氏　　名	PHAN LE SANG
学位(専攻分野)	博士(情報学)
学 位 記 番 号	総研大甲第 1799 号
学位授与の日付	平成27年9月28日
学位授与の要件	複合科学研究科　情報学専攻 学位規則第6条第1項該当
学 位 論 文 題 目	Event Detection from Video Using Segment-Based Approach
論 文 審 査 委 員	主　査　教授　杉本　晃宏 教授　佐藤　いまり 准教授　Duy-Dinh Le 准教授　Gene Cheung 教授　佐藤　真一　国立情報学研究所

(別紙様式 2)
(Separate Form 2)

論文内容の要旨
Summary of thesis contents

Event Detection from Video Using Segment-Based Approach

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 segmentbased 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

(別紙様式 3)
(Separate Form 3)

博士論文の審査結果の要旨

Summary of the results of the doctoral thesis screening

Event Detection from Video Using Segment-Based Approach

本論文は、Event Detection from Video Using Segment-based Approach (セグメントに基づくアプローチによる映像からのイベント検出)と題し、映像中の複雑なイベントの検出技術について述べている。映像中のイベントとは主として映像中の人物の行動に基づくものであり、人物の動作、複数の人物間の相互作用、人物と物体や状況との相互作用などからなり、映像により記述される重要な情報である。従って、映像中のイベントの検出は、映像検索をはじめ様々な応用に必要不可欠な技術である。本論文は、映像を構成するセグメントに基づくイベント検出のアプローチについて検討しており、セグメントに基づく最適な映像表現方法、その表現に基づいて映像特徴量を構成するための統合手法、またセグメントに基づくイベント検出を最適化するための機械学習技術という三つの視点から広範囲な検討を行い、英文にてまとめている。

第一章 Introduction(序論)では、本研究の動機、対象とした映像からのイベント検出という問題、その課題、本論文の貢献についてまとめている。

第二章 Background(背景)では、本論文の背景となる事項についてまとめており、具体的に取り組んだ TRECVID マルチメディアイベント検出について説明した上で、映像処理、映像特徴量、それらを統合するエンコーディング技術、機械学習等、関連する研究領域についてまとめている。

第三章 Event Detection Using Segment-based Feature Representation(セグメントに基づく特徴表現によるイベント検出)では、セグメント表現に基づくイベント検出について検討しており、特にイベント検出性能を最適化するセグメント表現方式について網羅的に実験を行い、実際に高いイベント検出性能を達成している。本章の内容は、Pacific-Rim Conference on Multimedia (2012)並びに Journal of Signal Processing Systems (2014)にて発表している。

第四章 Event Detection Using Sum-Max Feature Aggregation(Sum-Max 特徴統合によるイベント検出)では、映像特徴記述の統合法について検討しており、画像特徴量や軌跡特徴量をセグメントにおいて統合する方法、セグメント特徴を映像において統合する方法の両段階において、Sum pooling（統合時に元の特徴記述すべての影響を考慮する）方法と Max pooling（統合時には元の特徴量のうち最も重要なもののみを考慮する）方法について、すべての組み合わせを検討し、セグメントレベルでは Sum pooling、映像レベルでは Max pooling を行う方法が最も適当であることを、包括的な実験に基づき解明している。本章の内容は International Conference on Image Processing (2014)にて発表している。

第五章 Event Detection Using Event-Driven Multiple Instance Learning(イベント駆動 Multiple-Instance 学習によるイベント検出)では、イベントを記述する言語表現が与えられている場合を想定し、各セグメントからディープラーニングに基づいて概念記述を求め、これらとイベント記述との類似度をディープラーニングに基づく単語埋め込み表現（ベクトル表現）を用いて求めた上、各セグメントをインスタンス、映像をバッグとみなした Multiple Instance Learning 技術によりセグメントごとにイベントに対する適合度を適応的に学習する方法を考案し、実際にイベント検出性能を顕著に向上させることに成功している。本章の内容は ACM Multimedia (2015)の Short paper（ポスター発表）として採択

(別紙様式 3)
(Separate Form 3)
済みである。
第六章 Conclusion(結論)にて本論文の成果をまとめている。
本論文で検討している、映像からのイベント検出の高度化に対し、セグメントに基づく映像表現、映像特徴の統合化、機械学習によるイベント検出の高精度化という三つの観点は、映像意味解析の高度化のためには極めて重要であり、得られた知見は研究コミュニティにも大変有用である。このように、本論文の映像イベント検出をはじめとする映像内容解析技術に関連する学術的・社会的貢献は少なくないと考えられる。