Event Detection from Video Using Segment-Based Approach

PHAN LE SANG

Recognizing event in unconstrained videos is one of the most important tasks in multimedia retrieval. It has many potential applications such as video indexing, searching, and event recounting. However, this is a challenging task due to the large content variation and uncontrolled capturing condition. This leads to the fact that these videos often contain irrelevant information to the event of interest. The straightforward way to solve this problem is to decompose the original video into smaller segments and build the event detectors from these segment representations. This dissertation follows the aforementioned direction to study event detection methods in real videos. Essentially, we study three complementary approaches including *feature representation*, *feature aggregation* and *feature learning*.

In the first approach, we propose to use the segment-based (**SB**) *feature representation* to overcome the limitation of the traditional video-based approach. In the video-based approach, local features are extracted from the entire video and then aggregated to form the final video representation. However, this video-based representation is ineffective when used for realistic videos because the video length can be very different and the clues to determine an event may happen in only a small segment of the entire video. To handle this problem, our segment-based divides the original videos into segments for feature extraction and classification, while still keeping the evaluation at the video level. We investigate several strategies to divide a video into segments including non-overlapping uniform segment sampling, overlapping uniform segment sampling, and segments that based on the shot boundary detection. We also study the optimal segment length for event detection, which is close to the mean average length of the training videos.

The second approach handles the aforementioned problem by proposing a new video pooling strategy for *feature aggregation*. We consider a video as a layered structure where the lowest layer are frames, the top layer is the entire video, and the middle layers are

the sequences of consecutive frames or the concatenation of lower layers. While it is easy to find local discriminative features in video from lower layers, it is non-trivial to aggregate these features into a discriminative video representation. In literature, people often use sum pooling to obtain reasonable recognition performance on artificial videos. However, the sum pooling technique does not work well on complex videos because the region of interests may reside within some middle layers. In this approach, we leverage the layered structure of video to propose a new video pooling method, named sum-max video pooling (**SM**), to handle this problem. Basically, we apply sum pooling at the low layer representation while using max pooling at the high layer representation. Sum pooling is used to keep sufficient relevant features at the low layer, while max pooling is used to retrieve the most relevant features at the high layer, therefore it can discard irrelevant features in the final video representation.

In the third approach, we focus on *feature learning* method to learn the key segments for video representation. In fact, a complex event can be recognized by observing necessary evidences. It is not easy to locate supportive evidences because they can happen anywhere in a video. A straightforward solution is to decompose the video into several segments and search for the evidences in each segment. This approach is based on the assumption that segment annotation can be assigned from its video label. However, this is a weak assumption because the importance of each segment is not considered. On the other hand, the importance of a segment to an event can be obtained by matching its detected concepts against the evidential description of that event. Leveraging this prior knowledge, we propose a new method, Event-driven Multiple Instance Learning (**EDMIL**), to learn the key evidences for event detection. We treat each segment as an instance and quantize the instance-event similarity into different levels of relatedness. Then the instance labels are learned by jointly optimizing the instance classifier and its related level. Finally the optimal instance classifiers are used to detect event.

We verify the effectiveness of our approaches on the large scale TRECVID Multimedia Event Detection 2010, 2011 and 2012 datasets. Our approaches can not only detect event, but also provide evidences for event detection. Compared to other segment-

based approaches, our solutions achieve significant improvements. For example, when comparing in the MED 2011 dataset with a same setting, the baseline method (traditional video-based approach) has the average precision of 6.74 %, while our methods (SB, SM and EDMIL) have the performance of 8.26 %, 6.92 % and 9.68 % respectively.