Foundations for Summarizing and Learning Latent Structure in Video

Presenter: Kevin Pitstick, MTS – Engineer

PI: Ed Morris, MTS – Senior Engineer

Problem

DoD Operational Deficiency

- Volume of streaming and archived surveillance video is outpacing the ability of analysts (e.g. USAF Tactical PED teams) to manually monitor and view it
- Our collaborators, Darrell Lochtefeld and Daniel Zelik from AFRL's Human-Centered ISR Division, confirmed there is a lack of automated tools to assist analysts in monitoring real-time video or analyzing archived video
- Issue acknowledged by Project Maven, an initiative started by past Deputy Defense Secretary Bob Work to apply computer vision analysis to DoD video
  - Our approach is a candidate algorithm for consideration

Solution

Background: Video Summarization

- Computer vision task to condense a long video into a shorter “trailer” which contains the key or unique segments
- Various techniques: (1) key frames, (2) key sub-shots (whole frames), (3) key objects

Key Object-Motion Clip Video Summarization

We propose a new video summarization task that aims to generate video summaries based on the key objects in motion

The summaries should answer the following questions:

1. What are the representative objects residing in the video?
2. What attractable actions of these objects are occurring in the video?

Approach

Object-Level Video Summarization via Online Motion Auto-Encoder

Design and prototype a novel unsupervised video summarization pipeline which functions on extracted clips of objects in motion

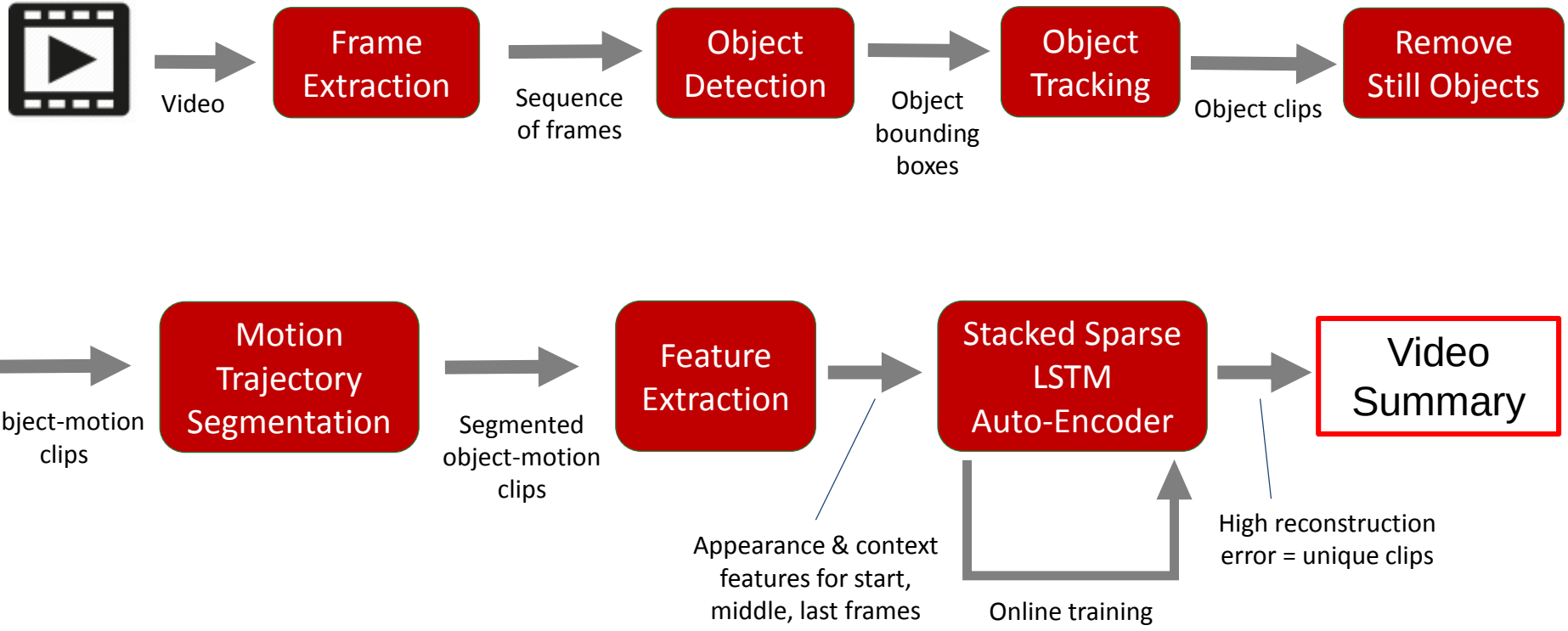
1. Extract clips of objects in motion from video
   - Object detection, object tracking, and frame segmentation
2. Feed each object clips' features through auto-encoder
   - Auto-encoder attempts to reconstruct the input
3. Clips with highest reconstruction error (adjustable threshold) become the summary

Key Contributions

1. Utilizing key object motion clips to depict whole video and generate video summaries
2. Unsupervised online motion auto-encoder model – encode and learn object motion patterns

CMU Machine Learning Dept. Collaborators: Xiaodan Liang and Eric Xing

Video Summarization Pipeline



Experiments

- **Datasets:** Orangeville (new), Base Jumping, SumMe, TVSum
- **Key Metrics:** Area under ROC curve (AUC), Average Precision (AP), F-measure (at threshold = 0.5)
- **Object-level:** Orangeville, **Subshot-level:** Base Jumping, SumMe, TVSum

Original: 100 seconds From “Orangeville” dataset (described in paper submission) **Summary:** ~17 seconds

Orangeville Results

Quantitative - Table 1

- Ground-truth annotated manually for key clips (fast moving cars, people crossing road, cars turning)
- Comparison with competing unsupervised, online approaches: sparse coding, alternate auto-encoders

Qualitative – Figure 1

- 15 subjects watching original at 3x speed followed by summary
- Assign rating from 1 to 10

	Sparse Coding	Stacked Sparse Auto-encoder	Stacked LSTM Auto-encoder	Stacked Sparse LSTM Auto-encoder (OURS)
AUC Score	0.4252	0.4354	0.5680	**0.5908**
AP Score	0.1542	0.1705	0.2638	**0.2850**
F-measure	0.1284	0.1662	0.2795	**0.2901**

Table 1: Object-level summarization results between competing approaches on **Orangeville** dataset

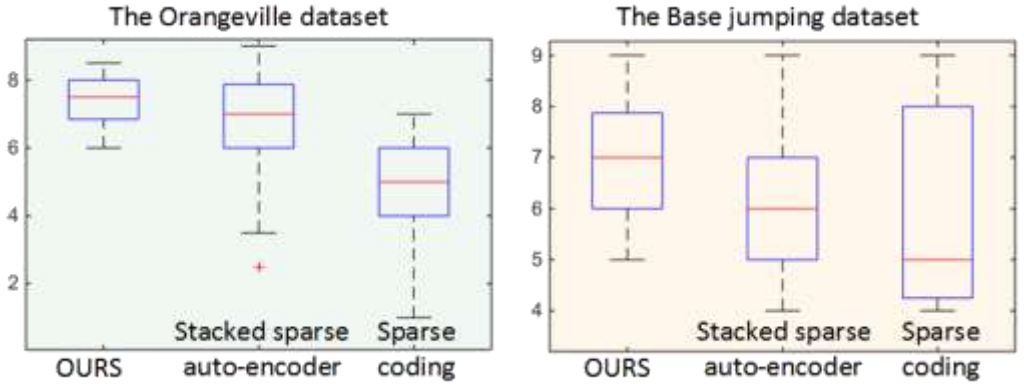


Figure 1: User study evaluation scores between competing approaches on **Orangeville** and **Base Jumping** datasets

SumMe and TVSum Results

- Adapt pipeline for subshot-level summarization to compare our auto-encoder against subshot-level approaches (e.g., TVSum, LiveLight, etc)

Method	F-measure
Video MMR	0.266
TVSum	0.266
$VSUMM_1$	0.328
$VSUMM_2$	0.337
Stacked GRU Auto-Encoder	0.354
Online Motion AE (OURS)	**0.377**

Table 1: Subshot-level summarization results on **SumMe** dataset

Method	F-measure
Web Image Prior	0.360
LiveLight	0.460
TVSum	0.500
Stacked GRU Auto-Encoder	0.510
Online Motion AE (OURS)	**0.515**

Table 2: Subshot-level summarization results on **TVSum** dataset

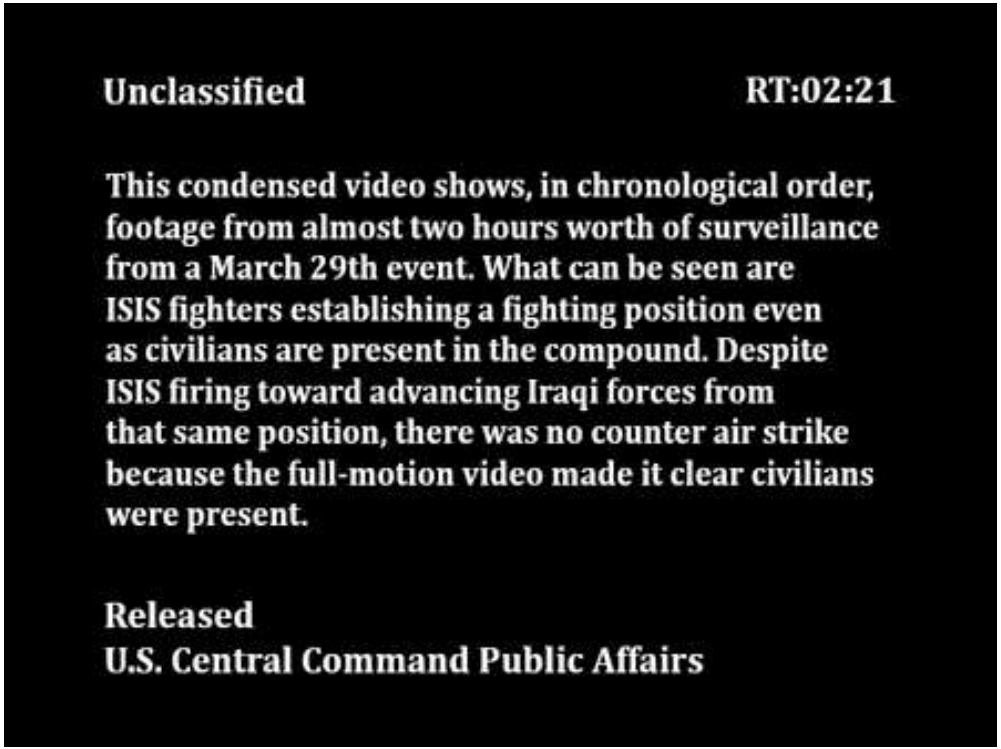
Analyzing DoD Full Motion Video (FMV)

While results are promising, DoD full motion video (FMV) differs from ground surveillance

- Infra-red (IR) vs electro-optical (EO) switches
- Moving camera vs. stationary camera
- Aerial viewpoint vs. ground viewpoint
- Changing zoom levels and rapid panning

AFRL Human-Centered ISR Division Collaboration

Darrell Lochtefeld and Daniel Zelik

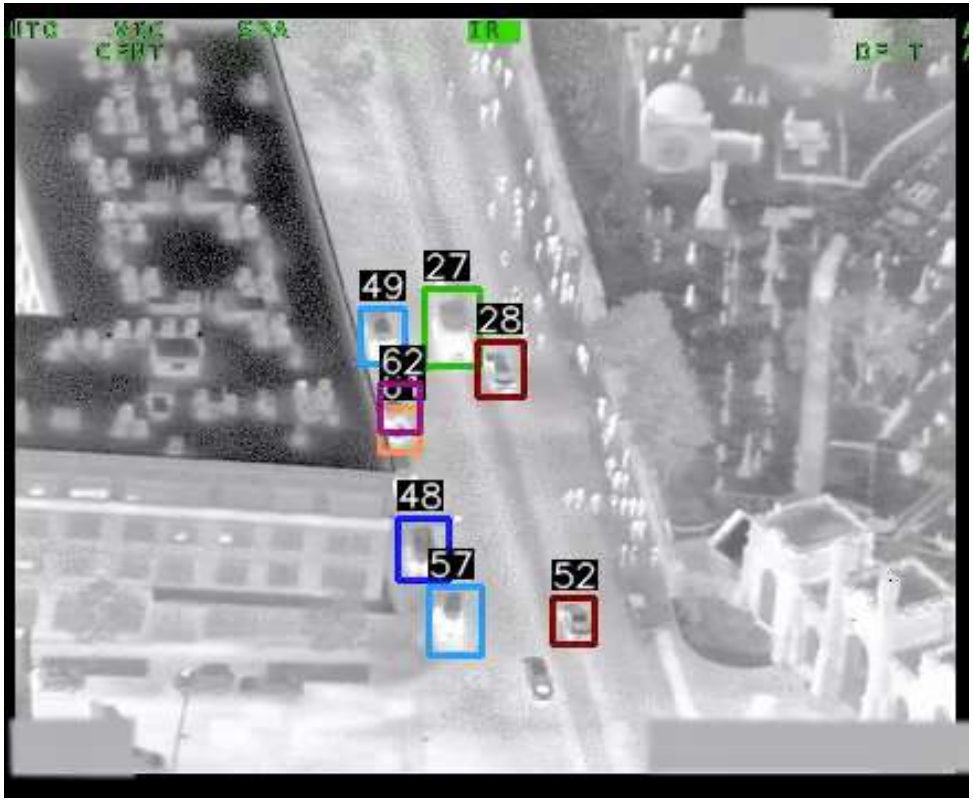


Publicly released by U.S. Central Command Public Affairs on CENTCOM's website - http://www.centcom.mil/MEDIA/VIDEO-AND-IMAGERY/VIDEOS/videoid/520438/

FBI Surveillance Video

Using FBI video of protests in Baltimore as first aerial surveillance dataset

- Labeled ~300 images with ground-truth vehicle annotations
- “Fine-tune” ImageNet object detection model to detect IR vehicles
  - Replace classifier layer and retrain it with 300 labeled images
- Detection model’s average precision: **0.89**



“Protests in Baltimore, Maryland 2015, Aerial Surveillance Footage.” FBI Records: The Vault. https://vault.fbi.gov/protests-in-baltimore-maryland-2015/unedited-versions-of-video-surveillance-footage

Project Artifacts

- **Software**
  - Prototype utilizing the pipeline for unsupervised, online, object-level video summarization
  - Video Markup Tool for annotating spatial-temporal object clips within video
- **Paper**
  - Submission to IEEE Transactions on Cybernetics: “Unsupervised Object-Level Video Summarization with Online Motion Auto-Encoder”
- **Dataset**
  - “Orangeville” benchmark for object-level summarization – dataset and annotations
  - Annotations and model for detecting vehicles in infra-red (IR) surveillance data released by FBI

Conclusion

Summary

- *Problem:* Lack of automated tools to assist analysts in processing the increasing volume of DoD surveillance video
- *Goal:* Investigate applying object-level video summarization techniques to identify key clips occurring in video
- *Results:* Video summarization pipeline meets or exceeds competing algorithms on benchmark datasets and should adapt to FMV aerial datasets

Future Work – FY18 Project: Summarizing and Searching Video

- Finish investigation of applying current pipeline to summarization of FMV datasets
- Unsupervised activity clustering – utilize object-motion clips as basis
- AFRL collaboration to explore applying analysis techniques to existing DoD problems
  - e.g., Nothing Significant to Report (NSTR) task

Contact Information

Presenter
Kevin Pitstick
MTS – Engineer
Email: kapitstick@sei.cmu.edu

Primary Investigator
Ed Morris
MTS – Senior Engineer
Email: ejm@sei.cmu.edu

SEI Team
- Jeffery Hansen
- Mark Klein
- Mike Konrad
- Keegan Williams
- Javier Vazquez-Trejo

CMU Machine Learning Department Collaborators
- Eric Xing, Professor
- Xiaodan Liang, Postdoctoral

AFRL Human-Centered ISR Division Collaborators
- Darrell Lochtefeld
- Daniel Zelik

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

DM17-0793