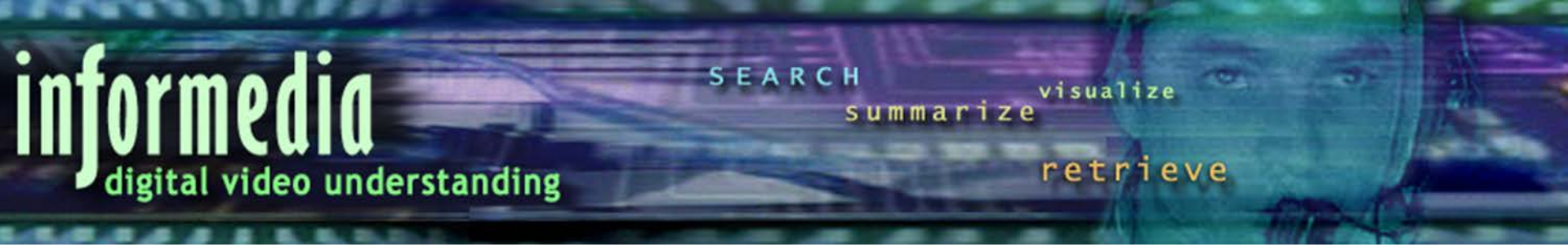


CMU-Informedia @ TRECVID 2014 Surveillance Event Detection

Xingzhong Du, Yang Cai, Yicheng Zhao,
Huan Li, Yi Yang, and Alexander Hauptmann
CMU Informedia Group
Carnegie Mellon University

Carnegie Mellon University

Outline

- General System
- Experiments on Features
- Experiments on Bounding Boxes

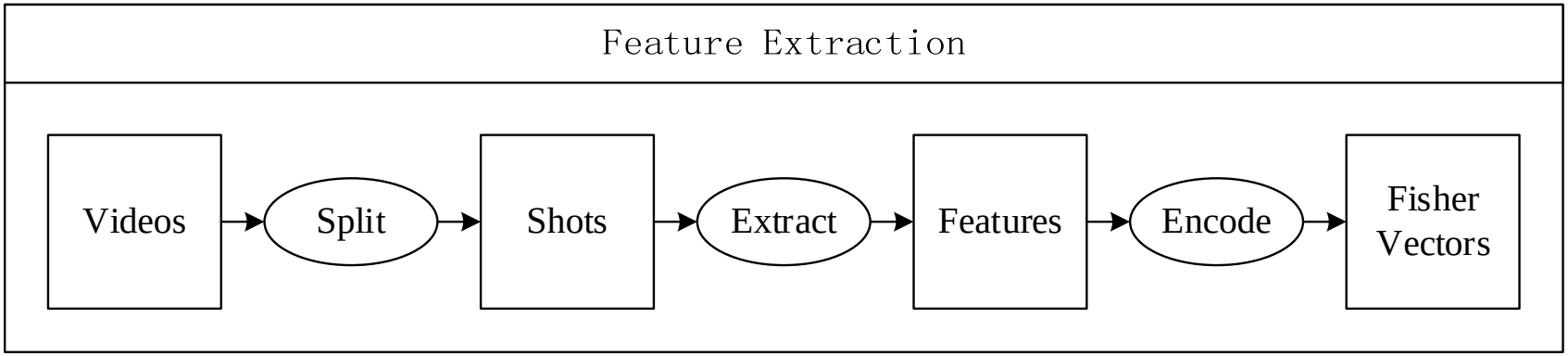
Carnegie Mellon University

Outline

- General System
- Experiments on Features
- Experiments on Bounding Boxes

General System

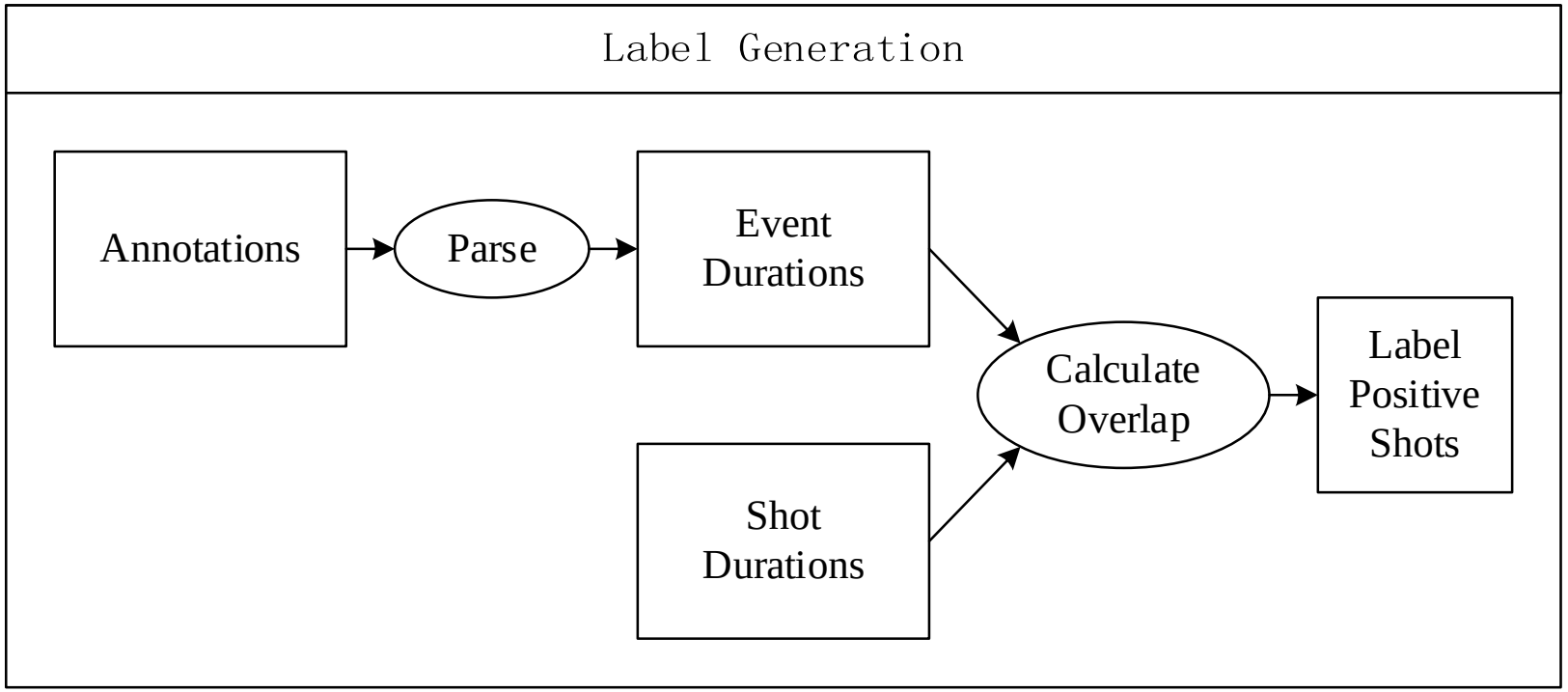
- Feature Extraction



  - Each video is resized to 320 * 240 pixels
  - The videos are sliced into windows (“shots”) of 60 frames with a step of 30 frames overlap.

General System

- Label Generation



  - The shots whose middle frame is located inside the event durations are labeled as positive in the experiments

Carnegie Mellon University

General System

- Training
  - Linear SVM
  - Two-fold cross validation
  - Non Maximum Compression
    - filter the shots by the thresholds from cross validation
    - attribute the adjacent shot label to the shot whose confidence is the local maximum
- Interactive System
  - Ranking the shots according to the detection scores
  - Play the previous and next shots to help the judgment

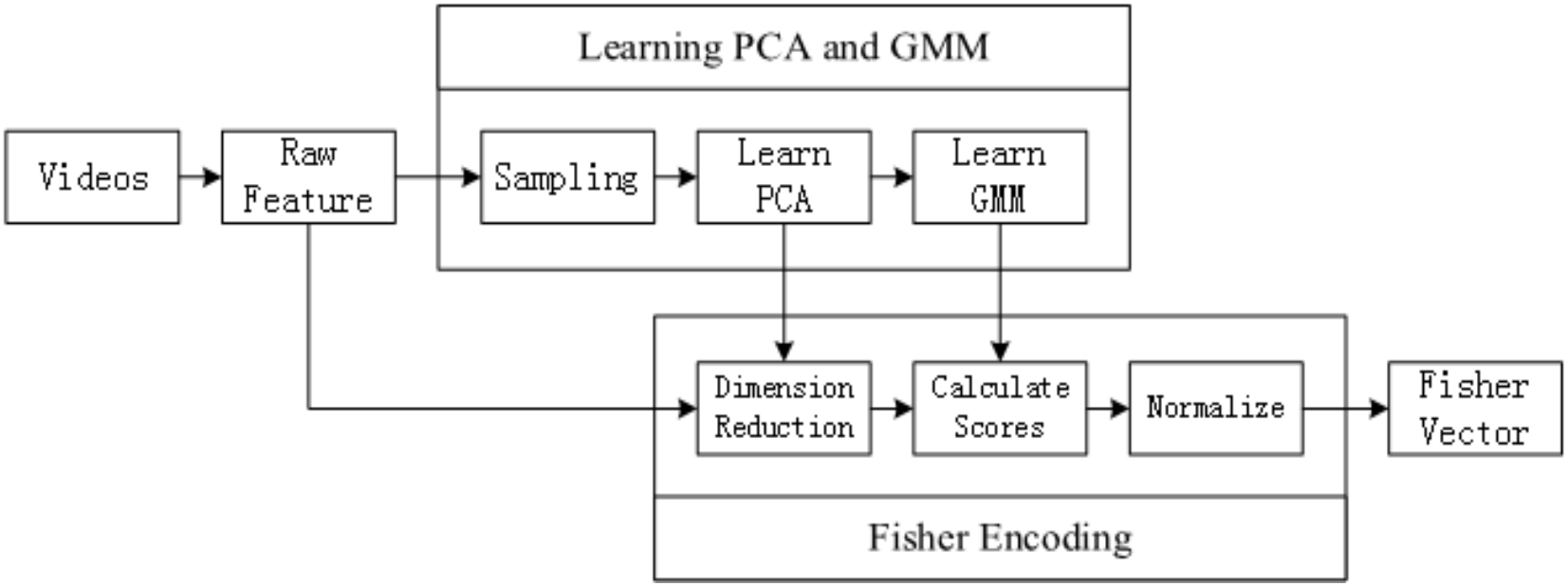
Carnegie Mellon University

Outline

- General System
- Experiments on Features
- Experiments on Bounding Boxes

Improved Dense Trajectory

- Key processing steps (Wang, 2013):
  - Perform PCA to reduce the dimensions by half
  - Learn GMMs of 256 mixture components
  - Fisher encoding
  - Power normalization and L_2 normalization



Performance on evaluation data

	MoSIFT_FV		IDT_FV	
	aDCR	mDCR	aDCR	mDCR
PersonRuns	0.8676	0.8065	0.7835	0.7497
CellToEar	1.0090	0.9993	0.9905	0.9891
ObjectPut	**1.0072**	1.0001	1.0127	0.9994
PeopleMeet	0.9927	0.9652	0.9581	0.9501
PeopleSplitUp	0.9665	0.9456	0.9555	0.9324
Embrace	**0.9671**	0.9305	1.0218	0.9520
Pointing	1.0000	0.9955	0.9965	0.9875

- The improved dense trajectory (IDT) is the best single feature of our system, which was MoSIFT last year

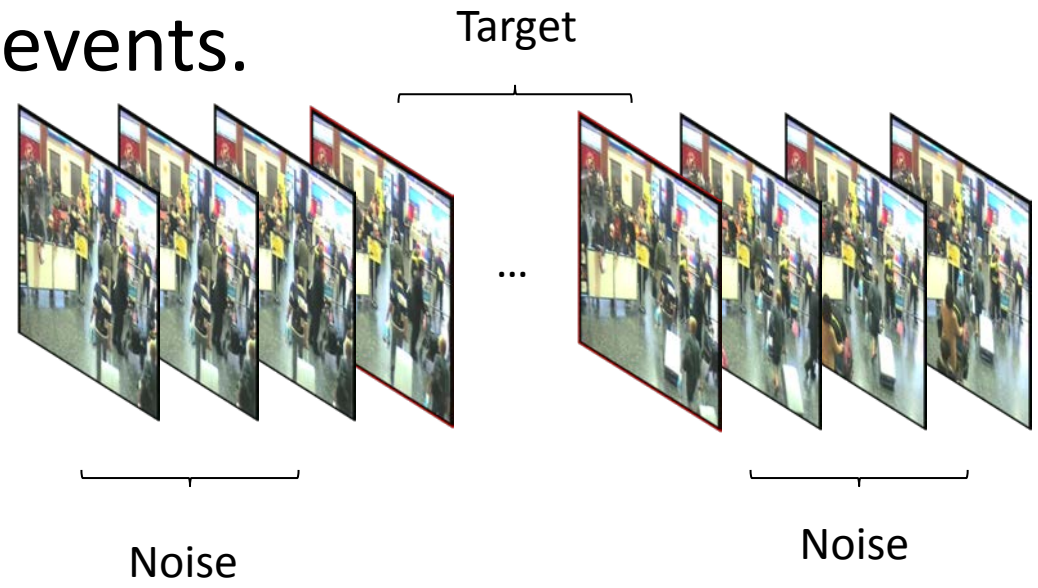
Carnegie Mellon University

Outline

- General System
- Experiments on Features
- Experiments on Bounding Boxes

Temporal Noise in the Encoding

- A shot may not contain exactly the event.
  - Non-event information outside the shot
  - Non-event information inside the shot
- Hypothesis 1: Noise from non-event frames contaminates the positive data especially for short events.



Spatial Noise in the Encoding

- We are only interested in the event-related features located inside the red box.
- Hypothesis 2: All non-event features within a frame constitute noise.

Implementation
Simple process on evaluation data

- Hypothesis 1: Extract only the features from the event shots rather than the fixed length shots for training

- Hypothesis 2: Extract only the features from the bounding box
  - Manually annotated bounding boxes

Evaluation Results

	IDT_FV		IDT_FV_T		IDT_FV_S	
	aDCR	mDCR	aDCR	mDCR	aDCR	mDCR
PersonRuns	0.7835	0.7497	0.8466	0.7843	0.8655	0.8337
CellToEar	0.9905	0.9891	1.0075	0.9865	1.0540	0.9928
ObjectPut	1.0127	0.9994	1.0104	1.0005	1.0801	1.0006
PeopleMeet	0.9581	0.9501	0.9810	0.9710	0.9759	0.9627
PeopleSplitUp	0.9555	0.9324	0.9786	0.9514	1.0029	0.9779
Embrace	1.0218	0.9520	1.0408	0.9871	1.0321	0.9999
Pointing	0.9965	0.9875	1.0101	0.9972	1.0655	0.9972

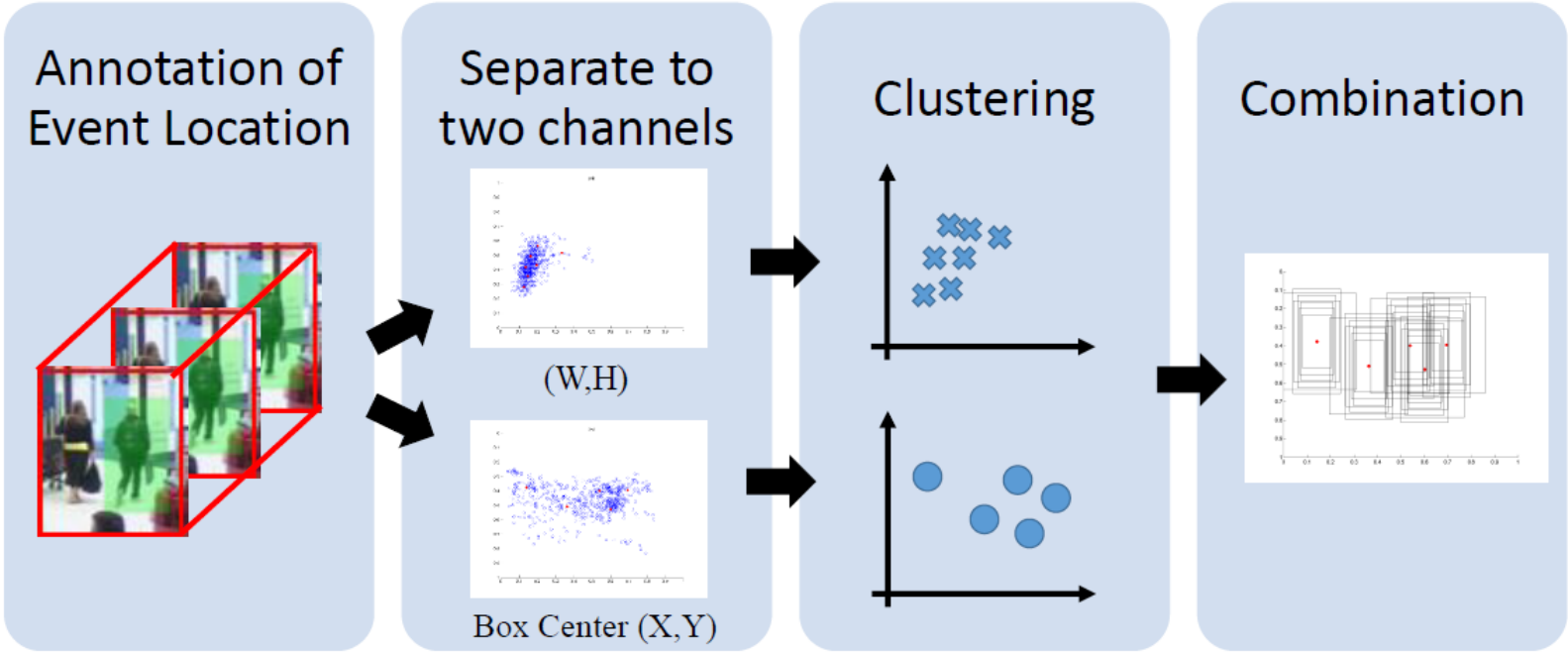
- Result: The simple process did not improve the performance
- T- one vector for each positive example
- S- one vector for each spatially constrained positive example

Carnegie Mellon University

Another Attempt
Template Bounding Boxes

- Learn template bounding boxes over training data
  - Position
  - Width and height
- Detect with template bounding boxes
  - Sliding window of the boxes
  - **Take the union of the area over all boxes**

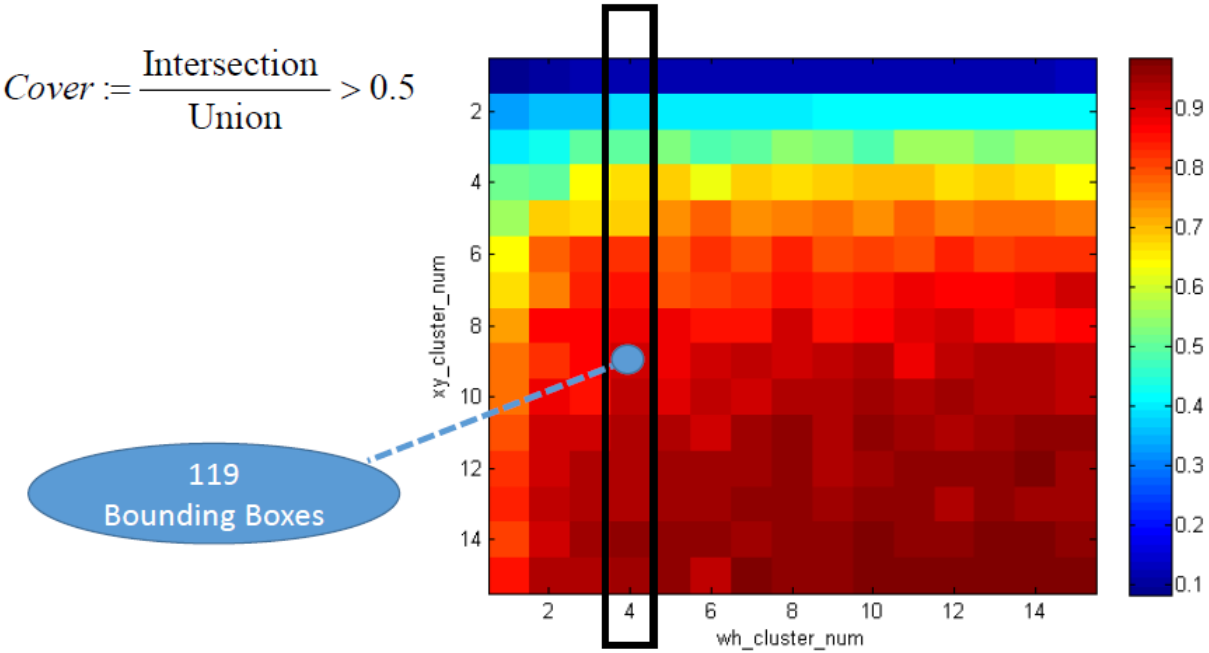
Template Bounding Box



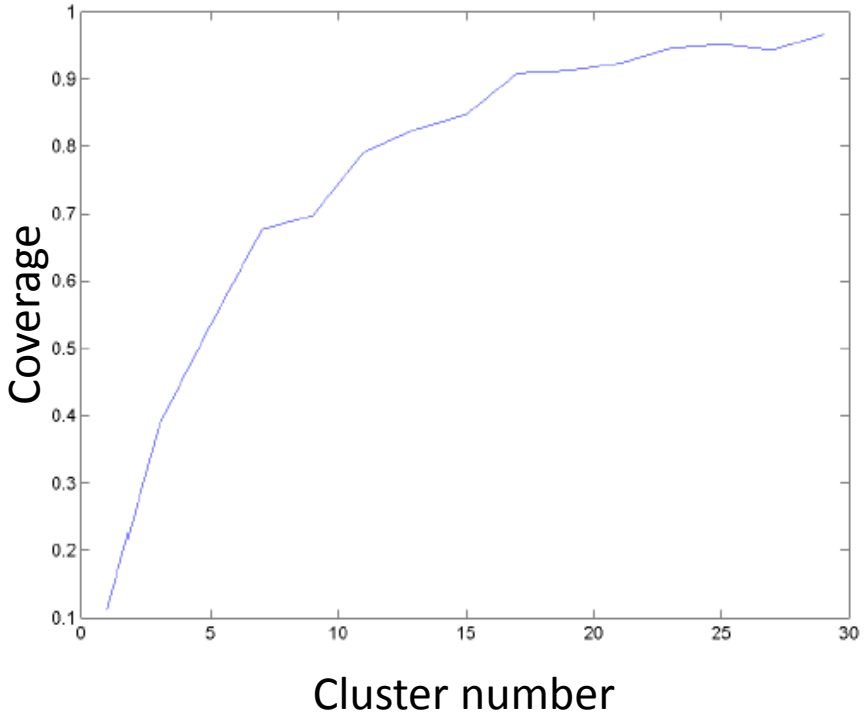
Template Bounding Box

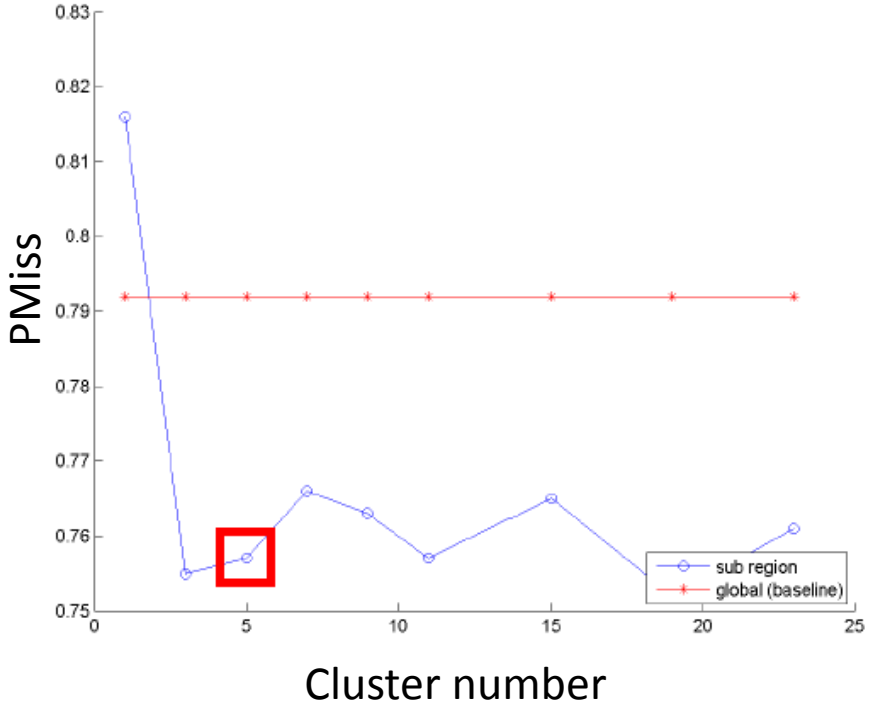
- Event coverage at different numbers of (x,y) cluster and (w,h) cluster

$$Cover := \frac{\text{Intersection}}{\text{Union}} > 0.5$$



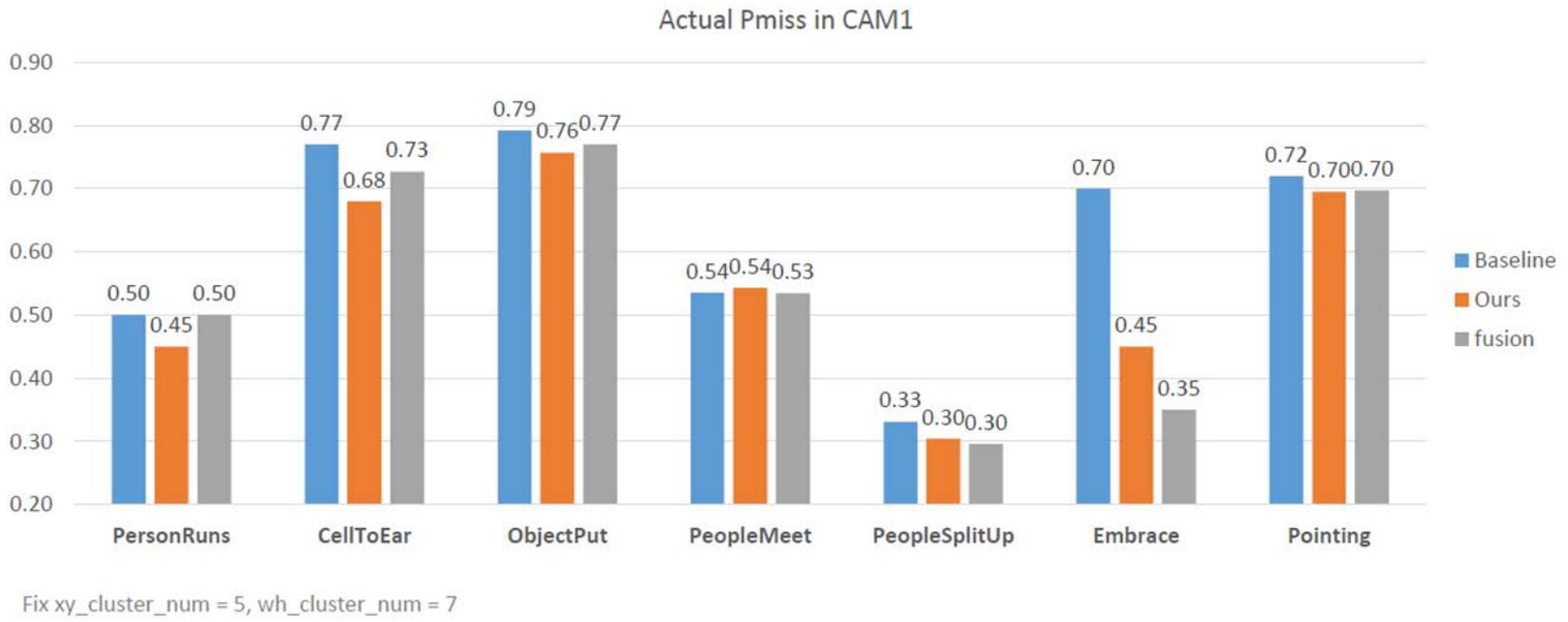
Select the cluster number





Preliminary Results

- Feature : Dense Trajectory, Camera 1 only



Additional work is necessary….

Discussions

- Alternate template bounding box method may improve performance significantly, mostly when the events have strong correlations to specific locations
- The feature tracking methods are not accurate all the time. Some of the meaningful feature points may be located just outside the bounding boxes. So taking the union may improve this problem.
- Preliminary results are supportive

This year's results

Retrospective

	CMU14		Best non-CMU	
	aDCR	mDCR	aDCR	mDCR
PersonRuns	0.8551	0.8500	0.8301	0.8301
CellToEar	1.0032	1.0005	0.9921	0.9911
ObjectPut	1.0023	1.0005	0.9713	0.9761
PeopleMeet	0.9008	0.8975	0.8587	0.8583
PeopleSplitUp	**0.8353**	0.8330	0.8698	0.8594
Embrace	0.8503	0.8462	0.8113	0.8113
Pointing	1.0035	0.9959	0.9998	0.9953

Interactive

	CMU14		Best non-CMU	
	aDCR	mDCR	aDCR	mDCR
PersonRuns	**0.7361**	0.7356	0.7895	0.7895
CellToEar	1.0041	1.0009	0.9555	0.9555
ObjectPut	**0.9280**	0.9276	0.9641	0.9641
PeopleMeet	0.8872	0.8849	0.7960	0.7960
PeopleSplitUp	**0.8115**	0.8097	0.8390	0.8390
Embrace	0.8417	0.8357	0.6978	0.6978
Pointing	0.9746	0.9745	0.9744	0.9744

Significant improvement in the interactive task due to reduced false alarms.
This was only possible because IDT found more positives than last year's STIP and MoSIFT

Carnegie Mellon University

Thank you