Zero-Example Multimedia Event Detection and Recounting with Unsupervised Evidence Localization

Yi-Jie Lu
Department of Computer Science, City University of Hong Kong
yijie.lu@my.cityu.edu.hk

ABSTRACT

Retrieval of a complex multimedia event has long been regarded as a challenging task. Multimedia event recounting, other than event detection, focuses on providing comprehensible evidence which justifies a detection result. Recounting enables "video skimming", which not only enhances video exploration, but also makes human-in-the-loop possible for improving the detection result. Most existing systems treat event recounting as a disjoint post-processing step over the result of event detection. Unlike these systems, this doctoral research aims to provide an in-depth understanding of how recounting, i.e., evidence localization, helps in event detection in the first place. It can potentially benefit the overall design of an efficient event detection system with or without human-in-the-loop. More importantly, we propose a framework for detecting and recounting everyday events without any needs of training examples. The system only takes a text description of an event as input, then performs evidence localization, event detection and recounting in a large, unlabelled video corpus. The goal of the system is to take advantage of event recounting which eventually improves zero-example event detection. We present preliminary results and work in progress.

Keywords

Multimedia Event Detection; Video Search; Zero Example; Event Recounting

1. INTRODUCTION

The exponential growth of web videos in the past decade has posed significant needs in developing broad video understanding techniques for content indexing and searching. Multimedia event detection is one of the challenging task intended for understanding high-level content in videos, which is typically characterized by objects, actions, activities, scenes, and the complex interactions between them [14]. Techniques for recognizing a complex multimedia event have been developing for years, and due to its difficulty, still attracting a lot of attention among researchers [25]. While event detection focuses on retrieval of videos that contain relevant event, event recounting complements event detection by providing comprehensible evidence that justifies a detection result. For example, what confides a video clip belonging to the event of "changing a car tyre"? One may expect appearances of a car, a zoomed-in shot of a tyre, and a lug wrench. However, any single appearance of such evidence does not necessarily justify the video clip. *Key evidence* such as an action of a person removing an old tyre with a wrench is crucial to discriminate a true positive. Notwithstanding the difference of importance, the evidence provides an explanation for a detection result, which can be used to enhance user experience for video browsing.

Most existing works treat event recounting as a disjoint post-processing step on top of the event detection result [16, 29]. Widely-used approaches in event detection such as average and max pooling take all sampled frames of a video into account [18, 19]. But the content relevant to an event in a video is usually scattered along the timeline [33]. Collecting information from all the frames would inevitably bring noise into the video representation. Therefore, we think that event recounting should benefit event detection in the first place by localizing key evidence in videos, thus forming a purified video representation. Although several works have already explored either ranking the importance of temporal regions [33], or jointly optimizing both detection and recounting [3], they need training examples to discover the discriminative evidence. In contrast, we conduct study under the zero-example scenario, which does not use training examples. We propose zero-example event detection with unsupervised evidence localization that helps improve the detection accuracy.

On the other hand, although there are known studies which automatically discover discriminative evidence given training examples, few provide insights on how evidence can help human make decisions. We stress the benefit of involving human-in-the-loop especially for zero-example event detection because it is an applicable way to quickly distinguish the *near-miss* videos – videos of a different event that are visually similar to the query event. Figure 1 illustrates examples of near-miss videos. We find near-miss videos widely exist in everyday events, but are difficult to be discriminated by a machine learner. While it is still possible to learn a specialized classifier to distinguish them, the fine details that can distinguish a near-miss video are trivial to learn and are not generalizable across events, not to mention it requires extensive human labeling to identify the near-miss

MM'16 October 15–19, 2016, Amsterdam, The Netherlands
© 2016 ACM ISBN 978-1-4503-3603-1/16/10 ...$15.00
DOI: http://dx.doi.org/10.1145/2964284.2971480

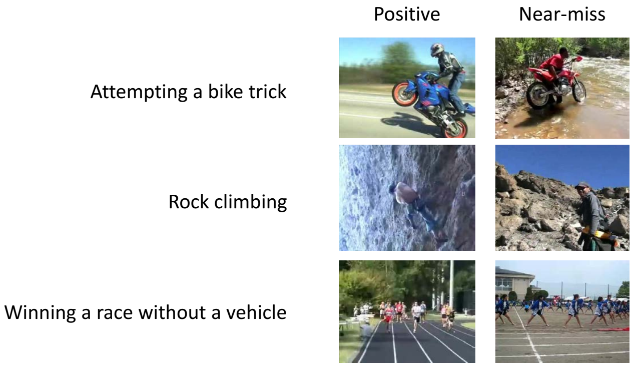

Figure 1: Examples of near-miss videos in different events. The near misses from top to bottom are: carrying a motorbike across the river, trekking over a rock hill, and doing setting-up exercise on the field.

videos. This problem becomes more crucial in zero-example scenario because a zero-example event detection system has to leverage a general knowledge base to relate textual information to visual information. Obviously, it is not feasible to prepare a knowledge base recognizing fine details that can distinguish the near-miss videos before hand. However, a human can identify decades of near-miss videos in a second as long as the evidence presented is sufficient for a human to make a decision. This renders the value of studying a better evidence presentation which eases human's judgement.

This doctoral research aims to have two outputs: First, an in-depth understanding of event recounting, both for how evidence localization can directly improve event detection performance, and for how evidence presentation can help a human make quick judgements. Second, an overall zero-example detection framework that can incorporate the benefit of event recounting with or without human-in-the-loop. The rest of the paper is organized as follows: Related work and contributions regarding the state of the art is discussed in the next section. Section 3 presents an overview of the proposed framework and approaches. Section 4 discusses the preliminary results and ongoing work.

2. STATE OF THE ART

Multimedia event detection usually needs training examples. Event detectors are trained by low-level features [7, 24, 30] or semantic features [9, 10, 26] that are directly extracted from training videos. *Zero-example* multimedia event detection (0Ex) is an emerging topic yet to be explored. In contrast to zero-shot learning which focuses on the recognition of images with unseen labels, 0Ex emphasizes the use of general and external knowledge for textual-to-visual relation. A decent 0Ex system usually consists of a semantic concept library as knowledge base, and a good search module that handles video ranking given a collection of concepts. A few pilot studies were proposed very recently [4, 31]. These works built a small concept library, typically hundreds of concepts, for textual-to-visual relation. More recent work starts to resort to a larger concept library. Ye et al. [32] collected a dataset with 500 events in which more than 4,000 concepts were hierarchically organized. Singh et al. [27] automatically discovered salient visual concepts by web search according to the text query. In event search phase, Habibian et al. [8] indexed concepts with AND/OR constraints. Chang et al. [2] proposed a rank aggregation framework that addressed the incomparable scales of scores when merging concepts from different feature spaces. Jiang et al. studied pseudo relevance feedback [12] and self-paced reranking [11] that further improved the performance by reranking. Jiang et al. [13] also systematically investigated 0Ex problem. This state-of-the-art work explored the contribution of multiple features including thousands of concepts, as well as the performance of several search models.

The primary concern of multimedia event recounting is the localization of key evidence that can discriminate an event. A related topic in recent research trends is semantic pooling [33], which suggests to only pool the evidential parts that are semantically important to an event query [20, 33]. The idea is based on the underlying assumption that important evidence is sparsely scattered in a video's timeline, thus aggregating all keyframes like average pooling [18, 19] would collect a bunch of junk information. Yu et al. [33] learned the concept importance of a small concept set and pooled the low-level features according to the importance of their related concepts. Mettes et al. [20] clustered the keyframes into fragment proposals and learned the importance of each proposal. Lately, more complex work tends to optimize the event detection by jointly discovering the evidence, given that a good recounting should assist detection in the first place [3, 28], rather than only interpret a detection result [6]. On the other hand, Bhattacharya et al. [1] conducted a user study and found that a human can recognize most events by only looking into very few sample segments of a video. Inspired by this finding, we develop unsupervised evidence localization which takes query words to locate key evidence in a video. We perform event detection merely on the evidence in order to improve the detection performance.

Our contributions relative to the state of the art are twofold:

- We propose unsupervised evidence localization under the framework of zero-example event detection, which is different from previous work that utilizes training examples. The evidence localization serves to improve event detection accuracy in the first place, rather than only explain the detection result.
- We provide insights on event recounting aiming to find a better evidence presentation that eases human's judgement.

3. PROPOSED APPROACH

Figure 2 illustrates the overall zero-example framework. The system only takes a text event query as input, then retrieves and recounts videos in a large, unlabelled video corpus. The core of the framework is a large concept bank containing concept detectors of objects, scenes, actions, and activities. The semantic concepts in the concept bank provide essential prior knowledge for text-to-visual relation.

In the offline phase, all frames of videos are uniformly extracted from the video corpus, and are represented by a concept-based representation in which each dimension represents the likelihood of presence of a particular concept. In the online phase, given an event query, e.g. "cleaning an appliance" with detailed text explanation, the noun and verb phrases are first extracted as tokens by an NLP parser. Then these tokens are mapped to an internal query representation called *semantic query* by concept matching. In

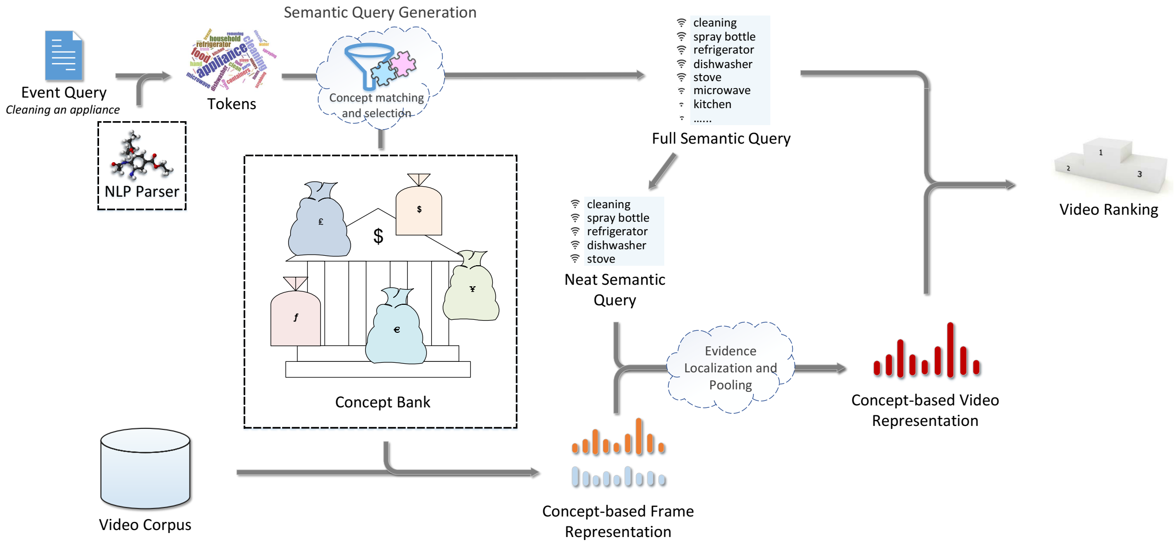

Figure 2: The overall framework for zero-example event detection with unsupervised evidence localization.

this way, the query is converted to the concept-based representation as well. Having the semantic query, we seek to derive a *neat semantic query* which only retains the most important relevant concepts. This neat semantic query is used for evidence localization in all videos, and then pooling over the evidence, thus forming a neat concept-based video representation. Finally, event search is performed having all the concept-based representations of both queries and videos.

There are two cruxes in this framework that have a large impact on the event detection performance. The first crux is the concept matching and selection in semantic query generation. We have conducted extensive study on how to pick up concepts in the concept bank for query representation [23, 17]. This includes various similarity measurements, such as TFIDF, WordNet [22] and Word2Vec [21]. Basically, the performance of TFIDF and Word2Vec similarities fall in the same scale. WordNet barely works in zero-shot event detection. We also find the selection of a few relevant concepts crucial to event detection performance [17]. The second crux is the representation of video by evidence. It includes an objective goal which aims to improve the detection accuracy and a subjective goal which aims to promote a better evidence presentation for human judgments. As this is an on-going study, we elaborate these two goals in the following sections.

3.1 Unsupervised evidence localization

Unsupervised evidence localization is intended to improve the performance of zero-example event detection. The localization process basically takes the neat semantic query as input and ranks the frames of a video according to their similarities to the query. The adjacent keyframes can be agglomerated to a shot if they are of similar importance. Then video representation is formed by average pooling over the evidence (keyframes or shots that are of top importance). Our preliminary study suggests that only three pieces of evidence are sufficient for a video representation that generates good performance (Table 1). This new representation has a notable benefit: It is robust to noisy concepts when using the full semantic query to perform the video search, which is the final step in Figure 2. In addition, this representation is not sensitive to the choice of the neat semantic query: We find the performance is relatively the same if simply choosing the top 2 to 16 concepts in the full semantic query as the neat semantic query.

We also study various methods for evidence localization. A typical variation is to cluster the frames of a video before hand, and rank the clusters by importance rather than rank the frames. Bisecting K-means is used to track each split of the clustering for analysis. We include the preliminary results in the experiment section, but the specific settings for clustering and pooling are not discussed in this paper.

3.2 Event recounting

Event recounting aims to provide a user with better understanding that can justify an event detection result. As discussed in Section 1, considering human-in-the-loop is an efficient way to address the near-miss problem. The key criterions to be considered here are the importance of evidence and user experience. Specifically, how fast a user can declare that a video is relevant by reading the presented evidence? General speaking, the goal is to display the minimal amount of evidence in the shortest possible time. To optimize both criterions, we will consider three aspects of information: (1) *concept-to-event relevancy* which prioritizes the importance of concepts to events, (2) *evidence diversity* which avoids redundancy by suggesting evidences of diverse content, and (3) *the shorter the better* which recommends only thumbnails that are just sufficient and necessary to evidence the presence of event. We will conduct user studies and design appropriate optimization algorithms on top of the evidence localization module that can integrate different criterions for event recounting.

4. EXPERIMENTS AND DISCUSSIONS

The experiments are conducted on the TRECVID Multimedia Event Detection (MED) datasets. We use the event kits that contain 20 event queries from E021 to E040. The

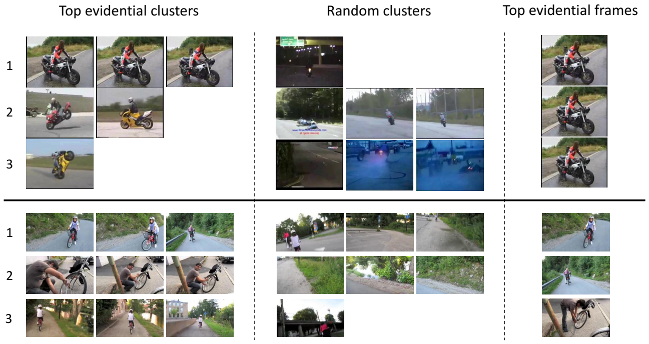

Figure 3: Visual examples of three evidential pooling methods for the event "attempting a bike trick". Two videos are listed: The one on the top is a true positive; bottom a false positive. Three pieces of evidence are presented for each video and for each method. Note that for clustering method, as the size of a cluster is varied, we present at most three frames for a cluster.

test set, named MED14Test, includes around 25,000 testing videos with no textual metadata. The performance is evaluated by the standard metric Mean Average Precision (MAP).

For preliminary experiments, we build a concept bank containing a total of 2,048 semantic concepts with varied granularity. They are sourced from off-the-shelf datasets with concept types covering most common objects, scenes, and actions. The four datasets are *ImageNet* 1000 concepts [5], *SIN* 346 concepts [34], *Places* 205 concepts [35] and *Research Collection* 497 concepts [23]. Except *ImageNet*, all the datasets are individually fine-tuned with AlexNet DCNN structure [15] on their own data. The concept responses are extracted for each frame uniformly sampled from a test video at the pace of two seconds one frame. The responses among different datasets are concatenated, forming a 2,048 dimensional feature vector. The individual performance of each dataset on MED14Test was reported in [17].

4.1 Obtained Results

Other than the results we previously reported [23, 17], the new results are mostly from the portion of the system discussed in Section 3.1. Table 1 summarizes the performance of various evidence localization settings. The best method is the video representation with only three evidential clusters, which improves average pooling by 88%. Surprisingly, the much simpler and time-efficient method using three evidential keyframes/shots without clustering can already achieve the MAP of 0.094. Meanwhile, we use the same number of random selected clusters as counterpart. The performance difference between evidential clusters and random selected clusters justifies the benefit of evidence localization. Figure 3 shows insights of three typical methods, which somewhat explains the result in Table 1. The key information for a human to justify a video is easily spotted in evidential clusters, while random clusters usually lie in segments not helpful to make a judgement. Evidential frames, on the other hand, may need more focus for a judge to make a decision than evidential clusters. It also has more tendencies to neglect key evidence due to imperfection of evidence localization.

Video Representation Method	*MAP*
3 evidential clusters	0.0979
3 evidential keyframes/shots	0.0936
10 evidential clusters	0.0900
10 evidential keyframes/shots	0.0879
Average pooling	0.0522
10 random clusters	0.0488
Max pooling	0.0485
3 random clusters	0.0461

Table 1: Performance summarization of different video representation methods on MED14Test.

The state of the art for automatic zero-example event detection has the MAP of 0.115 [13]. But different from our preliminary experiment settings, they used almost twice as many concepts, including event-level concept detectors such as Sports-1M. Sports-1M was proved to have significant contribution to the performance. However, due to the lacking of frame-level concept responses, we haven't yet included any event-level detectors into our concept pool for preliminary comparisons.

4.2 Work in Progress

Currently, we primarily focus on the research described in Section 3.2, that is to promote a better evidence presentation for involving human in the loop. This will include extensive user studies. We have already found that, although clustering does not significantly improve the detection performance, presenting the evidence in clusters can reduce both human's judgement time and judgement error. This is because clustering can help organize duplicate evidence, thus increasing the evidence diversity. We are still trying more ways of evidence presentation and comparing human's experience between the different ways.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China (No. 61272290), and the National Hi-Tech Research and Development Program (863 Program) of China under Grant 2014AA015102.

5. REFERENCES

[1] S. Bhattacharya, F. X. Yu, and S.-F. Chang. Minimally needed evidence for complex event recognition in unconstrained videos. In *ACM ICMR*, pages 105–112, 2014.
[2] X. Chang, Y. Yang, A. G. Hauptmann, E. P. Xing, and Y.-L. Yu. Semantic concept discovery for large-scale zero-shot event detection. In *IJCAI*, 2015.
[3] X. Chang, Y.-L. Yu, Y. Yang, and A. G. Hauptmann. Searching persuasively: Joint event detection and evidence recounting with limited supervision. In *ACM MM*, pages 581–590. ACM, 2015.
[4] J. Dalton, J. Allan, and P. Mirajkar. Zero-shot video retrieval using content and concepts. In *CIKM*, pages 1857–1860. ACM, 2013.
[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009.
[6] N. Gkalelis, V. Mezaris, I. Kompatsiaris, and T. Stathaki. Video event recounting using mixture subclass discriminant analysis. In *ICIP*, pages 4372–4376, Sept 2013.
[7] A. Habibian, T. Mensink, and C. G. Snoek. Videostory: A new multimedia embedding for few-example recognition and translation of events. In *ACM MM*, pages 17–26, 2014.

[8] A. Habibian, T. Mensink, and C. G. M. Snoek. Composite concept discovery for zero-shot video event detection. In *ACM ICMR*, pages 17–24, 2014.

[9] A. Habibian, K. E. van de Sande, and C. G. Snoek. Recommendations for video event recognition using concept vocabularies. In *ACM ICMR*, pages 89–96, 2013.

[10] L. Jiang, A. G. Hauptmann, and G. Xiang. Leveraging high-level and low-level features for multimedia event detection. In *ACM MM*, pages 449–458, 2012.

[11] L. Jiang, D. Meng, T. Mitamura, and A. G. Hauptmann. Easy samples first: Self-paced reranking for zero-example multimedia search. In *ACM MM*, pages 547–556, 2014.

[12] L. Jiang, T. Mitamura, S.-I. Yu, and A. G. Hauptmann. Zero-example event search using multimodal pseudo relevance feedback. In *ACM ICMR*, pages 297–304, 2014.

[13] L. Jiang, S.-I. Yu, D. Meng, T. Mitamura, and A. G. Hauptmann. Bridging the ultimate semantic gap: A semantic search engine for internet videos. In *ACM ICMR*, pages 27–34, 2015.

[14] Y.-G. Jiang, S. Bhattacharya, S.-F. Chang, and M. Shah. High-level event recognition in unconstrained videos. *International Journal of Multimedia Information Retrieval*, 2(2):73–101, 2013.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[16] K.-T. Lai, D. Liu, M.-S. Chen, and S.-F. Chang. Recognizing complex events in videos by learning key static-dynamic evidences. In *ECCV*, pages 675–688. Springer, 2014.

[17] Y.-J. Lu, H. Zhang, M. de Boer, and C.-W. Ngo. Event detection with zero example: Select the right and suppress the wrong concepts. In *ACM ICMR*, 2016.

[18] M. Mazloom, A. Habibian, and C. G. Snoek. Querying for video events by semantic signatures from few examples. In *ACM MM*, pages 609–612, 2013.

[19] M. Merler, B. Huang, L. Xie, G. Hua, and A. Natsev. Semantic model vectors for complex video event recognition. *IEEE Transactions on Multimedia*, 14(1):88–101, Feb 2012.

[20] P. Mettes, J. C. van Gemert, S. Cappallo, T. Mensink, and C. G. Snoek. Bag-of-fragments: Selecting and encoding video fragments for event detection and recounting. In *ACM ICMR*, pages 427–434, 2015.

[21] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119, 2013.

[22] G. A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, Nov. 1995.

[23] C.-W. Ngo, Y.-J. Lu, H. Zhang, T. Yao, C.-C. Tan, and L. Pang. VIREO-TNO @ TRECVID 2014: Multimedia event detection and recounting (MED and MER). In *NIST TRECVID Workshop*, 2014.

[24] S. Oh, S. Mccloskey, I. Kim, A. Vahdat, K. J. Cannons, H. Hajimirsadeghi, G. Mori, A. G. Perera, M. Pandey, and J. J. Corso. Multimedia event detection with multimodal feature fusion and temporal concept localization. *Mach. Vision Appl.*, 25(1):49–69, Jan. 2014.

[25] P. Over, G. Awad, M. Michel, J. Fiscus, W. Kraaij, A. F. Smeaton, G. Queenot, and R. Ordelman. TRECVID 2015 – an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *Proceedings of TRECVID 2015*. NIST, USA, 2015.

[26] B. Safadi, M. Sahuguet, and B. Huet. When textual and visual information join forces for multimedia retrieval. In *ACM ICMR*, pages 265–272, 2014.

[27] B. Singh, X. Han, Z. Wu, V. I. Morariu, and L. S. Davis. Selecting relevant web trained concepts for automated event retrieval. In *ICCV*, pages 4561–4569, Dec 2015.

[28] C. Sun and R. Nevatia. Discover: Discovering important segments for classification of video events and recounting. In *CVPR*, pages 2569–2576, June 2014.

[29] C.-Y. Tsai, M. L. Alexander, N. Okwara, and J. R. Kender. Highly efficient multimedia event recounting from user semantic preferences. In *ACM ICMR*, pages 419:419–419:422. ACM, 2014.

[30] F. Wang, Z. Sun, Y.-G. Jiang, and C.-W. Ngo. Video event detection using motion relativity and feature selection. *IEEE Transactions on Multimedia*, 16(5):1303–1315, Aug 2014.

[31] S. Wu, S. Bondugula, F. Luisier, X. Zhuang, and P. Natarajan. Zero-shot event detection using multi-modal fusion of weakly supervised concepts. In *CVPR*, pages 2665–2672, June 2014.

[32] G. Ye, Y. Li, H. Xu, D. Liu, and S.-F. Chang. Eventnet: A large scale structured concept library for complex event detection in video. In *ACM MM*, pages 471–480. ACM, 2015.

[33] Q. Yu, J. Liu, H. Cheng, A. Divakaran, and H. Sawhney. Semantic pooling for complex event detection. In *ACM MM*, pages 733–736, 2013.

[34] W. Zhang, H. Zhang, T. Yao, Y. Lu, J. Chen, and C.-W. Ngo. VIREO @ TRECVID 2014: instance search and semantic indexing. In *NIST TRECVID Workshop*, 2014.

[35] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *NIPS*, pages 487–495, 2014.