Exploring Semantic Concepts for Complex Event Analysis in Unconstrained Video Clips

Xiaojun Chang

Centre for Quantum Computation & Intelligent Systems

University of Technology, Sydney

A thesis submitted for the degree of

Doctor of Philosophy

July, 2016

I would like to dedicate this thesis to my loving parents and wife.

Acknowledgements

This thesis cannot be done without many important people in my work and life. My supervisor, Dr. Yi Yang, is the one who led me into the realm of machine learning and multimedia analysis. Under his supervision, I have learned a lot of theoretic knowledge and research skills. His passion for exploring new ideas and rigorous attention to detail affected me profoundly. Dr. Feiping Nie is my associate supervisor. I am always impressed by the way he thinks about every research problem. His self-motivation, hard-working spirit and rigorous research attitude have set an excellent example for me. My friend, Dr. Yao-Liang Yu, is a master of maths and statistics. I always admire how knowledgeable he is on maths and statistics. He has taught me a lot of mathematical theory and skills that are very useful in my work. Dr. Alexander G. Hauptmann was my supervisor when I visited Carnegie Mellon University. I have appreciated his guidance immensely.

I would also like to thank many mates at UTS and CMU. Zhongwen Xu, Yan Yan, Pingbo Pan and Linchao Zhu have helped me so much when I was in the group. Their sincerity and kind heart impressed me to give warm care to every other person. I will show my gratitude to Shoou-I Yu, Zhigang Ma and Xuanchong Li. Their acompany makes my research and life much easier.

Lastly, I would like to thank my wife and my parents for their continuous support for my academic pursuit. I wish them health and happiness.

Certificate of Original Authorship

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.
I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Student: Xiaojun Chang
13/07/2016

Abstract

Modern consumer electronics (*e.g*. smart phones) have made video acquisition convenient for the general public. Consequently, the number of videos freely available on the Internet has been exploding, thanks also to the appearance of large video hosting websites (*e.g*. Youtube). Recognizing complex events from these unconstrained videos has been receiving increasing interest in the multimedia and computer vision field. Compared with visual concepts such as actions, scenes and objects, event detection is more challenging in the following aspects. Firstly, an event is a higher level semantic abstraction of video sequences than a concept and consists of multiple concepts. Secondly, a concept can be detected in a shorter video sequence or even in a single frame but an event is usually contained in a longer video clip. Thirdly, different video sequences of a particular event may have dramatic variations.

The most commonly used technique for complex event detection is to aggregate low-level visual features and then feed them to sophisticated statistical classification machines. However, these methodologies fail to provide any interpretation of the abundant semantic information contained in a complex video event, which impedes efficient high-level event analysis, especially when the training exemplars are scarce in real-world applications. A recent trend in this direction is to employ some high-level semantic representation, which can be advantageous for subsequent event analysis tasks. These approaches lead to improved generalization capability and allow zero-shot learning (*i.e*. recognizing new events that are never seen in the training phase). In addition, they provide a meaningful way to aggregate low-level features, and yield more interpretable results, hence may facilitate other video analysis tasks such as retrieval on top of many low-level features, and have roots in object and action recognition.

Although some promising results have been achieved, current event analysis systems still have some inherent limitations. 1) They fail to consider the fact that only a few shots in a long video are relevant to the event of interest while others are irrelevant or even misleading. 2) They are not capable of leveraging the mutual benefits of Multimedia Event Detection (MED) and Multimedia Event Recounting (MER), especially

when the number of training exemplars is small. 3) They did not consider the differences of the classifier's prediction capability on individual testing videos. 4) The unreliability of the semantic concept detectors, due to lack of labeled training videos, has been largely unaddressed. To solve these challenges, in this thesis, we aim to develop a series of statistical learning methods to explore semantic concepts for complex event analysis in unconstrained video clips. Our works are summarized as follows:

In Chapter 2, we propose a novel semantic pooling approach for challenging tasks on long untrimmed Internet videos, especially when only a few shots/segments are relevant to the event of interest while many other shots are irrelevant or event misleading. The commonly adopted pooling strategies aggregate the shots indifferently in one way or another, resulting in a great loss of information. Instead, we first define a novel notion of semantic saliency that assess the relevance of each shot with the event of interest. We then prioritize the shots according to their saliency scores since shots that are semantically more salient are expected to contribute more to the final event analysis. Next, we propose a new isotonic regularizer that is able to exploit the constructed semantic ordering information. The resulting nearly-isotonic SVM classifier exhibits higher discriminative power in event detection and recognition tasks. Computationally, we develop an efficient implementation using the proximal gradient algorithm, and we prove new and closed-form proximal steps.

In Chapter 3, we develop a joint event detection and evidence recounting framework with limited supervision, which is able to leverage the mutual benefits of MED and MER. Different from most existing systems that perform MER as a post-processing step on top of the MED results, the proposed framework simultaneously detects high-level events and localizes the indicative concepts of the events. Our premise is that a good recounting algorithm should not only explain the detection result, but should also be able to assist detection in the first place. Coupled in a joint optimization framework, recounting improves detection by pruning irrelevant noisy concepts while detection directs recounting to the most discriminative evidences. To better utilize the powerful and interpretable semantic video representation, we segment each video into several shots and exploit the rich temporal structures at shot level. The consequent computational challenge is carefully addressed through a significant improvement of the current ADMM algorithm, which, after eliminating all inner loops and equipping novel closed-form solutions for all intermediate steps, enables us to efficiently process extremely large

video corpora.

In Chapter 4, we propose an **E**vent-**D**riven **C**oncept **W**eighting framework to automatically detect events without the use of visual training exemplars. In principle, zero-shot learning makes it possible to train an event detection model based on the assumption that events (*e.g.* birthday party) can be described by multiple mid-level semantic concepts (*e.g.* "blowing candle", "birthday cake"). Towards this goal, we first pre-train a bundle of concept classifiers using data from other sources, which are applied on all test videos to obtain multiple prediction score vectors. Existing methods generally combine the predictions of the concept classifiers with fixed weights, and ignore the fact that each concept classifier may perform better or worse for different subset of videos. To address this issue, we propose to learn the optimal weights of the concept classifiers for each testing video by exploring a set of online available videos which have free-form text descriptions of their content. To be specific, our method is built upon the local smoothness property, which assumes that visually similar videos have comparable labels within a local region of the same space.

In Chapter 5, we develop a novel approach to estimate the reliability of the concept classifiers without labeled training videos. The **EDCW** framework proposed in Chapter 4, as well as most existing works on semantic event search, ignore the fact that not all concept classifiers are equally reliable, especially when they are trained from other source domains. For example, "face" in video frames can now be reasonably accurately detected, but in contrast, the action "brush teeth" remains hard to recognize in short video clips. Consequently, a relevant concept can be of limited use or even misuse if its classifier is highly unreliable. Therefore, when combining concept scores, we propose to take their relevance, predictive power, and reliability all into account. This is achieved through a novel extension of the spectral meta-learner, which provided a principled way to estimate classifier accuracies using purely unlabeled data.

Contents

List of Figures

Publications

This thesis consists of the following publications:

- Chapter 2:
  - **X. Chang**, Y. Yang, E. P. Xing and Y. Yu: "Complex Event Detection using Semantic Saliency and Nearly-Isotonic SVM", International Conference on Machine Learning (ICML), 2015. *ERA Rank A**
- Chapter 3:
  - **X. Chang**, Y. Yu, Y. Yang, and A. Hauptmann: "Searching Persuasively: Joint Event Detection and Evidence Recounting with Limited Supervision", ACM International Conference on Multimedia (ACM MM), 2015. *ERA Rank A**
- Chapter 4:
  - **X. Chang**, Y. Yang, G. Long, C. Zhang and A. Hauptmann: "Dynamic Concept Composition for Zero-Example Event Detection", AAAI Conference on Artificial Intelligence (AAAI), 2016. *ERA Rank A**
  - **X. Chang**, Y. Yang, A. Hauptmann, E. P. Xing and Y. Yu: "Semantic Concept Discovery for Large-Scale Zero-Shot Event Detection", International Joint Conferences on Artificial Intelligence (IJCAI), 2015. *ERA Rank A**
- Chapter 5:
  - **X. Chang**, Y. Yu, Y. Yang, and E. P. Xing: "They Are Not Equally Reliable: Semantic Event Search using Differentiated Concept Classifiers", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. *ERA Rank A*

The following are the papers published during the course of the Ph.D but not included in this thesis:

- **X. Chang** and Y. Yang: "Semi-Supervised Feature Analysis by Mining Correlations among Multiple Tasks", IEEE Transactions on Neural Networks and Learning Systems, 2016. *ERA Rank A**

- **X. Chang**, Z. Ma, Y. Yang, Z. Zheng and A. Hauptmann: "Bi-Level Semantic Representation Analysis for Multimedia Event Detection", IEEE Transactions on Cybernetics, 2016. *ERA Rank A*
- **X. Chang**, F. Nie, Y. Yang, C. Zhang and H. Huang: "Convex Sparse PCA for Unsupervised Feature Analysis", ACM Transactions on Knowledge Discovery from Data, 2016. *ERA Rank B*
- M. Luo, F. Nie, **X. Chang**, Y. Yang, A. Hauptmann and Q. Zheng: "Avoiding Optimal Mean Robust PCA/2DPCA with Non-greedy L1-norm Maximization", International Joint Conferences on Artificial Intelligence (IJCAI), 2016. *ERA Rank A**
- **X. Chang**, F. Nie, Y. Yang and X. Zhou: "A Convex Formulation for Spectral Shrunk Clustering", AAAI Conference on Artificial Intelligence (AAAI), 2015. *ERA Rank A**
- **X. Chang**, F. Nie, Y. Yang, X. Zhou and C. Zhang: "Compound Rank-k Projections for Bilinear Analysis", IEEE Transactions on Neural Networks and Learning Systems, 2015. *ERA Rank A**
- Y. Yang, Z. Ma, F. Nie, **X. Chang** and A. Hauptmann: "Multi-Class Active Learning by Uncertainty Sampling with Diversity Maximization", International Journal of Computer Vision, 2015. *ERA Rank A**
- S. Wang, F. Nie, **X. Chang**, L. Yao, X. Li and Q. Zheng: "Unsupervised Feature Analysis with Class Margin Optimization", European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD), 2015. *ERA Rank A*
- **X. Chang**, F. Nie, Y. Yang and H. Huang: "A Convex Formulation for Semi-Supervised Multi-Label Feature Selection", AAAI Conference on Artificial Intelligence (AAAI), 2014. *ERA Rank A**