2010

Event recognition based on top-down motion attention

Li Li  
*Chinese Academy of Sciences*

Weiming Hu  
*Chinese Academy of Sciences*

Bing Li  
*Chinese Academy of Sciences*

Chunfeng Yuan  
*Chinese Academy of Sciences*

Pengfei Zhu  
*Chinese Academy of Sciences*

*See next page for additional authors*

Follow this and additional works at: [https://ro.uow.edu.au/infopapers](https://ro.uow.edu.au/infopapers)

Part of the Physical Sciences and Mathematics Commons

**Recommended Citation**

Li, Li; Hu, Weiming; Li, Bing; Yuan, Chunfeng; Zhu, Pengfei; and Li, Wanqing: Event recognition based on top-down motion attention 2010.  

Research Online is the open access institutional repository for the University of Wollongong. For further information contact the UOW Library: research-pubs@uow.edu.au
Event recognition based on top-down motion attention

Abstract
How to fuse static and dynamic information is a key issue in event analysis. In this paper, a top-down motion guided fusing method is proposed for recognizing events in an unconstrained news video. In the method, the static information is represented as a Bag-of-SIFT-features and motion information is employed to generate event specific attention map to direct the sampling of the interest points. We build class-specific motion histograms for each event so as to give more weight on the interest points that are discriminative to the corresponding event. Experimental results on TRECVID 2005 video corpus demonstrate that the proposed method can improve the mean average accuracy of recognition.

Disciplines
Physical Sciences and Mathematics

Publication Details

Authors
Li Li, Weiming Hu, Bing Li, Chunfeng Yuan, Pengfei Zhu, and Wanqing Li

This conference paper is available at Research Online: https://ro.uow.edu.au/infopapers/3471
Event Recognition based on Top-Down Motion Attention

Li Li, Weiming Hu, Bing Li, Chunfeng Yuan, Pengfei Zhu and Wanqing Li†
National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
†SCSSE, University of Wollongong, Australia
Email: {lili,wmhu,cfyuan}@nlpr.ia.ac.cn and wanqing@uow.edu.au

Abstract

How to fuse static and dynamic information is a key issue in event analysis. In this paper, a top-down motion guided fusing method is proposed for recognizing events in an unconstrained news video. In the method, the static information is represented as a Bag-of-SIFT-features and motion information is employed to generate event specific attention map to direct the sampling of the interest points. We build class-specific motion histograms for each event so as to give more weight on the interest points that are discriminative to the corresponding event. Experimental results on TRECVID 2005 video corpus demonstrate that the proposed method can improve the mean average accuracy of recognition.

1. Introduction

Event detection or recognition is a key task in automatic video analysis, including semantic summarization, annotation and retrieval, and has received increasing attention in the past decades. Since 2001, the National Institute of Standards and Technology (NIST) has started benchmarking content-based-video retrieval technologies, known as TRECVID, in which event detection is one of the evaluation tasks. NIST provides a benchmark of annotated video corpus for detecting a set of predefined events. Despite much effort has been devoted to video based event recognition [9, 10, 11] and some success has been achieved, the problem is still far away from being solved. This is particularly due to the within-event variations caused by many factors, such as unconstrained motion, cluttered background, occlusion, environmental illumination and geometric variance of the objects involved in the events.

In a video clip, an event is usually has two important attributes: what and how. The what attribute refers to the appearance information which can be obtained from static images. SIFT feature by Lowe [6] has been proved to be an effective way to describe the static information due to their high performance and relatively easy to extract. Zhou [11] proposed a generative-to-discriminative framework by encoding each video clip as a bag-of-SIFT-features. On the other hand, the how attribute refers to the dynamic information of the event, which can be described by the motion of objects or subjects involved in the event. For instance, an event is modeled in [5] by the volumetric features derived from optical flow in a video sequence. However, how to effectively combine both what and how attributes is still a central task to any event recognition technique. To address this, a set of motion and bag-of-visual-words combination methods are proposed in [9] to exploit the relativeness of the motion information and the relatedness of the static visual information.

Figure 1. The block diagram of the proposed method

In this paper, we follow the principle of the top-down human visual system [4] and propose a method to combine the static visual and motion cues by selecting only a subset of information that is relevant to the events because we believe that not all interest points make the same contribution in recognizing different types of events. Some interest points may carry more information for a particular class of events. If we weight more in the recognition to the interest points that are highly rel-
event to the event, the recognition is expected to be improved compared with the case where all interest points are treated equally. To achieve this, we construct a class-specific motion histogram for each type of events.

The rest of the paper is organized as follows. Section 2 describes the proposed features and their extraction from a video clip. Section 3 presents the classifiers adopted in this paper. Experimental results are given in Section 4, followed by conclusions in Section 5.

2. Motion Attention based Bag of Words Representation (MA-BoW)

Bag-of-words has been proved to be a powerful tool for various image analysis tasks. In this paper, interest points are first detected for each frame in a video clip using the DoG method [7]. At each interest point, SIFT feature is extracted as the appearance descriptor and optical flow is estimated as the motion descriptor. Then, the appearance and motion vocabulary are constructed respectively through the k-means clustering algorithm. Each cluster is defined as a visual word. Fig. 1 shows the block diagram of the proposed method.

We employ Lucas and Kanade’s method implementation in OpenCV[1] to estimate the optical flow vector at each interest point. Fig. 2 shows a few examples of interest points and the estimated optical flow. Noticed that there are many noisy key points that are irrelevant to event. To reduce the influence of the noisy interest points, we propose the motion attention based representation for event recognition.

2.1 Motion Attention based Feature Representation

Inspired by Khan’s work combining shape and color cues for object recognition, we employ the top-down human visual attention mechanism to recognize events. As seen in Fig. 2, for a given event, not all interest points contribute equally in characterizing the event. For example, for “People Marching” event, only the interest points which do depict person and contain in the action “March” carry the information relevant to the event. The rest of interest points contributes little. To deal with this, we utilize learned class-specific motion information to construct an attention map of the corresponding event. SIFT feature will be used as a descriptive cue and motion feature as an attention cue. We build appearance and motion vocabularies independently and associate a visual word label and a motion word label with each interest point. Then, for a visual word v, a class-specific motion attention based Bag of Words histogram is calculated as:

\[ H_v^{Mag}(i) = \sum_{p \in N_v} P(i \mid m_p^{Mag}), i \in \text{Labels} \] (1)

\[ H_v^{Orient}(i) = \sum_{p \in N_v} P(i \mid m_p^{Orient}), i \in \text{Labels} \] (2)

where \( N_v \) is the collection of SIFT features which are mapped to the visual word \( v \) and \( m_p \) is the motion word of interest point \( p \), \( i \) is the category label of the event. \( H_v^{Mag} \) and \( H_v^{Orient} \) are computed based on the magnitude and orientation components of optical flow respectively. As seen in Eq. 1 and Eq. 2, the motion cue directly guides our prior knowledge about which event types we are looking for in a top-down manner. The probabilities \( P(i \mid m_p^{c_{motion}}) \), \( c_{motion} \in \{Mag, Orient\} \) can be computed by using the Bayesian rule,

\[ P(i \mid m_p^{c_{motion}}) \propto P(m_p^{c_{motion}} \mid i)P(i) \] (3)

where \( P(m_p^{c_{motion}} \mid i) \) is the prior empirical distribution. Given a motion word \( m_p \), the probability can be calculated by summing over the training videos containing this word and event class \( i \). The prior class probability \( P(i) \) can be obtained from the training data. Then \( P(i \mid m_p^{c_{motion}}) \) can be represented as

\[ P(i \mid m_p^{c_{motion}}) \propto \sum_{Video^i} \sum_{p} m_p^{c_{motion}} \] (4)

where \( Video^i \) refers to all of the video in the category \( i \). If we combine the Magnitude and Orientation

![Image 2](https://example.com/image2.jpg)

**Figure 2.** Examples of keypoints and the corresponding optical flow vector.
components of optical flow, the class-specific motion attention based Bag of Words histogram can be computed
\[ H_{p}^{MO}(i) = \sum_{p \in N_{v}} P(i \mid m_{p}^{Mag}) P(i \mid m_{p}^{Orient}), i \in Labels \] (5)

Notice from Eq. 1, 2 and Eq. 5 that the motion information is regarded as the weight of the SIFT features. In particular, for a given motion word, its probabilities for different events are different. For example, the motion words which depict the “Running” action should be paid more attention for event category “People running” than “People walking”. On the other hand, for some event recognition tasks such as “Riot” the motion information is irrelevant, the probability \( P(i \mid m_{p}^{motion}) \) is almost uniform.

3. Event Recognition

3.1 Similarity Between video clips

Given a video clip \( P \), once the motion attention based Bag of Words histogram is obtained between every two neighboring frames. \( P \) can be represented by a signature \( P = \{(p_{i}, w_{p_{i}}), 1 \leq i \leq m\} \), where \( p_{i} \) denotes Motion Attention based Bag of Words histogram extracted from the \( i \)th frame, \( w_{p_{i}} \) is the weight of frame \( i \), and satisfies \( \sum_{i=1}^{n} w_{p_{i}} = 1, 0 < w_{p_{i}} \leq 1 \), with default value being \( 1/m \). We employ the Earth Mover’s Distance (EMD)[8] to measure the distance between two video clips. EMD has been proved to effective in image retrieval and visual tracking because it can find optimal signature alignment. Moreover, the EMD based temporal matching method has outperformed in [10] the keyframe and multiframe-based methods by a large margin. For arbitrary two signatures \( P \) and \( Q \), \( P = \{(p_{i}, w_{p_{i}}), 1 \leq i \leq m\} \), \( Q = \{(q_{i}, w_{q_{i}}), 1 \leq i \leq n\} \), where \( m \) and \( n \) are the number of frames in video \( P \) and \( Q \), respectively. The EMD between video \( P \) and \( Q \) is computed by

\[ D(P, Q) = \sum_{i=1}^{m} \sum_{j=1}^{n} d_{ij} f_{ij} \] (6)

where \( d_{ij} \) is the Euclidean distance between \( p_{i} \) and \( q_{j} \), and \( f_{ij} \) is the optimal match between two signatures \( P \) and \( Q \) that can be computed by solving the Linear Programming problem.

3.2 Classifiers

For classification, we employ LibSVM [3] with “one against all” approach. The EMD distance between video clips is incorporated into the kernel function of the SVM classification framework by Gaussian function:

\[ K(P, Q) = exp(-\frac{1}{\lambda M} D(P, Q)) \] (7)

where \( M \) is a normalization factor which is the mean value of the EMD distances between all training samples. \( \lambda \) is a scaling factor which is empirically decided by cross validation.

4. Experiments

4.1 Data Sets

We conduct experiments on TRECVID 2005 video corpus to evaluate the performance of the proposed method. The ground truth is based on LSCOM annotated event concepts. After removing the events that has small number of positive samples, nine events were chosen as our evaluation set in this paper. Because the LSCOM[2] annotation labels are not sufficient for our dynamic concepts, we re-annotated the events by watching all frames in the video shot. As a result, there are 1671 positive clips for the nine events: Existing Car, Handshaking, Running, Demonstration Or Protest, Walking, Riot, Dancing, Shooting, People Marching. Half of the clips were used for training and the remaining for testing. The frame rate was down-sampled to five frames per second. We measured the performance using the Average Precision (AP) measure, which is the standard evaluation metric adopted in TRECVID. The Mean Average Precision (MAP) is defined as the mean of APs over all nine events.

4.2 Results

To verify the efficacy of the proposed features, we compared the video-based features rather than keyframe based algorithm. First, we compared Motion Attention based Bag of Words (MA-BoW) with static features-the Bag-of Visual-Word(BOW) that are widely used in concept detection. Then, we compared two ways of combining static and dynamic information methods: the Orientation Motion Histogram of Visual Words (OMH-BoW) proposed in [9] and Magnitude Motion Histogram of Visual Words(MMH-BoW).

For the OMH-BoW,

\[ OMH_{v}(i) = \sum_{p \in N_{v}} O_{i}(m_{p}), i = 1, 2, 3, 4, 5 \] (8)

where \( i \) is the bin number of Orientation component of motion vector, here we use 5 bins. Function \( O_{i}(\cdot) \) maps \( m_{p} \) to the \( i \)th direction.

For the MMH-BoW,

\[ MMH_{v} = \sum_{p \in N_{v}} M(m_{p}) \] (9)

where \( M(\cdot) \) is the magnitude of the corresponding visual word \( v \). Notice that each word bin is weighted by the magnitude.
The results on TRECVID have demonstrated the effectiveness of the proposed method.

References


Table 1. Comparison of Average Precision (%) using different features. BoW: Bag-of-Words; OMH-BoW: Orientation Motion Histogram of BoW; MMH-BoW: Magnitude Motion Histogram of BoW; MMA-BoW: Magnitude of Motion Attention based BoW; OMA-BoW: Orientation of Motion Attention based BoW; MOMA-BoW: Magnitude and Orientation of Motion Attention based BoW

<table>
<thead>
<tr>
<th>Event Name</th>
<th>BoW</th>
<th>OMH-BoW</th>
<th>MMH-BoW</th>
<th>MMA-BoW</th>
<th>OMA-BoW</th>
<th>MOMA-BoW</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exiting-car</td>
<td>34.8</td>
<td>30.0</td>
<td>15.8</td>
<td>38.1</td>
<td>40.2</td>
<td>37.1</td>
</tr>
<tr>
<td>Handshaking</td>
<td>50.9</td>
<td>47.0</td>
<td>36.1</td>
<td>46.6</td>
<td>47.1</td>
<td>47.1</td>
</tr>
<tr>
<td>Running</td>
<td>82.8</td>
<td>77.3</td>
<td>73.0</td>
<td>83.4</td>
<td>82.8</td>
<td>84.5</td>
</tr>
<tr>
<td>Demonstration-Protest</td>
<td>46.1</td>
<td>41.1</td>
<td>27.0</td>
<td>50.0</td>
<td>51.1</td>
<td>59.4</td>
</tr>
<tr>
<td>Walking</td>
<td>59.3</td>
<td>61.5</td>
<td>53.0</td>
<td>58.7</td>
<td>58.4</td>
<td>61.3</td>
</tr>
<tr>
<td>Riot</td>
<td>34.7</td>
<td>36.0</td>
<td>28.7</td>
<td>31.4</td>
<td>33.0</td>
<td>31.9</td>
</tr>
<tr>
<td>Dancing</td>
<td>31.1</td>
<td>33.4</td>
<td>20.6</td>
<td>45.0</td>
<td>45.0</td>
<td>47.4</td>
</tr>
<tr>
<td>Shooting</td>
<td>67.0</td>
<td>76.7</td>
<td>73.7</td>
<td>71.3</td>
<td>71.0</td>
<td>72.3</td>
</tr>
<tr>
<td>People-Marching</td>
<td>34.7</td>
<td>39.0</td>
<td>28.0</td>
<td>35.7</td>
<td>35.1</td>
<td>39.4</td>
</tr>
<tr>
<td>Mean Average Precision</td>
<td>49.0</td>
<td>49.1</td>
<td>39.6</td>
<td>51.1</td>
<td>51.5</td>
<td>53.4</td>
</tr>
</tbody>
</table>

Table 1 summarizes the experimental results for different features. From Table 1, it can be observed:

1. Among these features, the best performance gain is obtained by combining Magnitude and Orientation based motion attention (MOMA-BoW) with the highest MAP of 53.4%. The combination of the two attention cues, Magnitude and Orientation, can generally further improve the average precision compared with the single cue cases.

2. Compared with BoW, an improvement of 4.4% has been achieved. This may be due to the fact that BoW only captures the what attribute of an event and ignores the how attribute.

3. For the events such as Running, Demonstration-Protest, Dancing and People-Marching, the MMA-BoW, OMA-BoW and MOMA-BoW outperformed OMH-BoW and MMH-BoW. Especially, the motion attention based features are significantly better than OMH-BoW for the event Exiting-car, Running, Demonstration-Protest and Dancing. This verified that the motion attention based feature did guide the recognition in a top-down manner.

4. Our method did not perform as good as OMH-BoW for the events Riot and Shooting. This may be because that the motion information is not important in recognizing these events.

5. Disappointing results for MMH-BoW might be caused by the confusion of magnitude component of optical flow in this data set. In addition, for event Handshaking, static visual(BOW) alone performed better than the cases where motion features were also included due to the small motion of the event.

5 Conclusions

In this paper, we have proposed a top-down method to combine static and dynamic information based on the bag-of-words representation for event recognition. In particular, a class-specific motion attention based histogram is proposed. The results on TRECVID have demonstrated the effectiveness of the proposed method.