Event Detection Based on Noisy Object Information

Yuzuko Utsumi, Misa Katte, Masakazu Iwamura and Koichi Kise
Graduate School of Engineering
Osaka Prefecture University
1-1, Naka, Sakai, Osaka 599–8531, Japan
Email: {yuzuko, masa, kise}@cs.osakafu-u.ac.jp, misakatte@gmail.com

***Abstract*—This paper proposes an event detection method using noisy object information. Some events have a close connection with objects, and the objects related to the event often appear with the event in a video. For example, if an event "Grooming an animal" appears in a video, an animal and people should appear in the video. If we detect the objects that have a close connection with the events, we can detect the events based on the object detection results. However, it is inevitable that the object detector gives false alarms and the object information would be noisy. Thus, we use the information about objects which is robust against false alarms of object detection. In our experiments, we evaluated how the information on objects is effective for event detection of videos. From the results, the proposed method showed better or equivalent detection results than a state-of-the-art method in some events, even the object detector gave false alarms.**

I. Introduction

If an event "changing a vehicle tire" happens in a video, it is natural to appear a tire in the video. If an event "making a sandwich" happens, it probably happens in a kitchen. As you can see from the exapmples above, events have a close connection with objects and backgrounds. In order to utilize the connection between events and objects for event detection in videos, we propose an event detection method from videos by using object and background information related to events. The primary contribution of this paper is to show the effectiveness of the use of object and background information (e.g., a tire and kitchen) in the event detection task in videos.

A lot of methods of event detection have been proposed in recent years. One of the main methods is using temporal information like optical flow, motion information and temporal state transition [1], [2], [3]. Motion information is effective for detecting events like human actions. Appearance information also has been used for event detection. Local features such as the SIFT [4] and the histogram of orientation (HOG) [5] features are used for describing video appearance. However, those local features just show gradient orientation of keypoints and more semantic features are needed to express events. In order to represent semantic information with local features, two approaches were proposed. One is making semantic local feature groups automatically by using learning methods and expressing events by the groups. Yang et al. [6] made semantic feature groups called "concepts" and described events by the concepts. The other is learning local features by objects and describing events by the object detection results. Object information has a strong connection with events as mentioned above. Moreover, object information is easy to understand for human. Thus, object information is used for event detection. Snoek et al. [7] used a pixel-wise object detection method for event detection. Li et al. [8] used a lot of object detectors and the outputs of the detector as features for scene detection from still images. In [9], [10], they employed the objects which appeared in a video as features to detect events. The drawback of the object information is the accuracy of the object detectors. It is inevitable for the object detectors to give false alarms, and false alarms affect the accuracy of event detection. In order to overcome the drawback, we use two object features which cope with the false alarms. One object feature is based on the number of objects detected in frame images. For example, a lot of people appear in an event "Parade." The detector detects more persons than other events. Another object feature is based on the confidence value of a detector. For example, in an event "Changing a vehicle tire", a tire can be detected as a car, bicycle, train and bus. The detection result seems useless. However, there is a tendency that a tire is misdetected as objects that contain tires and wheels. Thus the tendency can be a clue to detect the event. The background information is also used in the proposed method. Specifying backgrounds is useful for event detection. If an event happens in a kitchen, the event is estimated to be related to cooking.

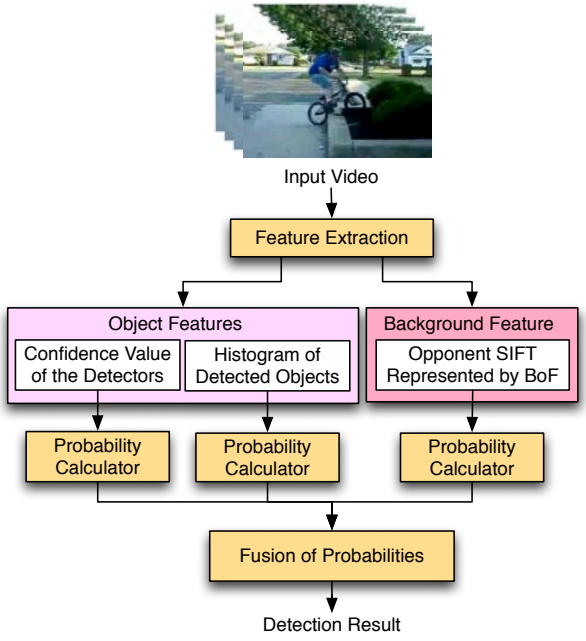


Fig. 1. The overview of the proposed event detection method.

In our experiments, we evaluated the proposed method

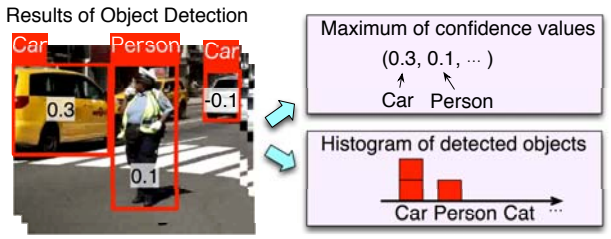


Fig. 2. Feature extraction.

by using TRECVID 2011 Multimedia event detection database [11]. From the results, the proposed method showed better average precision than the state-of-the-art method [3] in some events, that have close relationships with typical objects and backgrounds even the detector gave false alarms.

II. Overview of the Proposed Event Detection

The overview of the proposed event detection method is shown in Fig. 1. First, some frames are sampled from an input video, and three types of features are extracted from the sampled frames. The first two features represent object information, and are calculated by using the results of the object detection. The first feature is a vector consisting of the maximum confidence values of the object detectors. The second feature is a histogram of detected objects. The third feature represents background information; the third feature is a histogram of the Opponent SIFT features based on a bag-of-feature (BoF) model. After extracting features, a probability of the event occurrence is calculated for each feature independently. Finally, a detection result is determined based on the average of the three probabilities.

III. Feature Extraction

The proposed method extracts three types of features to represent videos: the maximum confidence values of the detectors, the histogram of detected objects and the histogram of the Opponent SIFT features based on the BoF model. In this section, we explain how the proposed method extracts features precisely.

A. Object features

Two object features are the maximum confidence values of the detector and the histogram of the detected objects. The outline of the object features is shown in Fig. 2. Object features are extracted by using the object detector. The scanning window approach is used for the detection. The proposed method applies the object detector to some frames from a video.

We employ the detector proposed in [12]. The detector consists of two kinds of filters; a root filter and a part filter. The root filter reacts to the shape of whole objects. The part filter reacts to the part of the objects and searches details of objects. The scores of the root and part filters are calculated as the dot products between a set of weights and the HOG features within a window. The confidence value of the detector is the sum of the score of the root and part filters and the score of the placement of each part relative to the root filter as shown in [12].

We use the maximum confidence values of the object detector for each object in the sampled frames as the first feature. Thus, the dimensionality of the first feature is the same as the number of learned objects. We use a histogram of detected objects as the second feature. The dimensionality of the second feature is also the same as the number of learned objects.

B. Background features

In the proposed method, we define a holistic feature of sampled images from a video as the background feature. We employ the Opponent SIFT features [13] with BoF representation to represent background information. The Opponent SIFT is the SIFT features extracted from images of the Opponent color space and the conventional paper [14] shows that the Opponent SIFT features are effective for scene detection of color videos. The images in the Opponent color space have three channels; one has red and green color information, one has yellow and blue color information and the other is equivalent to lightness of the HSV color space. Feature points are detected by the Harris Laplace detector and features are extracted around the detected feature points from three channels. The dimensionality of the Opponent SIFT features is 384.

After extracting the Opponent SIFT features, we describe the features by the BoF model. The Opponent SIFT features are represented by the frequencies of the visual words which are chosen from learning data beforehand.

IV. Probability of Event Occurrence

In this section, we explain how the probability of an event occurrence is calculated. The k-Nearest Neighbor (k-NN) search method is employed for calculating the probability. Let $\boldsymbol{v}_{max}$ be a vector of maximum confidence values of the object detector, $\boldsymbol{v}_{hist}$ be a histogram of detected objects, $\boldsymbol{v}_{sift}$ be a histogram of the Opponent SIFT features and $\boldsymbol{v} \in \{\boldsymbol{v}_{hist}, \boldsymbol{v}_{max}, \boldsymbol{v}_{sift}\}$ be a feature. The probability $P(E|\boldsymbol{v})$ that a feature $\boldsymbol{v}$ belongs to an event E is represented by

$$P(E|\boldsymbol{v}) = \frac{\#\mathrm{KNN}(E)}{k}, \tag{1}$$

where #KNN(E) stands for the number of features that belong to an event E in the k-nearest neighbors to the input feature and k is the parameter of the k-NN search different for each feature. After calculating the probabilities with three features, the average of three probabilities is calculated as a occurrence probability of the event E on the video.

V. Experimental Results

We executed event detection experiments to evaluate the proposed method. For the experiments, we used TRECVID 2011 Multimedia event detection (MED) database [11]. The videos in the database were collected from the web, and the resolution and lengths of the videos were different from each video. The database consists of two parts; one is the DEV-T set and the other is the DEV-O set. In the DEV-T set, there are 5 events which are "Attempting a board trick," "Feeding an animal," "Landing a fish," "Wedding ceremony" and "Working on a woodworking project." In the DEV-O set, there are 10

TABLE I. AVERAGE PRECISION OF DETECTION ON THE MED DEV-T DATASET.

Event Class	Tang et al. [3]	Our method
Attempting a board trick	15.44%	5.04%
Feeding an animal	3.55%	1.87%
Landing a fish	14.02%	5.41%
Wedding ceremony	15.19%	3.69%
Working on a woodworking project	8.17%	2.00%

TABLE II. AVERAGE PRECISION OF DETECTION ON THE MED DEV-O DATASET.

Event Class	Tang et al. [3]	Our method
Birthday party	4.38%	2.11%
Changing a vehicle tire	0.92%	1.53%
Flash mob gathering	15.29%	5.78%
Getting a vehicle unstuck	2.04%	3.89%
Grooming an animal	0.74%	1.5%
Making a sandwich	0.84%	1.46%
Parade	4.03%	5.48%
Parkour	3.04%	1.88%
Repairing an appliance	10.88%	1.75%
Working on a sewing project	5.48%	0.74%

events which are "Birthday party," "Changing a vehicle tire," "Flash mob gathering," "Getting a vehicle unstuck," "Grooming an animal," "Making a sandwich," "Parade," "Parkour," "Repairing an appliance," and "Working on a sewing project." Both the DEV-T and DEV-O sets consist of a learning set and test set. The learning set has about 150 videos for each event. The test set has a large number of videos; some contain the events in the DEV-T and DEV-O sets, and the others do not contain the events. The DEV-T set has 10402 videos, and the DEV-O set has 31820 videos as test sets. We constructed probability calculators with learning sets and evaluated the detection results with the DEV-T and DEV-O sets, respectively.

For learning the detectors, we used 21 objects which were "Aeroplane," "Bicycle," "Bird," "Boat," "Bottle," "Bus," "Car," "Cat," "Chair," "Cow," "Dining table," "Dog," "Horse," "Motorbike," "People," "Potted plant," "Sheep," "Sofa," "Train," "TV/monitor," and "Person". We used PascalVOC 2009 database [15] and INRIA Person Dataset [16] for learning the detector. We applied the detector to 3 images that were sampled from a video randomly. After that, the first and second features were extracted from the images. The histogram of the number of the detected objects were normalized before learning and classification. The dimensionality of both features was 21.

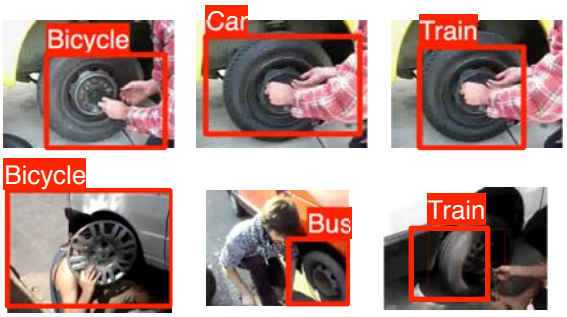
We constructed BoF model for describing the third feature. We sampled a frame for every 5 seconds from the test videos for feature extraction. We extracted the Opponent SIFT features from the learning sets of the DEV-T and DEV-O sets, and chose 3969 visual words according to [14]. We applied Principal Component Analysis (PCA) to the feature vectors that described based on the visual words, and reduced the dimensionality of the background feature to 3204. The parameter of the k-NN search method were selected with fifth of learning sets for each feature.

We compared the proposed method to the state-of-the-art method that uses motion features [3] in average precision (AP). We chose AP, which is different from the criteria of TRECVID 2012 MED Task, because we measured the performance in each video category. Let N be the number of samples in each video category, NR be the number of relevant samples in each video category and NR_l be the number of relevant samples found in the top l ranked samples by an event detection method. Let I_l be an indicator that is 1 if the lth sample is a relevant sample and is 0 otherwise. The AP defined as $(1/NR) \cdot \sum_{l=1}^{N}(I_l \cdot NR_l/l)$. The AP of the DEV-T and DEV-O set with the proposed method and state-of-the-art is shown in Table I and II. From Table I and II, the proposed method showed better average precision than the state-of-the-art method in some events. The events that the proposed method showed better average precision were "Changing a vehicle tire," "Getting a vehicle unstuck," "Grooming an animal," "Making a sandwich" and "Parade". Some results of the object detector with those events are shown in Fig. 3. From the Fig. 3 (b), in the images of the event "Getting a vehicle unstuck," cars were detected correctly. The event showed good detection results because the related objects were detected correctly. From Fig. 3 (a), (c), the objects were detected wrongly, however, the proposed method succeeded in detecting the events "Changing a vehicle tire" and "Grooming an animal." That is because the false alarms of the object detector had a tendency. In the event "Changing a vehicle tire," the object detector detected cars as car, bus and train and in the event "Grooming an animal," the object detector detected animals as dog, cat, horse and cow. Thanks to use the maximum confidence values of all objects as features, we can use the tendency to event detection effectively.

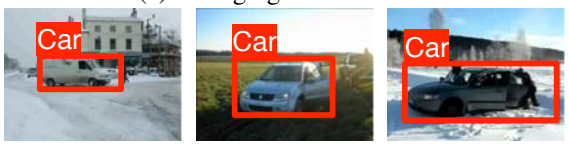
In the event "Parade," the histograms of the number of detected objects were effective for detection because a lot of people were appeared on the video of the event as shown in Fig. 4. Detection results contained false alarms. However, people were detected correctly and the number of detected people was good clue of detect the event.

In the event "Making a sandwich," the Opponent SIFT features were effective for detecting the event. Some frames of the event "Making a sandwich" are shown in Fig. 5. The event was always occurred in kitchens, and the Opponent SIFT features extracted information about kitchens.

On the other hand, the proposed method did not detect the events "Birthday party," "Flash mob gathering," "Parkour," "Repairing an appliance," "Working on a sewing project" and all the events of the DEV-T set. There are objects related to those events, however, the related objects were not learned by the object detector. If the detector learns some objects like birthday cakes and can detects the learned objects, the proposed method would succeed in detecting the events. However, some related objects like cloth are hard to detect because the shape is deformable. Other clues like motion features would be effective for the events that have undetectable related objects. The events "Flash mob gathering" and "Repairing an appliance" were also hard to detect for the proposed method because those events do not have any typical objects related to the events strongly. Motion information is more appropriate features to detect those events.



(a) Changing a vehicle tire



(b) Getting a vehicle unstuck

(c) Grooming an animal

Fig. 3. The object detection results on the events. Red rectangles show the regions that gave maximum confidence values of the detectors.

Fig. 4. The object detection results in the event "Parade". Red rectangles show the region detected as a person by the detector

VI. Conclusion

This paper proposed an event detection method based on noisy object information. In order to cope with false alarms of the object detector, the proposed method used the maximum confidence values of the detector and the histogram of detected objects as object features. Moreover, the Opponent SIFT features were used as background features. The k-NN search method was used for calculating the probabilities of event occurrences in each feature, and the event occurrence probability was given by the average of the probabilities. The proposed method showed better AP than the state-of-the-art method even the detector gave false alarms. The events the proposed method can detect have strong relations with specific objects and the backgrounds, and the specific objects were able to be detected by the detector.

In the future, we have two problems. One is to increase the number of the detectable objects. If we can achieve it, the proposed method would improve the event detection results in the events that have related objects to the events. The other is to use not only the proposed method but also the motion based detection method. The experimental results showed that the proposed method and motion information based method had different tendencies of detection. If we can use those methods properly, the detection accuracy would be better.

Fig. 5. The sampled frames from the videos with the event "Making a sandwich."

References

[1] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[2] F. Wang, Y. Jiang, and C. Ngo, "Video event detection using motion relativity and visual relatedness," in *Proc. of the 16th ACM international conference on Multimedia*, 2008, pp. 239–248.

[3] K. Tang, L. Fei-Fei, and D. Koller, "Learning latent temporal structure for complex event detection," in *Proc. of 2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1250–1257.

[4] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," in *International Journal of Computer Vision*, vol. 60, 2004, pp. 91–110.

[5] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, 2005, pp. 886–893.

[6] Y. Yang and M. Shah, "Complex events detection using data-driven concepts," in *Proc. of 12th Europian Conference on Computer Vision*, 2012.

[7] C. G. M. Snoek, M. Worring, and A. W. M. Smeulders, "Early versus late fusion in semantic video analysis," in *Proc. of the 13th annual ACM international conference on Multimedia*, 2005, pp. 399–402.

[8] L.-J. Li, H. Su, E. P. Xing, and L. Fei-Fei, "Object bank: A high-level image representation for scene classification & semantic feature sparsification," in *Proc. of the Neural Information Processing Systems (NIPS)*, 2010.

[9] M. Merler, B. Huang, L. Xie, G. Hua, and A. Natsev, "Semantic model vectors for complex video event recognition," *IEEE Transactions on Multimedia*, vol. 14, no. 1, pp. 88–101, 2012.

[10] L. Jiang, A. G. Hauptmann, and G. Xiang, "Leveraging high-level and low-level features for multimedia event detection," in *Proc. of the 20th ACM international conference on Multimedia*, 2012, pp. 449–458.

[11] P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, B. Shaw, W. Kraaij, A. F. Smeaton, and G. Quéenot, "TRECVID 2012 – an overview of the goals, tasks, data, evaluation mechanisms and metrics," in *Proc. of TRECVID 2012*, 2012.

[12] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[13] K. E. van de Sande, T. Gevers, and C. G. Snoek, "Color descriptors for object category recognition," in *Proc. of 4th European Conference on Colour in Graphics, Imaging, and Vision*, 2008, pp. 378–381.

[14] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek, "Evaluating color descriptors for object and scene recognition," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1582–1596, 2010.

[15] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal Visual Object Classes Challenge 2009 (VOC2009) Results," http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2009/workshop/index.html, 2009.

[16] N. Dalal, "INRIA person dataset," http://pascal.inrialpes.fr/data/human/.