# R4-A.2: Rapid Similarity Prediction, Forensic Search & Retrieval in Video

## I. PARTICIPANTS

| Faculty/Staff | | | |
|---|---|---|---|
| **Name** | **Title** | **Institution** | **Email** |
| Venkatesh Saligrama | PI | BU | srv@bu.edu |
| David Castanon | PI | BU | dac@bu.edu |
| Hanxiao Wang | Post-doc | BU | hxw@bu.edu |
| **Graduate, Undergraduate and REU Students** | | | |
| **Name** | **Degree Pursued** | **Institution** | **Month/Year of Graduation** |
| Yannan Bai | MS | BU | 4/2018 |
| Di Wu | MS | BU | 12/2017 |

## II. PROJECT DESCRIPTION

### *A. Project Overview*

This project develops video analytics for maintaining airport and perimeter security. Our objectives include real-time suspicious activity detection, seamless tracking of individuals across sparse multi-camera networks, and the forensic search for individuals and activities in years of archived data. Surveillance networks are becoming increasingly effective in the public and private sector. Generally, use of these surveillance networks falls into a real-time or forensic capacity. For real-time use, the activities of interest are known *a-priori*, and the challenge is to detect those activities as they occur in the video. For forensic use, the data is archived until a user decides on an activity to search for. Forensic use calls for a method of content-based retrieval in large video corpuses based on user-defined queries. In general, identifying relevant information for tracking and forensics across multiple cameras with non-overlapping views is challenging. This is difficult given the wide range of variations, from the traditional pose, illumination, and scale issues to spatio-temporal variations of a scene itself.

It is worth focusing on the different goals of the forensic and real-time problem sets. In both problems, given the ubiquity of video surveillance, it is a fair assumption that the video to be searched grows linearly with time, and will stream in consistently. This mandates an ability to detect a predetermined activity in data as quickly as it streams in, for the real-time model. In the forensic model, this massive data requirement means that: (1) whatever representation is archived is computable as quickly as the data streams in, and (2) the search process scales sub-linearly with the size of the data corpus. The system will fall behind if it is not fulfilled. A user will have to wait too long for his/her results when searching a large dataset if it is not fulfilled.

The significance of a real-time activity monitoring effort to the Homeland Security Enterprise (HSE) is that these methods will enable the real-time detection of suspicious activities and entities throughout an airport by seamlessly tagging and tracking objects. Suspicious activities include baggage drops, behavior, and abandoning objects. The forensic search capability will significantly enhance current human-driven and relatively short horizon forensic capabilities, and allow for an autonomous search that matches user-defined activity queries in years of compressed data for detecting incidents such as a baggage drop, and identifying who/what was involved in that incident over large time-scales. Boston Logan International Airport (BOS) currently has the capability to store ~1 month's data, and much of the forensics requires significant human involvement. Our proposed research will generate new techniques for real-time activity recognition and tracking with higher probability of correct detection and reduced false alarms. Furthermore, it will enable the rapid search of historical video for the enhanced detection of complex activities in support of security applications.

### *A.1. Research Challenges*

1. Data lifetime: Since video is constantly streamed, there is a perpetual renewal of video data. This calls for a model that can be updated incrementally as video data is made available. The model must be capable of substantial compression for efficient storage. Our goal is to leverage the relatively stationary background and exploit dynamically changing traffic patterns to realize 1000X compression.
2. Unpredictable queries: The nature of queries depends on the field of view of the camera, the scene, the type of events being observed, and the user's preferences. The system should support queries of different natures that can retrieve both recurrent events, such as people entering a store, as well as infrequent events, such as abandoned objects or aimless lingering.
3. Unpredictable event duration: Within semantically equivalent events, there is significant variation. Events start anytime, vary in length, and overlap with other events. The system is nonetheless expected to return complete events regardless of their duration and whether or not other events occur simultaneously.
4. Clutter: Events in real surveillance videos rarely happen in isolation. Videos have a vast array of activities, so the majority of a video tends to be comprised of activities unrelated to any given search. This "needle in a haystack" quality differentiates exploratory search from many standard image and video classification problems.
5. Occlusions: Parts of events are frequently occluded, or do not occur. Trees, buildings, and other people often get in the way and make parts of events unobservable.

The challenges of search can be summarized as: (1) big data, (2) unknown query when the data arrives, (3) many false alarms, and (4) poor data quality. To tackle these challenges, we utilize a three-step process that generates a graphical representation of an activity, down-samples the video to the potentially relevant data, and then reasons intelligently over that data.

## *B. State of the Art and Technical Approach*

### *B.1. State of the Art*

We are presenting a survey of different approaches to the problems addressed by this project.

#### *B.1.a. Classification Methods*

Many methods [1-4] at run-time take a video-snippet (temporal video-segment) as input and outputs a score based on how well it matches the desired activity. During training, activity classifiers for video snippets are learned using fully labeled training data. In this context, several recent works have proposed deep neural network approaches for learning representations for actions and events [5, 6]. These works leverage the fact that, in some applications, object/attributes provide good visual signatures for characterizing activity.

In contrast to these methods, we do not utilize activity-level training data. Furthermore, while these methods are suited for situations where an activity manifests as a dominant signature in the video snippet, they are ill-suited for situations where the activity signature is weak, namely, the activity occurs among many other unrelated co-occurring activities, which is the typical scenario in surveillance problems.

### *B.1.b. Zero-shot Methods*

More recently, zero-shot methods have been applied to several visual tasks, such as event detection [7, 8], action recognition [9], and action localization [10]. These methods share the same advantage with our work in that activity level training data associated with the desired activity is not required. Nevertheless, zero-shot methods are trained based on source domain descriptions for a subset of activities that allow for forging links between activity components, which can then be leveraged for classification of unseen activity at test-time. Furthermore, the current set of approaches are only suitable in scenarios where the activity has strong visual signatures in low-clutter environments.

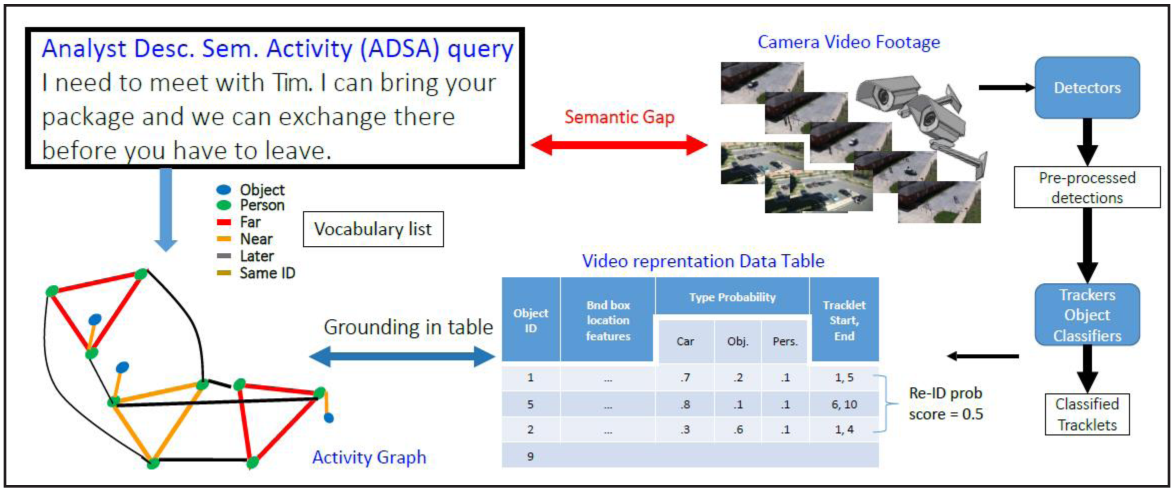
### *B.1.c. Activity Graphs*

It is worth pointing out that several works [3, 11] have developed structured activity representations, but they use fully annotated data as mentioned earlier. [3] describes a bipartite object/attribute matching method. [11] describes Boolean-Graphs based on aggregating sub-events for test-time activity recognition. Similar to classification-based methods, these methods only work well when the desired activity is dominant over a video snippet.

### *B.2. Salient Aspects of Our Approach*

We formulate a probabilistic activity graph that explicitly accounts for visual distortions, bridges the semantic gap through learning low-level concepts, and proposes an efficient probabilistic scoring scheme based on conditional random fields (CRFs). We propose to represent activities as graph queries. We utilize ground-truth data to reduce video to a large annotated graph. We propose probabilistic subgraph matching algorithms to probabilistically ground our CRF model in the video archive graph. We propose to handle the relationship semantic gap (for instance, nearness, proximity, etc.) through low-level learning. In this way, we account for visual distortions (mis-detections, track-loss etc.) and bridge the semantic gap between concepts and visual domain. We focus on retrieval of activity that matches analyst- or user-described semantic activity (ADSA) from surveillance videos. Surveillance videos pose two unique issues: (1) wide query diversity, and (2) the existence of many unrelated, co-occurring activities that share common components.

The wide-diversity of ADSAs limits our ability to collect sufficient training data for different activities and to learn activity models for a complete list of ADSAs. Methods that can transfer knowledge from detailed activity descriptions to the visual domain are required. To handle query diversity, we focus on a novel intermediate approach, wherein a user represents an activity as a semantic graph (see Fig. 1) with object attributes and inter-object semantic relationships associated with nodes and edges respectively. We propose to bridge the relationship semantic gap by learning relationship concepts with annotated data. At the object/node level, we utilize existing state-of-the-art methods to train detectors, classifiers, and trackers to obtain detected outputs, class-labels, track data, and other low-level outputs. This approach is practical because, in surveillance, the vocabulary of low-level components of a query are typically limited and can be assumed to be known in advance.


Analyst Desc. Sem. Activity (ADSA) query
I need to meet with Tim. I can bring your package and we can exchange there before you have to leave.
Semantic Gap
Camera Video Footage
Detectors
Pre-processed detections
Trackers Object Classifiers
Classified Tracklets
Object
Person
Far
Near
Later
Same ID
Vocabulary list
Grounding in table
Activity Graph
Video reprentation Data Table
Re-ID prob score = 0.5

| Object ID | Bnd box location features | Type Probability | | | Tracklet Start, End |
|---|---|---|---|---|---|
| | | Car | Obj. | Pers. | |
| 1 | ... | .7 | .2 | .1 | 1, 5 |
| 5 | ... | .8 | .1 | .1 | 6, 10 |
| 2 | ... | .3 | .6 | .1 | 1, 4 |
| 9 | | | | | |

**Figure 1: Illustration of building blocks for searching user defined activity in surveillance video.**

Our next challenge is to identify candidate groundings. By a grounding, we mean a mapping from archived video spatio-temporal locations to query nodes. Grounding is a combinatorial problem that requires searching over different candidate patterns that match the query. The difficulty arises from many unrelated, co-occurring activities that share node and edge attributes. Additionally, the outputs of low-level detectors, classifiers, and trackers are inevitably error-prone, leading to mis-detections, misclassifications, and loss of tracks. Additional uncertainties can also arise due to the semantic gap. Consequently, efficient methods that match the activity graph with high-confidence in the face of uncertainty are required.

We propose a novel probabilistic framework to score the likelihood of groundings that explicitly account for visual-domain errors and uncertainties. The probabilistic framework is based on a CRF model of semantic activity that combines the activity graph together with the confidence/margin outputs of our learned component-level classifiers, and calculates a likelihood score for each candidate grounding. We pose the combinatorial problem of identifying likely candidate groundings as a constrained optimization problem of maximizing precision at a desired recall rate. To solve this problem, we propose a successive refinement scheme that recursively attempts to find candidate matches at different levels of confidence. For a given level of confidence, we show that a two-step approach based on first finding subtrees of the activity graph that are guaranteed to have high-precision, followed by a tree based dynamic programming recursion to find the matches, leads to efficient solutions. Our method outperforms bag-of-objects/attributes approaches [3], demonstrating that objects/attributes are weak signatures for activity in surveillance videos unlike other cases [12, 13]. We compare against [3], which is based on manually encoding node/edge level relationships to bridge the visual domain gap and demonstrate that our semantic learning combined with probabilistic matching outperforms such methods.

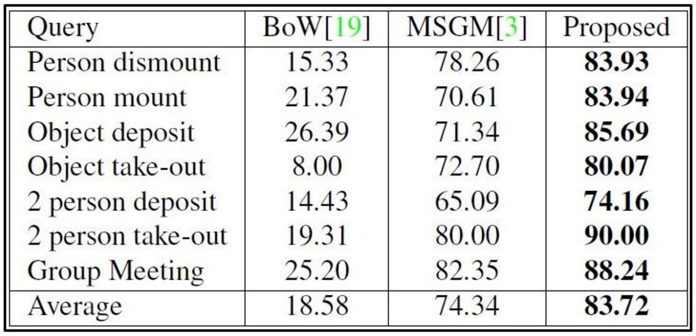
### *B.3. Experiments and Comparisons*

We performed semantic video retrieval experiments on the VIRAT Ground 2.0 dataset and the Air Force Research Laboratory (AFRL) benchmark Wide-Area Motion Imagery (WAMI) data. Given a set of activity graph queries, each algorithm is asked to return a ranked list of groundings in the archive video based on their likelihood scores. The minimal bounding spatio-temporal volume of the involved bounding boxes then represents each grounding. For the VIRAT dataset, where ground truth is provided, standard precision-recall curves are produced by varying the scoring threshold. For the unannotated AFRL data, a human operator evaluates the precision of top-k returns by watching the corresponding spatio-temporal window of the video. Each return is marked as a true detection if the overlap of the returned spatio-temporal volume and the true spatio-temporal volume is larger than 50% of the union.

We compared our performance with two approaches, a bag-of-words (BoW) scheme [3] and our previous Manually Specified Graph Matching (MSGM) scheme [12, 13], which was the approach we developed in Years 1-4. We first show the baseline performance of three methods on human-annotated track data from the VIRAT Ground 2.0 dataset with a set of seven queries. The VIRAT dataset is composed of 40 gigabytes of surveillance videos, capturing 11 scenes of moving people and vehicles interacting. Resolution varies, with about 50x100 pixels representing a pedestrian and around 200x200 pixels for vehicles. As shown in Tables 1 and 2, the proposed approach outperforms BoW and MSGM. On human annotated tracked data, where we assume no uncertainty at the object level, we can see that both MSGM and the proposed method significantly outperform BoW. The queries all include some level of structural constraints between objects; for example, there is an underlying distance constraint for the people, car, and object involved in object deposit.

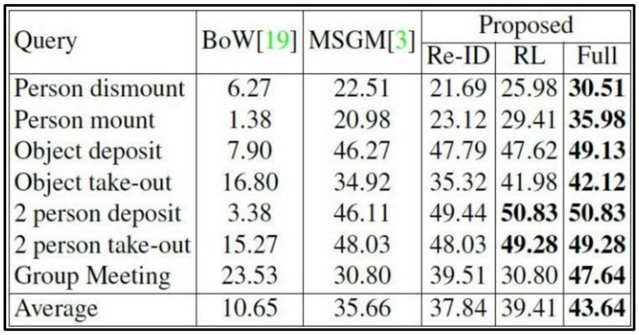
We performed an ablative analysis of our approach with detected and tracked bounding boxes in Table 2. To demonstrate the effect of re-ID and relationship learning, we report performance with only re-ID, only relationship learning, and both re-ID and relationship learning. Performance of all three methods degrade on tracked data due to missed detections/classifications and track errors. While our method degrades significantly, we still out-perform existing methods that train for an a priori known set of activities.

relationship learning, and both re-ID and relationship learning. Performance of all three methods degrade on tracked data due to missed detections/classifications and track errors. While our method degrades significantly, we still out-perform existing methods that train for an a priori known set of activities.



| Query | BoW[19] | MSGM[3] | Proposed |
|---|---|---|---|
| Person dismount | 15.33 | 78.26 | **83.93** |
| Person mount | 21.37 | 70.61 | **83.94** |
| Object deposit | 26.39 | 71.34 | **85.69** |
| Object take-out | 8.00 | 72.70 | **80.07** |
| 2 person deposit | 14.43 | 65.09 | **74.16** |
| 2 person take-out | 19.31 | 80.00 | **90.00** |
| Group Meeting | 25.20 | 82.35 | **88.24** |
| Average | 18.58 | 74.34 | **83.72** |

**Table 1: Area-Under-Curve (AUC) of precision-recall curves on VIRAT dataset with human annotated bounding boxes for BoW, MSGM, and our proposed approach.**



| Query | BoW[19] | MSGM[3] | Proposed | | |
|---|---|---|---|---|---|
| | | | Re-ID | RL | Full |
| Person dismount | 6.27 | 22.51 | 21.69 | 25.98 | **30.51** |
| Person mount | 1.38 | 20.98 | 23.12 | 29.41 | **35.98** |
| Object deposit | 7.90 | 46.27 | 47.79 | 47.62 | **49.13** |
| Object take-out | 16.80 | 34.92 | 35.32 | 41.98 | **42.12** |
| 2 person deposit | 3.38 | 46.11 | 49.44 | **50.83** | **50.83** |
| 2 person take-out | 15.27 | 48.03 | 48.03 | **49.28** | **49.28** |
| Group Meeting | 23.53 | 30.80 | 39.51 | 30.80 | **47.64** |
| Average | 10.65 | 35.66 | 37.84 | 39.41 | **43.64** |

**Table 2: Area-Under-Curve (AUC) of precision-recall curves on VIRAT dataset with automatically detected and tracked data for BoW, MSGM, and our proposed approach with only re-ID (Re-ID) with only relationship learning (RL), and the full system (Full) with both re-ID and RL.**

## *C. Major Contributions*

Our method for activity detection and similarity search in large surveillance video datasets is based on two aspects: (1) exploiting structural visual relationships to identify visual similarities to reduce visual ambiguity; and (2) developing a capability for inputting semantic queries that can interface with a visual domain. Unlike conventional approaches, our method is zero-shot, meaning it does not require knowledge of the activity classes contained in the video. Instead, we propose a user-centric approach that models queries through the creation of sparse semantic graphs based on attributes and discriminative relationships. Our eventual goal is to directly input textual descriptions of activity, and to search for activities in video that match these locations. We have posed search as a probabilistic grounding of a CRF model. Our key insight was to exploit the fact that the attributes and relationships in the query have different levels of discriminability to filter out bad matches.

The Forensic search system can be modularized into distinct components:

a. Query representation of activity models;
b. Re-identification, track maintenance, and target detection/recognition system;
c. A video archiving and hashing system with identified confidence values for targets and target-target relationships; and
d. A search system for identifying activity in video matching user defined input.

We first conceptualized a rudimentary system composed of (a), (c) and (d) described above. This system is based on the premise that we have access to high-performance algorithms for detection, recognition, and tracking. This rudimentary system is deterministic and sensitive to track and detection errors.

We next developed efficient subgraph matching techniques for matching graph-based input queries. In parallel, we developed re-identification algorithms for recovering targets. We developed novel structured prediction based Re-ID methods.

We built software for deterministic matching of input queries with a video archive for a limited vocabulary of target attributes and target-to-target relationships. This software demonstrated that under perfect tracking and detection, complex activities could be retrieved with high-accuracy.

To overcome sensitivity to track errors, we developed a probabilistic matching objective in order to account for detection and tracking errors. We then developed a probabilistic semantic retrieval system, which is robust when detection and track errors are small.

To deal with large track errors we incorporated Deep Neural Networks for low-level processing. This appears to significantly improve target detection performance. We integrated our methods into a fully functional Forensic system. The software is being tested for release.

### *D. Milestones*

Milestones achieved:

1. A graph representing user-defined activity with nodes representing objects and edges representing relationships between objects.
2. Search algorithm based on probabilistic grounding of CRF-based queries.
3. Incorporated object level mis-classifications and track-losses into the probabilistic search algorithm.
4. We tested our algorithm on both ground-level and airborne datasets. We demonstrated performance improvements in accuracy on both ground-truth and tracked data.
5. Improved recall-precision curves for simple and complex queries for tracked data through integration of re-ID and margin-based classifier scoring functions. Our goal is to realize over 15% area under curve (AUC) precision/recall improvement over conventional feature-accumulation algorithms.
6. Software integration of different modular components.
7. Investigated performance of textual queries to transition from the current graph-based query inputs.

### *E. Future Plans*

The work will reach the end-users as a software code to perform functions of interest. Our software will be capable of performing a number of video analytics functions that may be of interest to security or Transportation Security Administration (TSA) analysts. It will be capable of identifying target identity across multiple camera locations. It will be capable of descriptive or graph-based representations of suspicious activity as input, and searching against video feeds to identify time and locations of anomalous or suspicious activity.

We have validated our software on aerial and ground video in moderately cluttered environments. A major contribution of Year 5 has been to integrate low-level learning components, such as re-ID and object-level classifiers, into similarity search and updating the software to include these components. Currently, we see a significant performance gap between ground-truth and tracked data. We propose to improve our low-level object detectors, trackers, and re-ID algorithms, and to improve the integration of these methods into the similarity search algorithm. We propose to validate, test, and transition it to an airport scenario in Year 6, should funding become available. Our goal is to work with TSA to obtain representative video feeds to demonstrate the utility of our software.

## III. RELEVANCE AND TRANSITION

### *A. Relevance of Research to the DHS Enterprise*

DHS locations (i.e. airports, transit stations, border crossings, etc.) deploy thousands of cameras, which can be exploited to enable enhanced security. The relevant metrics for rapid similarity detection, forensic search, activity detection, and retrieval include: (1) negligible mis-identification probability of suspicious activity; (2) low computational complexity with respect to cameras and object density; and (3) large AUC for precision and recall for retrieving desired or suspicious activity.

### *B. Potential for Transition*

Interest has been expressed by BOS Massport for transition. Other airports and mass transit locations are also possible. We have developed a user-friendly software code for forensic search in Year 5. Our goal is to demo these capabilities to TSA and other interested parties during Year 6 so that they gain confidence with our activity detection and video analytics capabilities. Our software can also be used to enhance safety in other mass-transit scenarios. These capabilities are multi-use (i.e. they can be used by DHS, the National Geospatial-Intelligence Agency (NGA), or other Department of Defense (DoD) agencies).

### *C. Data and/or IP Acquisition Strategy*

A U.S. patent application for “Large Scale Video Search Using Queries that Define Relationships between Objects” filed through the Boston University (BU) patenting office was submitted in May 2017.

### *D. Transition Pathway*

We demonstrated capabilities of our software to ARO, ONR and NGA. The feedback we have received is that the software capabilities, when fully functional, would significantly enhance current security analyst capabilities.

Our recall rates are nearly 100% for simple and 80% for complex queries, but drops-off significantly to less than 50% with tracking and detection errors. One bottleneck is real-time computation that limits how well we can rule out false positives. We are working on robust search techniques to improve precision/recall rates.

Nevertheless, we must improve both the accuracy of our current method as well as improve user interface before it can be functional. We made significant improvements through tighter integration of state-of-the-art machine learning techniques with our video activity detection algorithms in Year 5. We plan to continue to demonstrate these capabilities to interested DHS parties in Year 6.

### *E. Customer Connections*

1. Martin Kruger, ONR Program Office, monthly meetings. Funding provided through ONR program.

2. Jason Schwendenmann, NGA Program, quarterly meetings. Funding through National University Research Initiative (NURI) Program.
3. Contacts with TSA at CLE and BOS along with RPI and NEU groups.

## IV. PROJECT ACCOMPLISHMENTS AND DOCUMENTATION

### *A. Education and Workforce Development Activities*

1. Student internships, job, and/or research opportunities
    a. Tolga Bolukbasi is joining Google Brain as an Applied Scientist.
    b. Yuting Chen has taken a research position at Adobe.

### *B. Peer Reviewed Journal Articles*

1. Zhang, Z., Liu, Y., Chen, X., Zhu, Y., Cheng, M.M., Saligrama, V., & Torr, P. "Sequential Optimization for Efficient High-Quality Object Proposal Generation." IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018.

**Pending-**

2. Y. Chen, J. Wang, Y. Bai, G. Castanon, V. Saligrama, Probabilistic Semantic Retrieval for Surveillance Videos with Activity Graphs, https://arxiv.org/abs/1712.06204. In revision for IEEE Transactions on Multi-Media)

### *C. Peer Reviewed Conference Proceedings*

**Pending-**

1. Zhu, P., Wang, H., Bolukbasi, T., Saligrama, V. "Zero-Shot Detection." Computer Vision and Pattern Recognition. Submitted March 19, 2018. https://arxiv.org/abs/1803.07113.

### *D. Student Theses or Dissertations Produced from This Project*

1. Bai., Y. "Video Analytics System for Surveillance Videos." MS Thesis, Electrical and Computer Engineering, Boston University, May 2018.

### *E. Software Developed*

1. Software for forensic search that uses a GUI interface input user to describe semantic activity by means of a graph representation, and performs probabilistic matching of user described activity video feed.

## V. REFERENCES

[1] K. Tang, L. Fei-Fei, and D. Koller. Learning latent temporal structure for complex event detection. In Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, pages 1250–1257. IEEE, 2012.

[2] Y. Yang, Z. Ma, Z. Xu, S. Yan, and A. G. Hauptmann. How related exemplars help complex event detection in web videos? In Proceedings of the IEEE International Conference on Computer Vision, pages 2104–2111, 2013

[3] D. Lin, S. Fidler, C. Kong, and R. Urtasun. Visual semantic search: Retrieving videos via complex textual queries. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2014.

[4] Z. Ma, Y. Yang, N. Sebe, and A. G. Hauptmann. Knowledge adaptation with partially shared features for event detection using few exemplars. IEEE transactions on pattern analysis and machine intelligence, 36(9):1789–1802, 2014

[5] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In NIPS, 2014.

[6] Z. Xu, Y. Yang, and A. G. Hauptmann. A discriminative cnn video representation for event detection. In CVPR, 2015

[7] 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, 2014.

[8] 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 AAAI, pages 2234–2240, 2015.

[9] Gan, M. Lin, Y. Yang, Y. Zhuang, and A. G. Hauptmann. Exploring semantic inter-class relationships (sir) for zeroshot action recognition. In AAAI, pages 3769–3775, 2015.

[10] M. Jain, J. C. van Gemert, T. Mensink, and C. G. M. Snoek. Objects2action: Classifying and localizing actions without any video example. In In ICCV, December 2015

[11] T. Shu, D. Xie, B. Rothrock, S. Todorovic, and S. Chun Zhu. Joint inference of groups, events and human roles in aerial videos. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015.

[12] G. D. Castanon, A. L. Caron, V. Saligrama, and P.M. Jodoin. Exploratory search of long surveillance videos. In Proceedings of the 20th ACM international conference on Multimedia, pages 309–318. ACM, 2012.

[13] G. D. Castanon, Y. Chen, Z. Zhang, and V. Saligrama. Efficient activity retrieval through semantic graph queries. In ACM Multimedia, 2015.

*This page intentionally left blank.*