CMU-IBM-NUS@TRECVID 2012: Surveillance Event Detection(SED)

Yang Cai †*, Qiang Chen ɨł*, Lisa Brown ɨ, Ankur Datta ɨ, Quanfu Fan ɨ, Rogerio Feris ɨ, Shuicheng Yan ł, Alex Hauptmann †, Sharath Pankanti ɨ

† Carnegie Mellon University

ɨ IBM Research

ł National University of Singapore

*Equal contributions by co-authors.

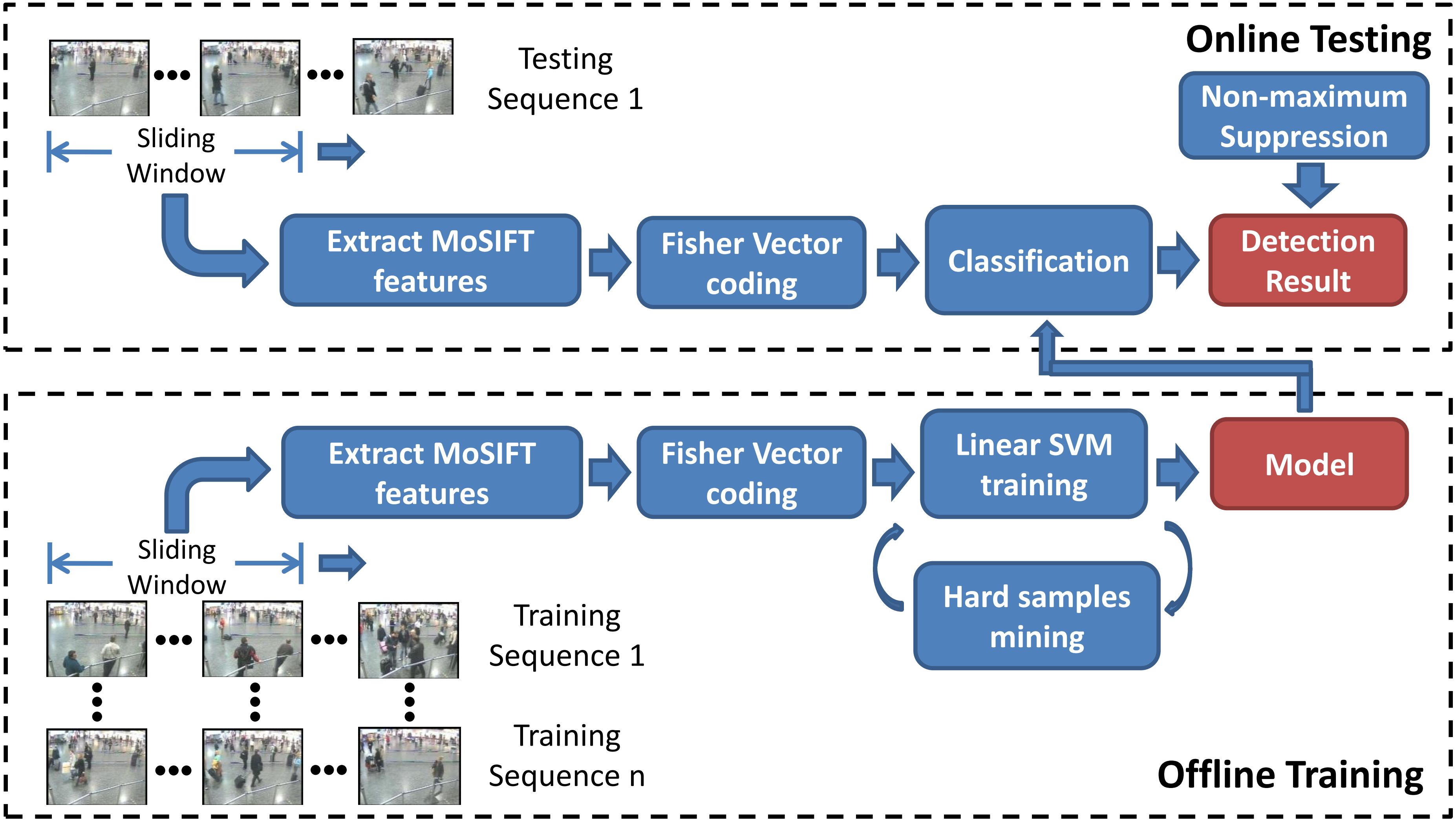
Outline

- Retrospective Event Detection
  - System Overview
  - Fisher Vector Coding for Event Representation
  - Performance Evaluation
- Interactive Event Detection
  - Detection Results Visualization
    - Event-specific Results Visualization
  - User Feedback Utilization
    - Temporal Locality Based Search
  - Performance Evaluation

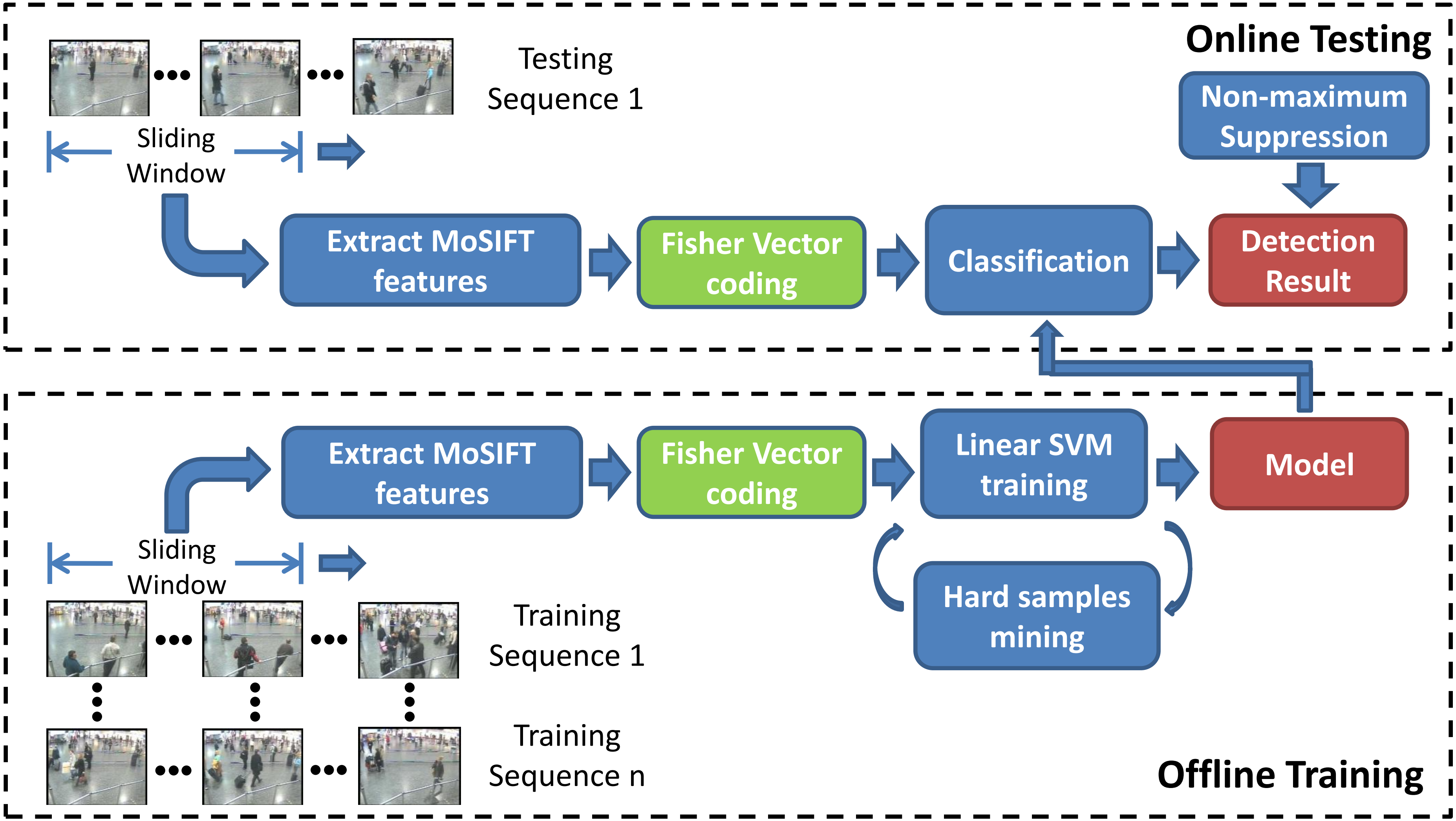
Outline

- Retrospective Event Detection
  - System Overview
  - Fisher Vector Coding for Event Representation
  - Performance Evaluation
- Interactive Event Detection
  - Detection Results Visualization
    - Event-specific Results Visualization
  - User Feedback Utilization
    - Temporal Locality Based Search
  - Performance Evaluation

System Overview



System Overview



Event Representation

- Fisher Vector (FV) Coding [1]:
  - A GMM is learnt to model each MoSIFT features.
  - For each feature point in a detection window, the gradients with respective to mean and standard deviation of the GMM are calculated.
  - FV is the concatenation of the two gradients averaged over all features in a detection window.
- Fisher Vector (FV) vs. Bag-of-Word(BoW) [2]
  - BoW is only about counting local descriptors assigned to each visual word while FV includes higher order statistics.
  - FV is faster to compute than BoW for a given feature dimension.

[1] F. Perronnin and T. Mensink. Improving the fisher kernel for large-scale image classification. In *ECCV*, 2010.

[2] F. Perronnin and H. Jégou. Tutorial on Large-Scale Visual Recognition, in CVPR, 2012.

Performance Evaluation

Primary Runs Results	CMU-IBM_FV2012		Others' Best 2012		CMU_BoW2011	
	ActDCR	MinDCR	ActDCR	MinDCR	ActDCR	MinDCR
CellToEar	1.0007	1.0003	1.004	0.9814	1.0365	1.0003
Embrace	0.8	0.7794	0.8247	0.824	0.884	0.8658
ObjectPut	1.004	0.9994	0.9983	0.9983	1.0171	1.0003
PeopleMeet	1.0361	0.949	0.9799	0.9777	1.01	0.9724
PeopleSplitUp	0.8433	0.7882	0.9843	0.9787	1.0217	1.0003
PersonRuns	0.8346	0.7872	0.9702	0.9623	0.8924	0.837
Pointing	1.0175	0.9921	0.9813	0.977	1.5186	1.0001

Performance Evaluation

Primary Runs Results	CMU-IBM_FV2012		Others' Best 2012		CMU_BoW2011	
	ActDCR	MinDCR	ActDCR	MinDCR	ActDCR	MinDCR
CellToEar	1.0007	1.0003	1.004	0.9814	1.0365	1.0003
Embrace	0.8	0.7794	0.8247	0.824	0.884	0.8658
ObjectPut	1.004	0.9994	0.9983	0.9983	1.0171	1.0003
PeopleMeet	1.0361	0.949	0.9799	0.9777	1.01	0.9724
PeopleSplitUp	0.8433	0.7882	0.9843	0.9787	1.0217	1.0003
PersonRuns	0.8346	0.7872	0.9702	0.9623	0.8924	0.837
Pointing	1.0175	0.9921	0.9813	0.977	1.5186	1.0001

- Compared to this year other teams' results (Others' Best 2012):
  - our system has better performance on 4/7 events (actual/minimum DCR of primary run).

Performance Evaluation

Primary Runs Results	CMU-IBM_FV2012		Others' Best 2012		CMU_BoW2011	
	ActDCR	MinDCR	ActDCR	MinDCR	ActDCR	MinDCR
CellToEar	1.0007	1.0003	1.004	0.9814	1.0365	1.0003
Embrace	0.8	0.7794	0.8247	0.824	0.884	0.8658
ObjectPut	1.004	0.9994	0.9983	0.9983	1.0171	1.0003
PeopleMeet	1.0361	0.949	0.9799	0.9777	1.01	0.9724
PeopleSplitUp	0.8433	0.7882	0.9843	0.9787	1.0217	1.0003
PersonRuns	0.8346	0.7872	0.9702	0.9623	0.8924	0.837
Pointing	1.0175	0.9921	0.9813	0.977	1.5186	1.0001

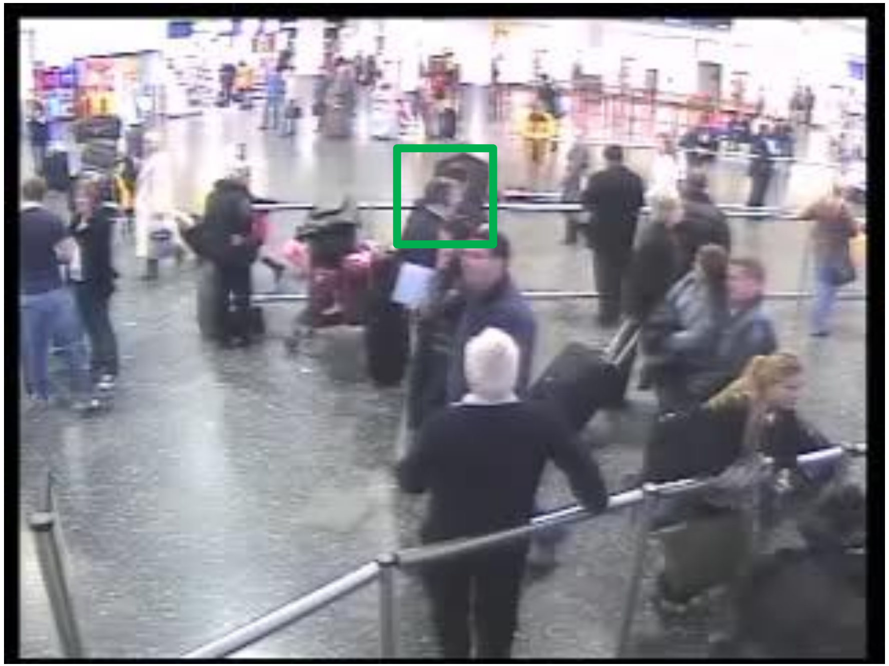
- Compared to this year other teams' results (Others' Best 2012):
  - our system has better performance on 4/7 events (actual/minimum DCR of primary run).
- Compared to our last year system based on BoW (CMU_BoW2011):
  - this year system gets improvement on 6/7 events (actual/min DCR of primary run).

Outline

- Retrospective Event Detection
  - System Overview
  - Fisher Vector Encoding for Event Representation
  - Performance Evaluation
- Interactive Event Detection
  - Detection Results Visualization
    - Event-specific Results Visualization
  - User Feedback Utilization
    - Temporal Locality Based Search
  - Performance Evaluation

Detection Results Visualization

- Problem:
  - Without a good visualization method, user-system interaction can be very ineffective and inefficient.
    - E.g. one may use several minutes to judge if a system detection is true positive or false alarm.

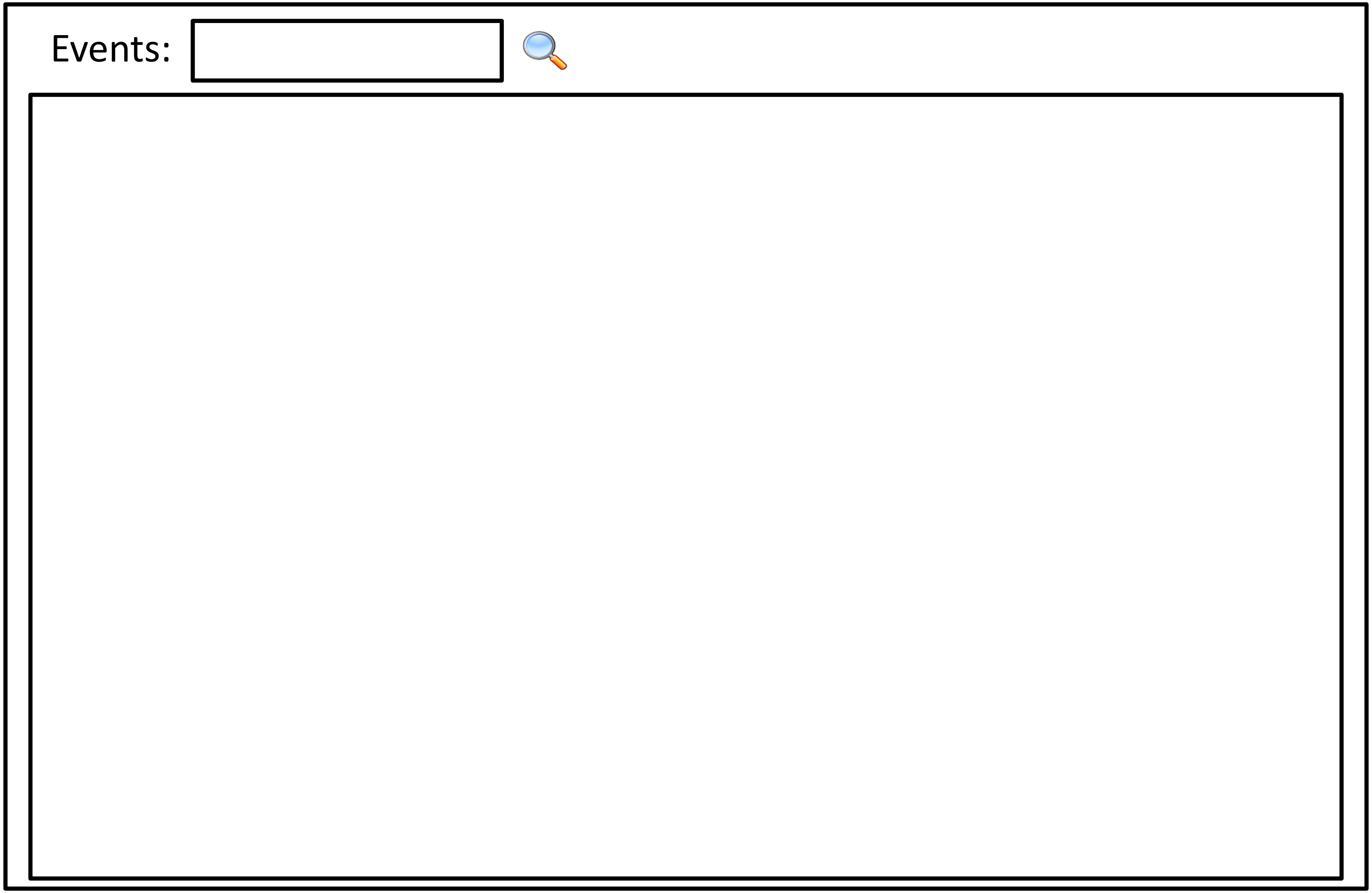


Is this a "CellToEar"?

Detection Results Visualization

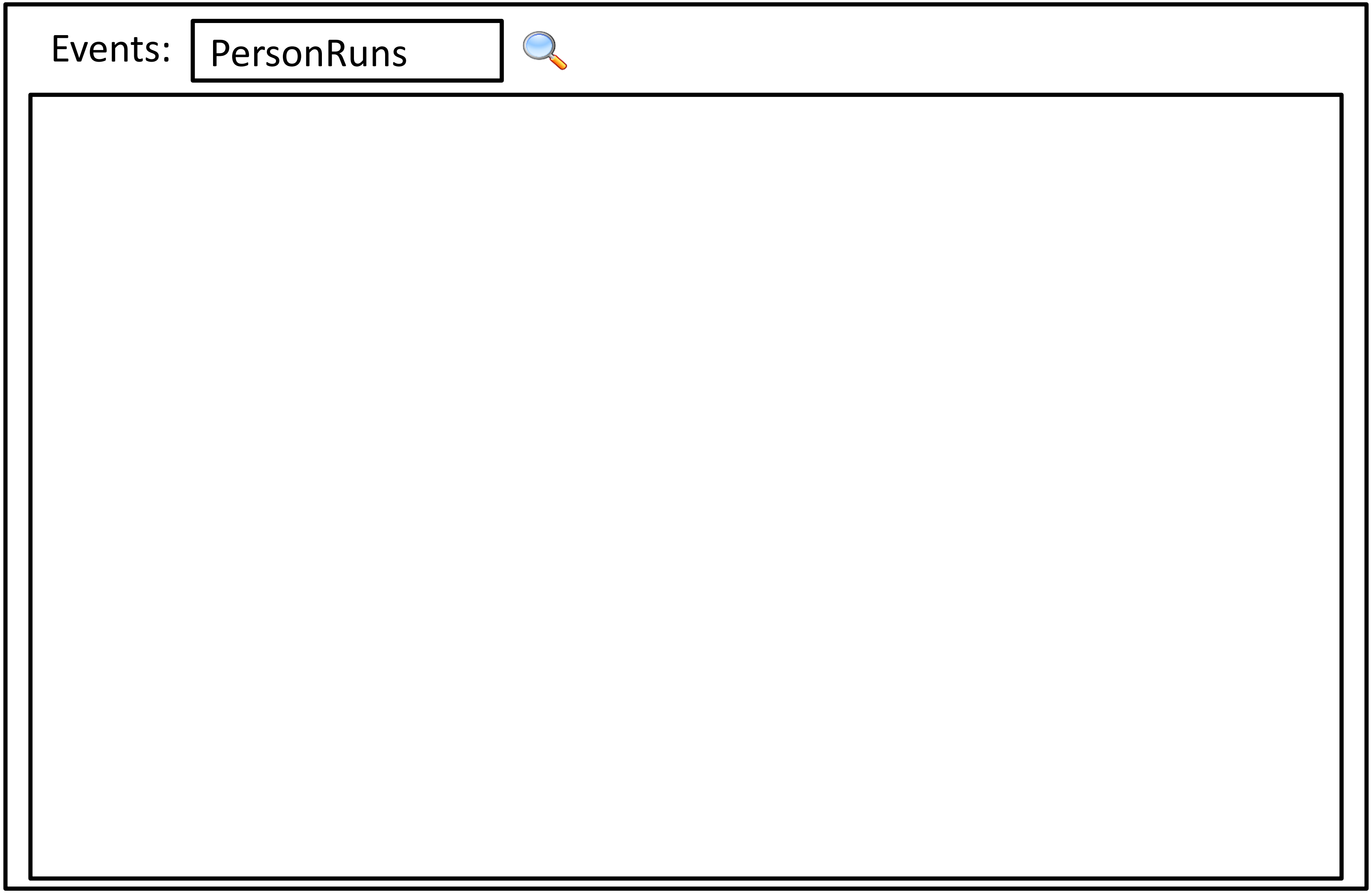
- Objective:
  - To find visualization methods that enable users to *accurately* and *quickly* understand detection results.

Event-specific Results Visualization



Events:

Event-specific Results Visualization



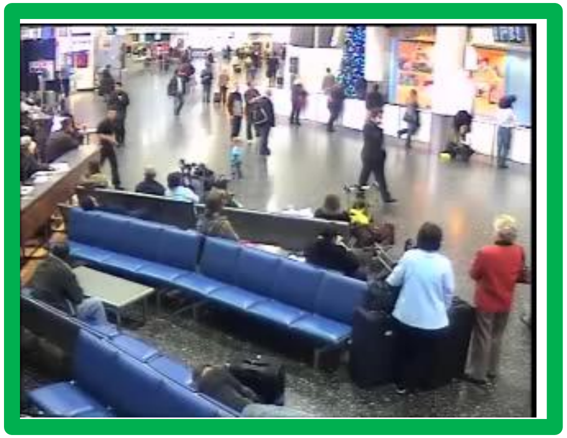
Events: PersonRuns 🔍

Event-specific Results Visualization

Events: PersonRuns

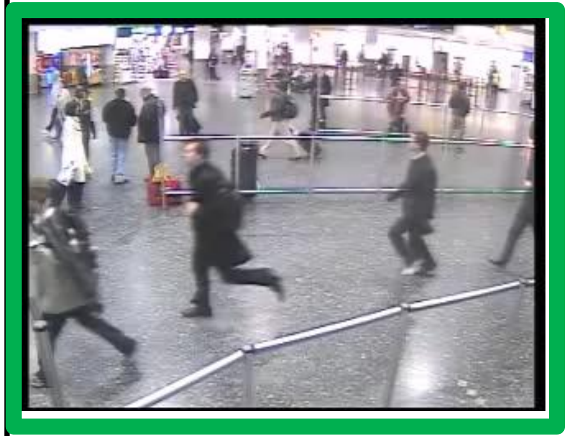
Which are true positives (PersonRuns)?

(A)



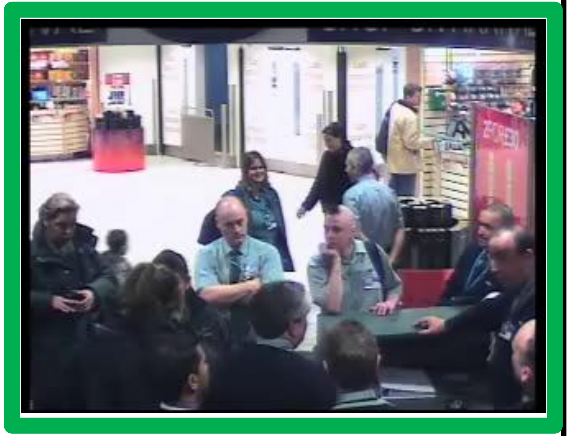
(B)

(C)



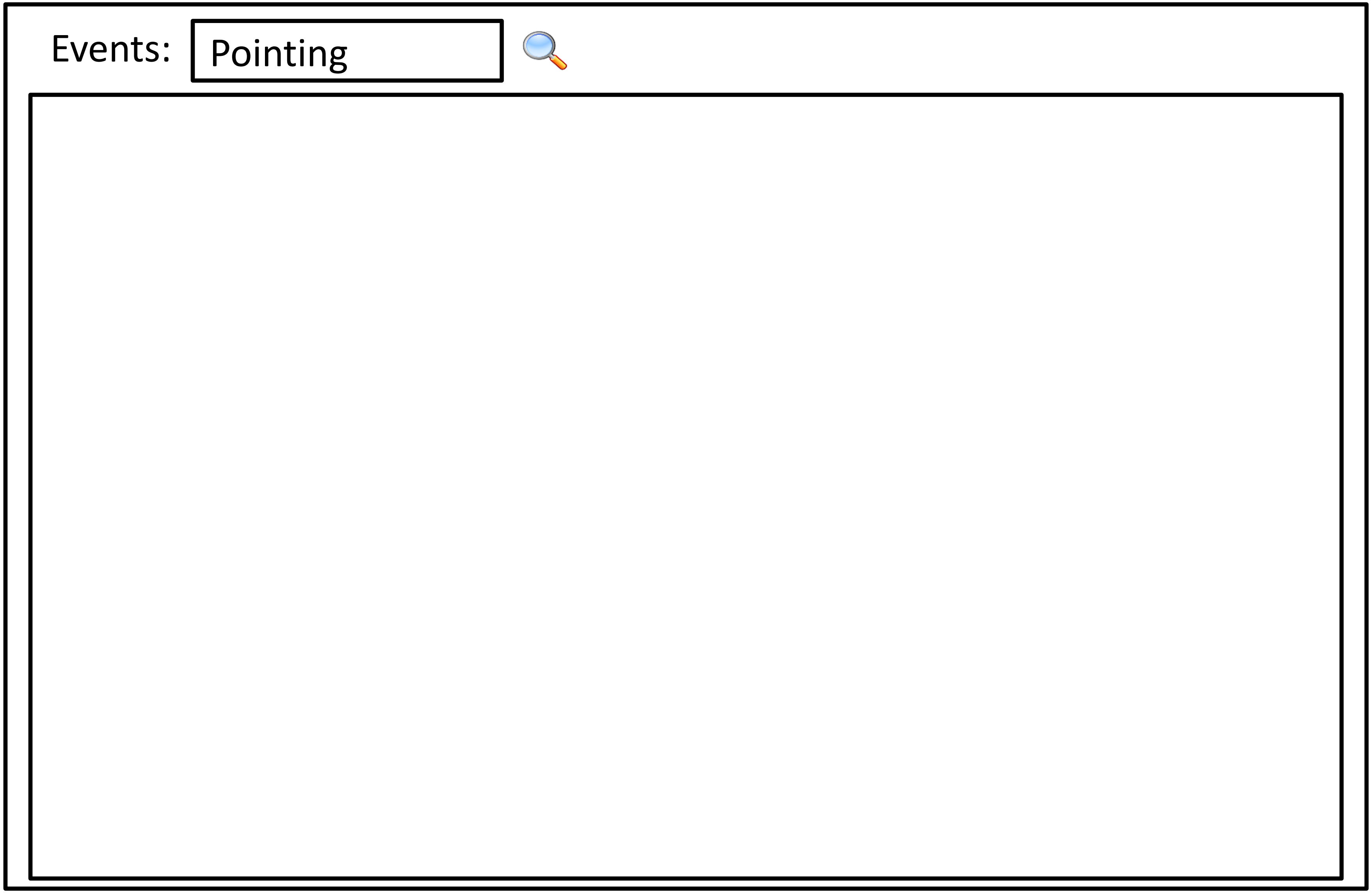
(D)

(E)



(F)

Event-specific Results Visualization



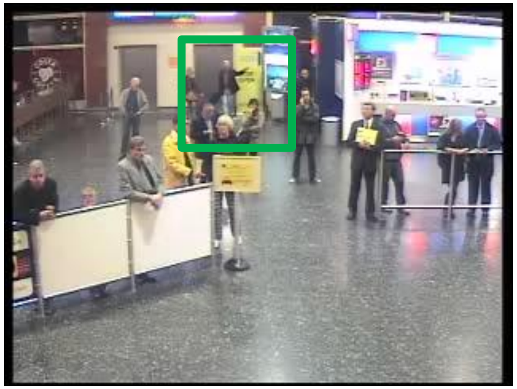
Events: Pointing 🔍

Event-specific Results Visualization

Events: Pointing 🔍 Which are true positives (Pointing)?

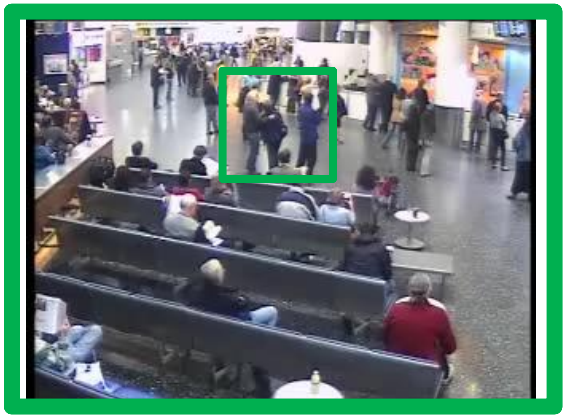
(A)

(B)



(C)

(D)



(E)

(F)

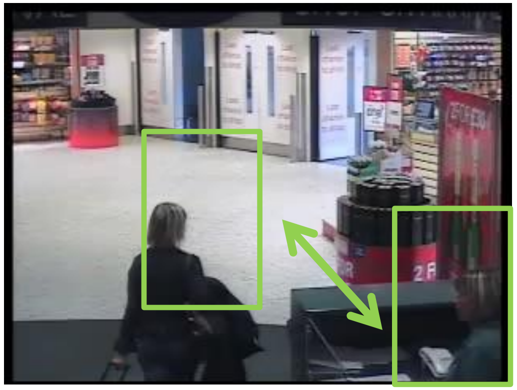
Event-specific Results Visualization

Events: Pointing

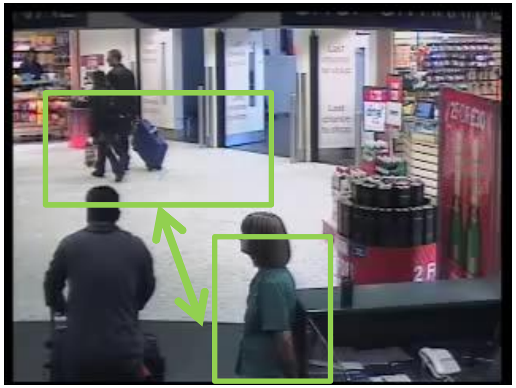
Which are true positives (Pointing)?

Event-specific Detection Visualization

Events: PeopleSplitUp 🔍 Are they "PeopleSplitUp"? Probably…

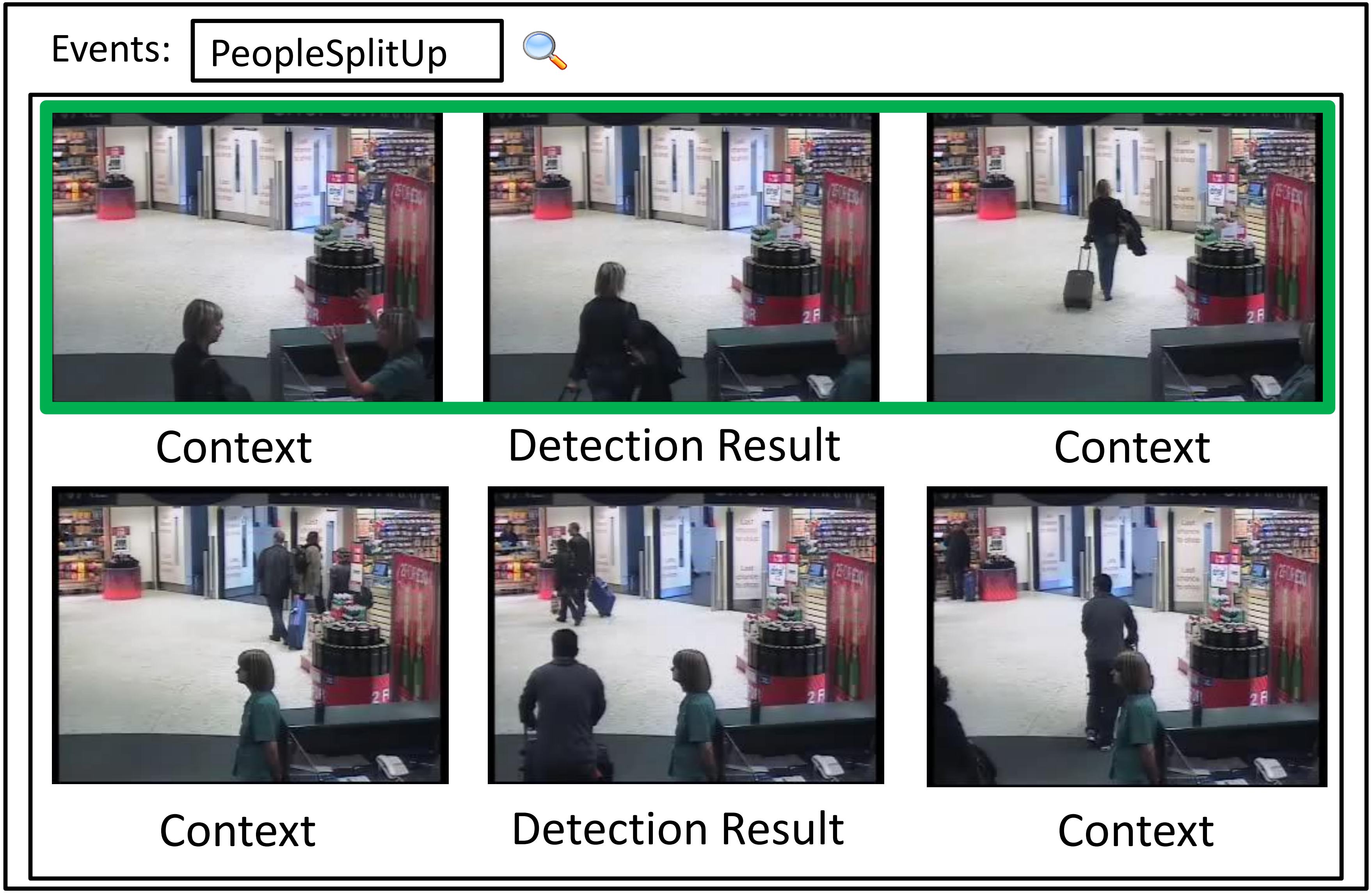


Detection Result



Detection Result

Event-specific Results Visualization



Events: PeopleSplitUp 🔍

Context

Detection Result

Context

Context

Detection Result

Context

Event-specific Results Visualization

- Different events are visualized using different schemes:
  - *many low-resolution* units:
    - Place multiple low-resolution units in a screen.
    - For events that can be captured by a glance.
      e.g. “PersonRuns”

many low-resolution units

  - *few high-resolution* units:
    - Place few high-resolution units in a screen.
    - For events that require careful checking.
      e.g. “CellToEar”, “ObjectPut”, “Pointing”.

few high-resolution units

  - *contextual* units:
    - Add context next to detections.
    - For group events with multiple phrases.
      e.g. “PeopleSplitUp”, “PeopleMeet”,
      “Embrace”.

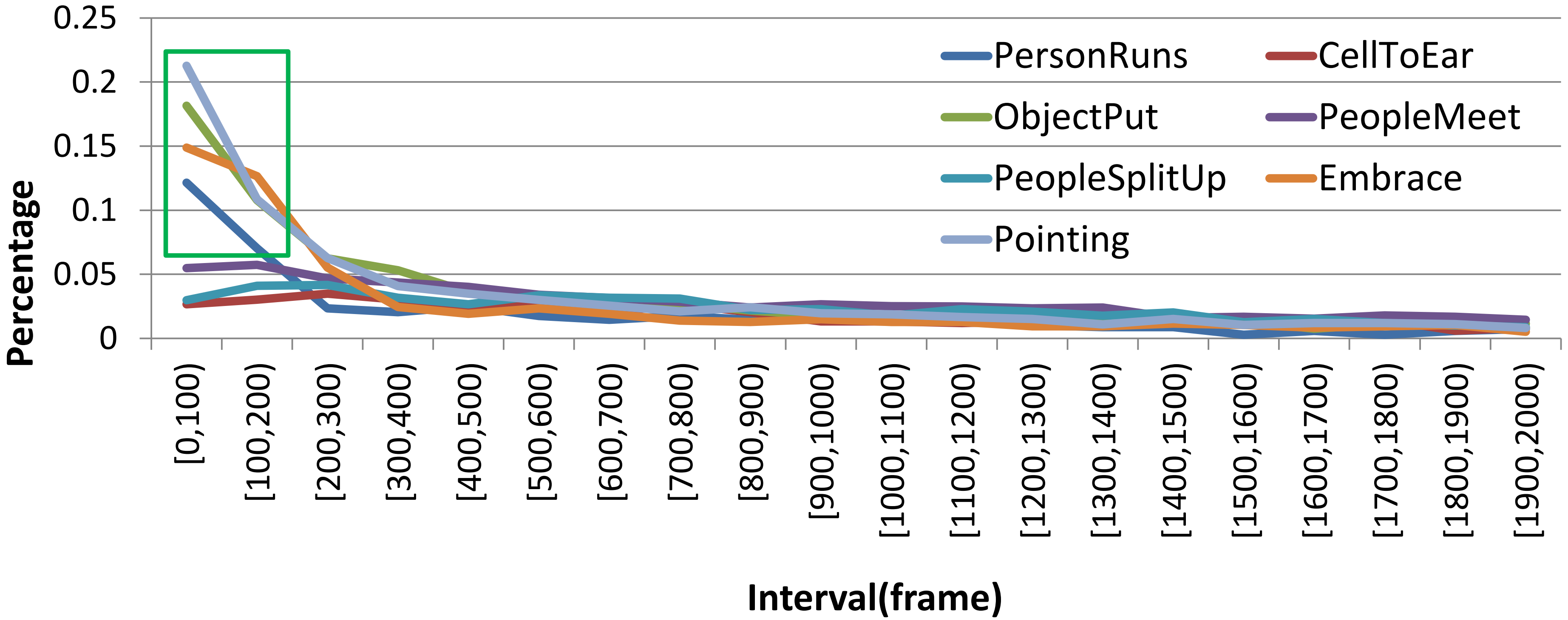
contextual units

User Feedback Utilization

- Problem:
  - Without feedback utilization, the interaction is nothing but removing *false alarms*.
- Objective:
  - To *efficiently* reduce *miss detections* as well by leveraging user feedbacks.

An Observation

- A temporally *clustered* distribution (*temporal locality*):
  - We calculated the interval between consecutive events of same class in development data.
  - For some events (e.g. “Pointing”, “ObjectPut”, “Embrace”, “PersonRuns”, etc.), most of the intervals are very small (< 200 frames/8 seconds).



Temporal Locality Based Search

- What does the observation tell us?
  - If we observe one positive at somewhere, we are likely to find another positive nearby.
- Temporal locality based search:
  - After receiving one positive feedback from user, the system returns user a set of neighbors living closely to the positive. Then user can quickly go through the neighbors to find potential miss detections.

Performance Evaluation

	Development Set **(Training: Dev08, Testing: Eval08, Wall time: 5 mins)**				**Evaluation Set** **(Primary Run)**	
Actual DCR	**Retro**	**Naive**	**ESpecVis**	**ESpecVis+TLSearch**	**Retro**	**ESpecVis+TLSearch**
CellToEar	1.0008	1.0014	1.0008	1.0009	1.0007	1.009
Embrace	0.9519	0.9547	0.9344	0.9115	0.8	0.6696
ObjectPut	1.0033	1.0026	1.0024	1.0023	1.004	1.0064
PeopleMeet	0.9381	0.9338	0.9334	0.9361	1.0361	0.9786
PeopleSplitUp	0.8972	0.9416	0.889	0.8863	0.8433	0.8177
PersonRuns	0.761	0.7528	0.7511	0.7366	0.8346	0.6445
Pointing	1.0168	1.0109	1.0134	1.0084	1.0175	0.9854

- **Retro**: retrospective event detection system output using fisher vector method.
- **Naïve**: the baseline interactive method, which linearly scans system output with only *"many low-resolution units"* visualization method for all events.
- **ESpecVis**: linearly scan system output with *event-specific visualization*.
- **ESpecVis+TLSearch**: scan the system output with both *event-specific visualization* and *temporal locality search*.

Performance Evaluation

	Development Set (Training: Dev08, Testing: Eval08, Wall time: 5 mins)				Evaluation Set (Primary Run)	
Actual DCR	**Retro**	**Naive**	**ESpecVis**	**ESpecVis+TLSearch**	**Retro**	**ESpecVis+TLSearch**
CellToEar	1.0008	1.0014	1.0008	1.0009	1.0007	1.009
Embrace	0.9519	0.9547	0.9344	0.9115	0.8	0.6696
ObjectPut	1.0033	1.0026	1.0024	1.0023	1.004	1.0064
PeopleMeet	0.9381	0.9338	0.9334	0.9361	1.0361	0.9786
PeopleSplitUp	0.8972	0.9416	0.889	0.8863	0.8433	0.8177
PersonRuns	0.761	**0.7528**	0.7511	0.7366	0.8346	0.6445
Pointing	1.0168	1.0109	1.0134	1.0084	1.0175	0.9854

- **Retro**: retrospective event detection system output using fisher vector method.
- **Naïve**: the baseline interactive method, which linearly scans system output with only *"many low-resolution units"* visualization method for all events.
- **ESpecVis**: linearly scan system output with *event-specific visualization*.
- **ESpecVis+TLSearch**: scan the system output with both *event-specific visualization* and *temporal locality search*.

Performance Evaluation

	Development Set **(Training: Dev08, Testing: Eval08, Wall time: 5 mins)**				**Evaluation Set** **(Primary Run)**	
Actual DCR	**Retro**	**Naive**	**ESpecVis**	**ESpecVis+TLSearch**	**Retro**	**ESpecVis+TLSearch**
CellToEar	1.0008	1.0014	1.0008	1.0009	1.0007	1.009
Embrace	0.9519	0.9547	**0.9344**	0.9115	0.8	0.6696
ObjectPut	1.0033	1.0026	1.0024	1.0023	1.004	1.0064
PeopleMeet	0.9381	0.9338	0.9334	0.9361	1.0361	0.9786
PeopleSplitUp	0.8972	0.9416	**0.889**	0.8863	0.8433	0.8177
PersonRuns	0.761	0.7528	0.7511	0.7366	0.8346	0.6445
Pointing	1.0168	1.0109	1.0134	1.0084	1.0175	0.9854

- **Retro**: retrospective event detection system output using fisher vector method.
- **Naïve**: the baseline interactive method, which linearly scans system output with only *“many low-resolution units”* visualization method for all events.
- **ESpecVis**: linearly scan system output with *event-specific visualization*.
- **ESpecVis+TLSearch**: scan the system output with both *event-specific visualization* and *temporal locality search*.

Performance Evaluation

	Development Set (Training: Dev08, Testing: Eval08, Wall time: 5 mins)				Evaluation Set (Primary Run)	
Actual DCR	**Retro**	**Naive**	**ESpecVis**	**ESpecVis+TLSearch**	**Retro**	**ESpecVis+TLSearch**
CellToEar	1.0008	1.0014	1.0008	1.0009	1.0007	1.009
Embrace	0.9519	0.9547	0.9344	**0.9115**	0.8	0.6696
ObjectPut	1.0033	1.0026	1.0024	1.0023	1.004	1.0064
PeopleMeet	0.9381	0.9338	0.9334	0.9361	1.0361	0.9786
PeopleSplitUp	0.8972	0.9416	0.889	0.8863	0.8433	0.8177
PersonRuns	0.761	0.7528	0.7511	**0.7366**	0.8346	0.6445
Pointing	1.0168	1.0109	1.0134	1.0084	1.0175	0.9854

- **Retro**: retrospective event detection system output using fisher vector method.
- **Naïve**: the baseline interactive method, which linearly scans system output with only *"many low-resolution units"* visualization method for all events.
- **ESpecVis**: linearly scan system output with *event-specific visualization*.
- **ESpecVis+TLSearch**: scan the system output with both *event-specific visualization* and *temporal locality search*.

Performance Evaluation

	Development Set **(Training: Dev08, Testing: Eval08, Wall time: 5 mins)**				**Evaluation Set** **(Primary Run)**	
Actual DCR	**Retro**	**Naive**	**ESpecVis**	**ESpecVis+TLSearch**	**Retro**	**ESpecVis+TLSearch**
CellToEar	1.0008	1.0014	1.0008	1.0009	1.0007	1.009
Embrace	0.9519	0.9547	0.9344	0.9115	0.8	**0.6696**
ObjectPut	1.0033	1.0026	1.0024	1.0023	1.004	1.0064
PeopleMeet	0.9381	0.9338	0.9334	0.9361	1.0361	0.9786
PeopleSplitUp	0.8972	0.9416	0.889	0.8863	0.8433	0.8177
PersonRuns	0.761	0.7528	0.7511	0.7366	0.8346	**0.6445**
Pointing	1.0168	1.0109	1.0134	1.0084	1.0175	0.9854

- **Retro**: retrospective event detection system output using fisher vector method.
- **Naïve**: the baseline interactive method, which linearly scans system output with only *"many low-resolution units"* visualization method for all events.
- **ESpecVis**: linearly scan system output with *event-specific visualization*.
- **ESpecVis+TLSearch**: scan the system output with both *event-specific visualization* and *temporal locality search*.

Conclusions

- Retrospective System:
  - Fisher Vector coding significantly improves detection performance (DCR) on some events. E.g "PersonRuns", "Embrace", "PeopleSplitUp".
  - The performances of "CellToEar", "Pointing" and "ObjectPut" are still not good.
- Interactive System:
  - Event-specific scheme should be used in detection results visualization.
  - Temporal locality search can improve the performance for event with *good temporal locality* and *reasonable system detection accuracy*.

Future Works

- Retrospective System:
  - “Interaction-oriented” detection methods which aim to facilitate user interaction need to be studied. E.g. event spatially localization.
- Interactive System:
  - Better visualization techniques need to be developed for difficult events. E.g. “CellToEar”, “ObjectPut”.
  - More user feedback utilization methods need to be studied.

Thanks!