple semantic events are considered, the mean of APs (*MAP*) is used.

3.1. Concept Score versus Low-level Visual Features

To show the effectiveness of the concept score representation in the semantic event detection algorithm (SE Detection), we first compare our proposed method with the approach where the BOF feature vectors are constructed based on original low-level visual features. Specifically, we use the same low-level visual features as in [3]. Fig. 3 gives the performance comparison. From the figure, both methods consistently outperform random guess. SE Detection with concept scores significantly outperforms SE Detection with low-level features in terms of AP over most concepts and by 20.7% in terms of MAP. This result confirms the power of using previous concept detection models to help detect semantic events.

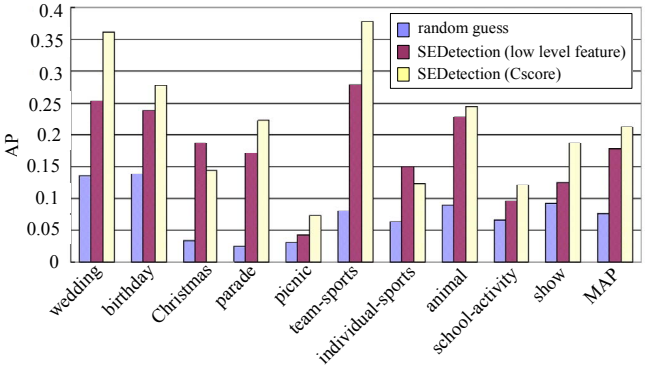

Fig. 3. Performance comparison of SE Detectors with concept scores and low-level visual features.

3.2. Event-Level versus Image-level Representations

In this experiment, we compare our SE Detection algorithm with two other detectors: (1) a baseline event detector (*baseline*), and (2) an SVM detector using image-level concept score representation directly (*SVM–Direct*).

• ***Baseline*** – In [3] we have developed SVM-based detectors to generate concept detection scores of 21 generic concepts. Six semantic events - "wedding", "birthday", "parade", "picnic", "animal", and "show" are included in the 21 concepts. These detectors can be directly applied for classifying photos and keyframes to get concept scores. Then for an event, the maximum detection score of the member images can be used as the detection score of the event. This event detection score gives a baseline detection result.

• ***SVM-Direct*** – By using concept detection scores over photos and keyframes, a one-vs.-all SVM classifier can be built at the image level to detect semantic events.

Fig. 4 gives AP comparison of different algorithms. From the figure, the proposed SE Detection performs the best over

semantic events	definition
wedding	bride and groom, cake, decorated cars, reception, bridal party, or anything relating to the day of the wedding
birthday	birthday cake, balloons, wrapped presents, and birthday caps, usually with the famous birthday song
Christmas	Christmas tree and the usual Christmas decorations, not necessarily taken on Christmas day
parade	processing of people or vehicles moving through a public place
picnic	outdoor, with or without a picnic table, with or without a shelter, people, and food in view
team sports	basketball, baseball, football, hockey and other team sports
individual sports	tennis, swimming, bowling and other individual sports
animal	pets (e.g., dogs, cats, horses, fish, birds, hamsters), wild animals, zoos, and animal shows
school activity	school graduation, school days (first or last day of school) and other events related to school
show	show and concerts, recitals, plays, and other events

Table 1. Definition of semantic events.

most of the semantic events and gets significant performance improvement of more than 20% compared with the second best method, over many semantic events like "wedding", "Christmas", and "school activity". This result confirms the success of the event-level BOF representation. Additionally, Fig. 5 gives the comparison of the number of support vectors from different algorithms. Generally, the less the support vectors the simpler the decision boundary. From the figure, decision boundaries are significantly simplified by event-level representation where SVM classifiers can separate semantic events quite well. Furthermore, Fig. 6 gives the top-5 detected events for "animal" by the baseline detector and our SE Detection method. From the figure, SE Detection can get 100% precision, while the image-based SVM-Direct method only gets 20% precision.

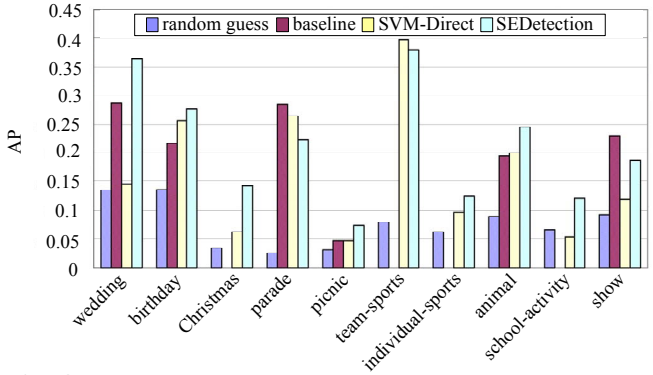

Fig. 4. AP comparison of the baseline detector, the image-level detector, and the proposed semantic event detector.

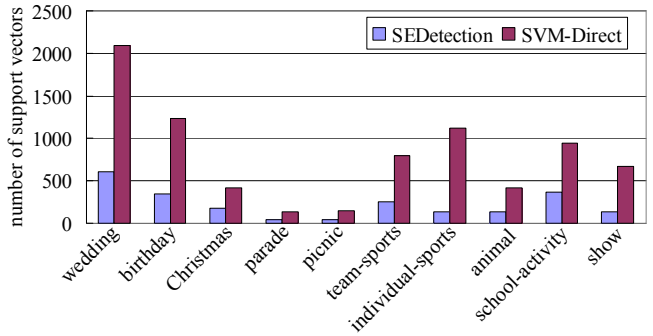

Fig. 5. Comparison of SVM models for different algorithms.

4. CONCLUSION

We propose a novel semantic event detection algorithm by developing an event-level BOF representation to reduce the influence of difficult and erroneous images and videos in measuring the event-level similarities. Based on this BOF representation, semantic event detectors are learned in a higher-level concept space instead of the original low-level feature space. Experimental results over a large-scale real consumer database confirm that the BOF representation can significantly reduce the complexity of the classification problem and provide better detection performance; also, the concept space significantly outperforms low-level visual feature space for our semantic event detection task.

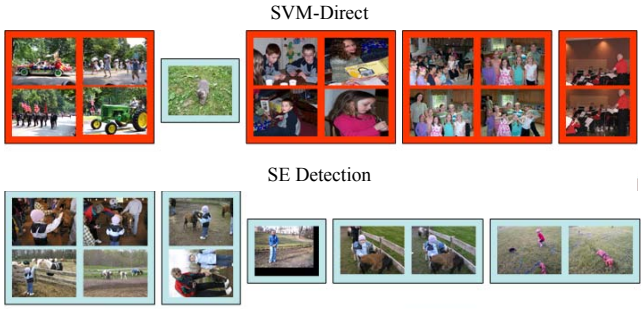

Fig. 6. Example of top-5 events by SE Detection and baseline. At most 4 photos or videos are shown for each event. Each event is cropped by one rectangular, green as correct and red as incorrect.

5. REFERENCES

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

[2] Shahram Ebadollahi and *et al.*, "Visual event detection using multi-dimensional concept dynamics," *IEEE ICME*, 2006.

[3] S.F. Chang and *et al.*, "Multimodal semantic concept detection for consumer video benchmark," *ACM MIR*, 2007.

[4] A.C. Loui and *et al.*, "Kodak consumer video benchmark data set: concept definition and annotation," *ACM MIR*, 2007.

[5] W. Jiang and *et al.*, "Similarity-based online feature selection in content-based image retrieval," *IEEE Trans. on Image Processing*, 15(3):702–727, 2006.

[6] J. Sivic and A. Zisserman, "Video google: a text retrieval approach to object matching in videos," *ICCV*, pp.1470-1477, 2003.

[7] A.Y. Ng, M. Jordan, and Y. Weiss, "On spectral clustering: analysis and an algorithm," *Advances in NIPS*, 2001.

[8] Y. Rubner, C. Tomasi, and L. Guibas, "The earth mover's distance as a metric for image retrieval," *IJCV*, 2000.

[9] NIST, "Trec video retrieval evaluation (trecvid)," 2001 - 2007, http://wwwnlpir.nist.gov/projects/trecvid.