IBM-Northwestern@TRECVID 2013: Surveillance Event Detection(SED)

Yu Cheng ɨ*, Lisa Brown ɨ, Quanfu Fan ɨ,
Rogerio Feris ɨ, Alok Choudhary *, Sharath Pankanti ɨ

ɨ IBM T. J. Watson Research Center
* Northwestern University

IBM Research

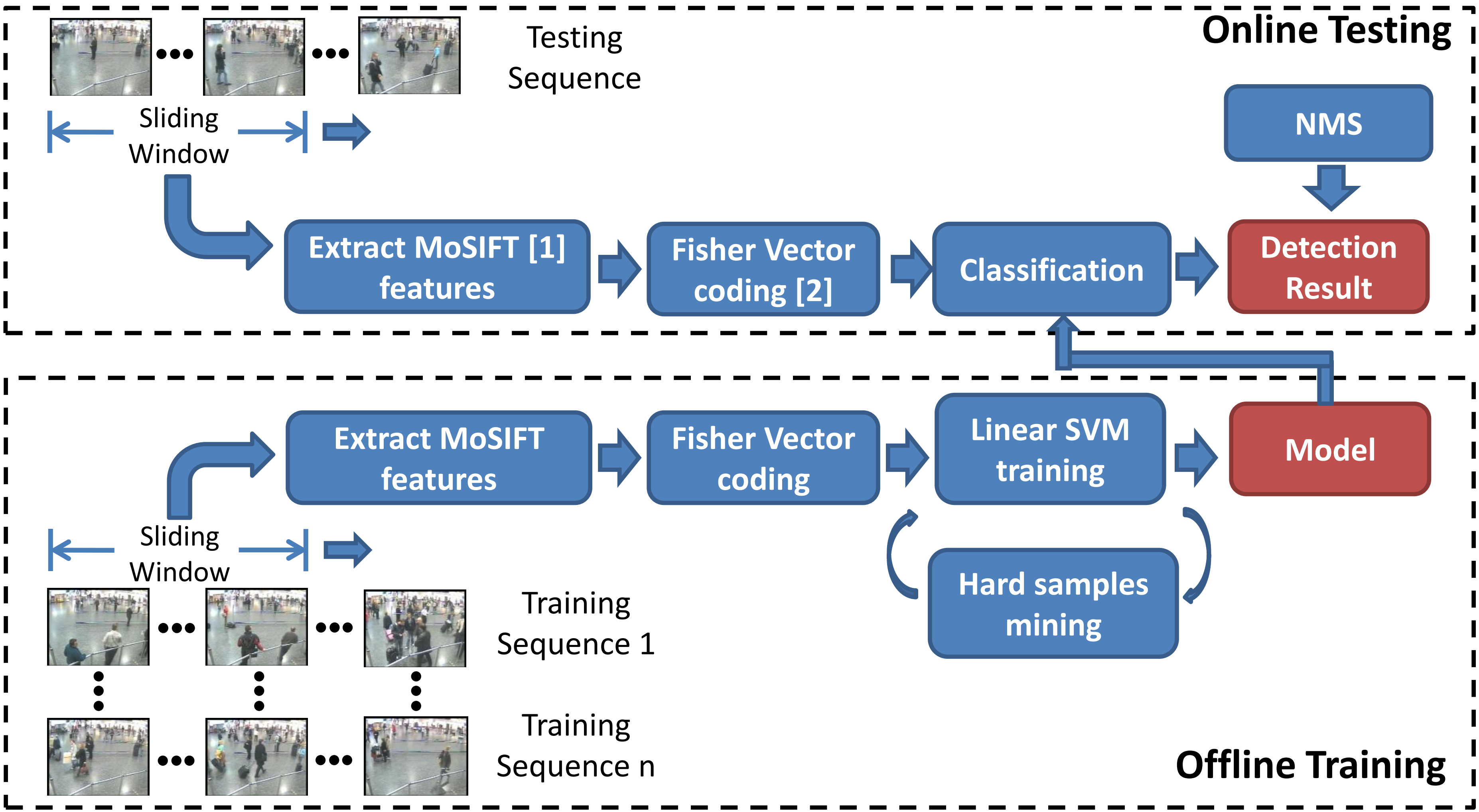
NORTHWESTERN UNIVERSITY

Outline

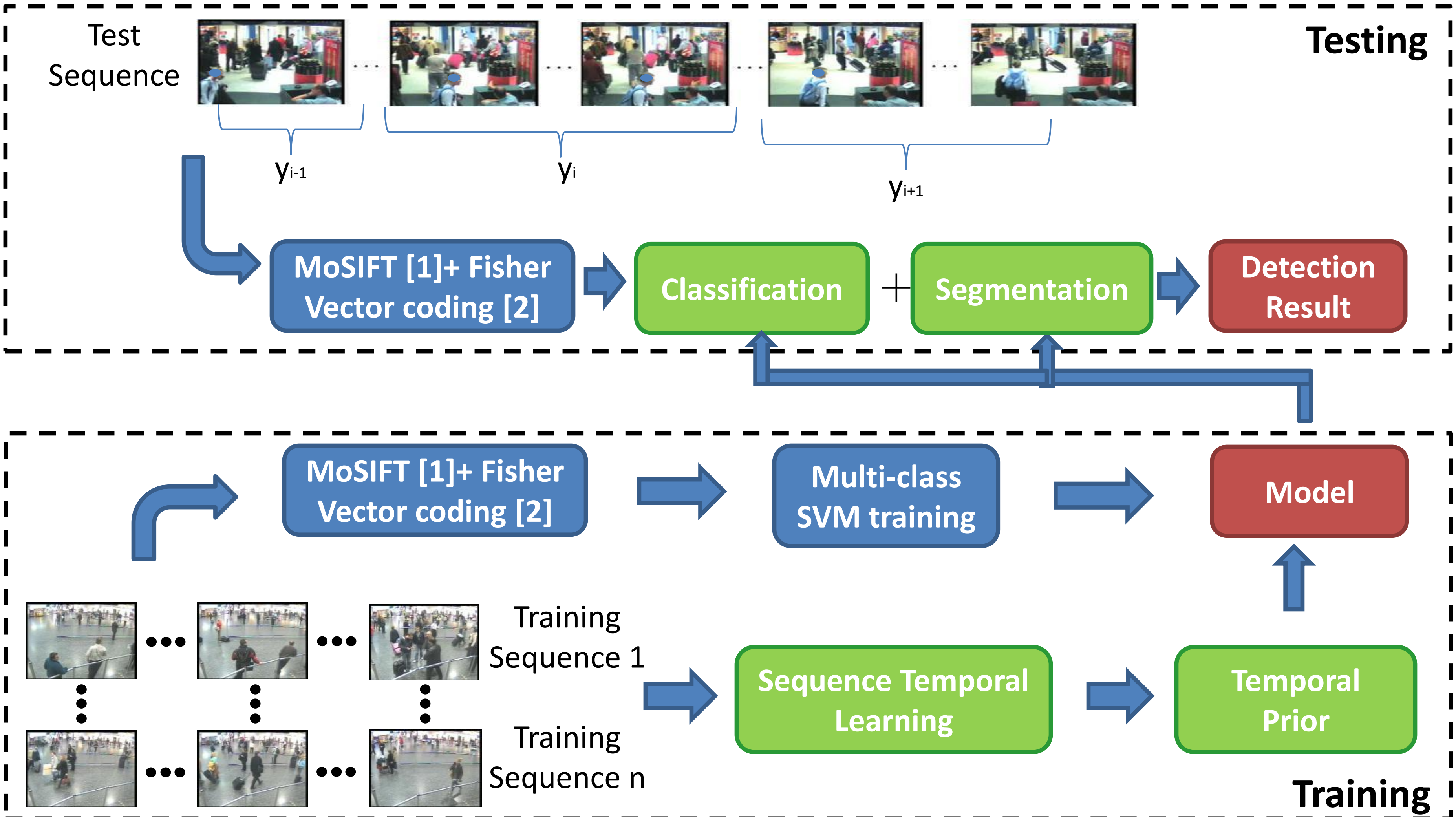
- Retrospective Event Detection
  – System Overview
  – Temporal Modeling for Event Detection
  – Performance Evaluation

- Interactive Event Detection
  – Interactive Visualization
  – Risk Ranking
  – Performance Evaluation

System Overview (CMU-IBM 2012)

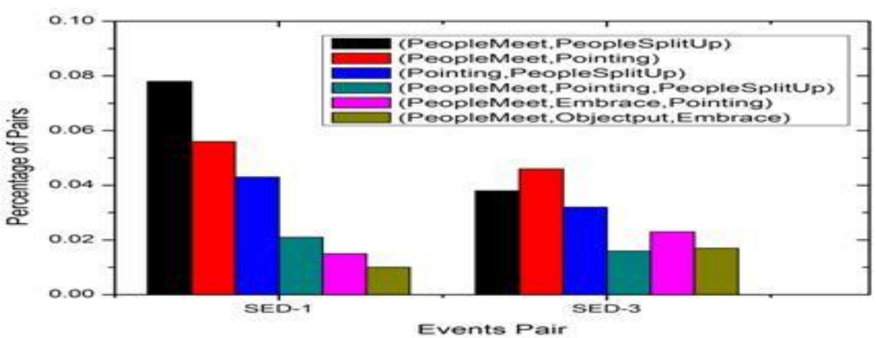


System Overview (IBM 2013)



Temporal Modeling

- Motivation:
  - Rich temporal patterns exhibit among visual events.
  - Exploiting temporal dependencies to enhance event detection .



Joint Segmentation and Detection

- Overall Framework:
  - A quadratic integer programming approach combining classification and temporal dependencies between events.
  - For an arbitrary segmentation $S = \{S_1, S_2, ..., S_n\}$ of X where $\mathrm{S}_i = \mathbf{X}(t_i, t_{i+1})$ ($t = \{t_1, t_2, ..., t_{n+1}\}$ are transition points, the quality of the segmentation can be measured by:

$$\mu \sum_{i=1}^{n} \sum_{k=1}^{K} \zeta_i^k \varphi^k(S_i) + (1-\mu) \sum_{j=1}^{n} \sum_{j^{'}=j+1}^{n^{'}} \sum_{k=1}^{K} \sum_{k^{'}=1}^{K} p(k, k^{'}) \zeta_j^k \zeta_{j^{'}}^{k^{'}}$$

$$\begin{aligned} &\mathrm{n}^{'} \leq n \\ &\forall i : \textstyle\sum_{k=1}^{K} \zeta_i^k \leq 1 \\ &\forall i, \forall i^{'}, \forall k, \forall k^{'} : \zeta_i^k + \zeta_{i^{'}}^{k^{'}} \leq 1 \qquad \textbf{if} \qquad S_i \bigcap S_j = 0 \end{aligned}$$

Joint Segmentation and Detection

- Classification Model:
  - Trained discriminatively using multiclass SVM [3] at different window sizes (30, 60, 90 and 120 frames)
  - Non-event is treated as a special null class
- Model Solver:
  - If only first-order dependency is considered, the objective function can be re-written as:

$$f(\mathbf{X}, K) = \mu \sum_{i=1}^{n} \varphi^{k}(S_i) + (1-\mu) \sum_{j=1}^{n} p(j-1, j)$$

  - The problem can be solved by dynamic programming [4], Given any vide flip $\mathrm{X}_{(0,u)}$ with length *u*:

$$\mathrm{f}(\mathrm{X}_{(0,u)}, K) = \mathrm{argmax}_{\mathrm{l_{min}} \leq \mathrm{l} \leq \mathrm{l_{max}}} \mathrm{f}(\mathbf{X}_{(0,\mathrm{u-l})}, \mathrm{K}) + \mathrm{f}(\mathbf{X}_{(\mathrm{u},\mathrm{u-l})}, \mathrm{K}).$$

l_{max} and l_{min} are the detection length of video frames.

[3] K. Crammer and Y. Singer. On the Algorithmic Implementation of Multi-class SVMs, JMLR, 2001.
[4] M. Hoai, Z.-Z. Lan, and F. De la Torre. Joint segmentation and classification of human actions in video. In CVPR, 2011.

Performance Evaluation

Primary Runs Results	IBM 2013			Others' Best 2013		CMU-IBM2012	
	Ranking	ActDCR	MinDCR	ActDCR	MinDCR	ActDCR	MinDCR
CellToEar	1	0.9985	0.9978	1.0069	0.9814	1.0007	1.0003
Embrace	1	0.7873	0.7733	0.8357	0.824	0.8	0.7794
ObjectPut	2	1.0046	1.002	0.9981	0.9783	1.004	0.9994
PeopleMeet	2	1.0267	0.9769	0.9474	0.9177	1.0361	0.949
PeopleSplitUp	1	0.8364	0.8066	0.8947	0.8787	0.8433	0.7882
PersonRuns	2	0.7887	0.7792	0.7708	0.7623	0.8346	0.7872
Pointing	3	1.0045	0.9904	0.9959	0.977	1.0175	0.9921

- Compared to our last year's results based on FV (CMU-IBM 2012):
  - this year's system got improvement over 6/7 events (primary run).
- Compared to other teams' results (Others' Best 2013):
  - our system leads in 3/7 events (primary run).

Outline

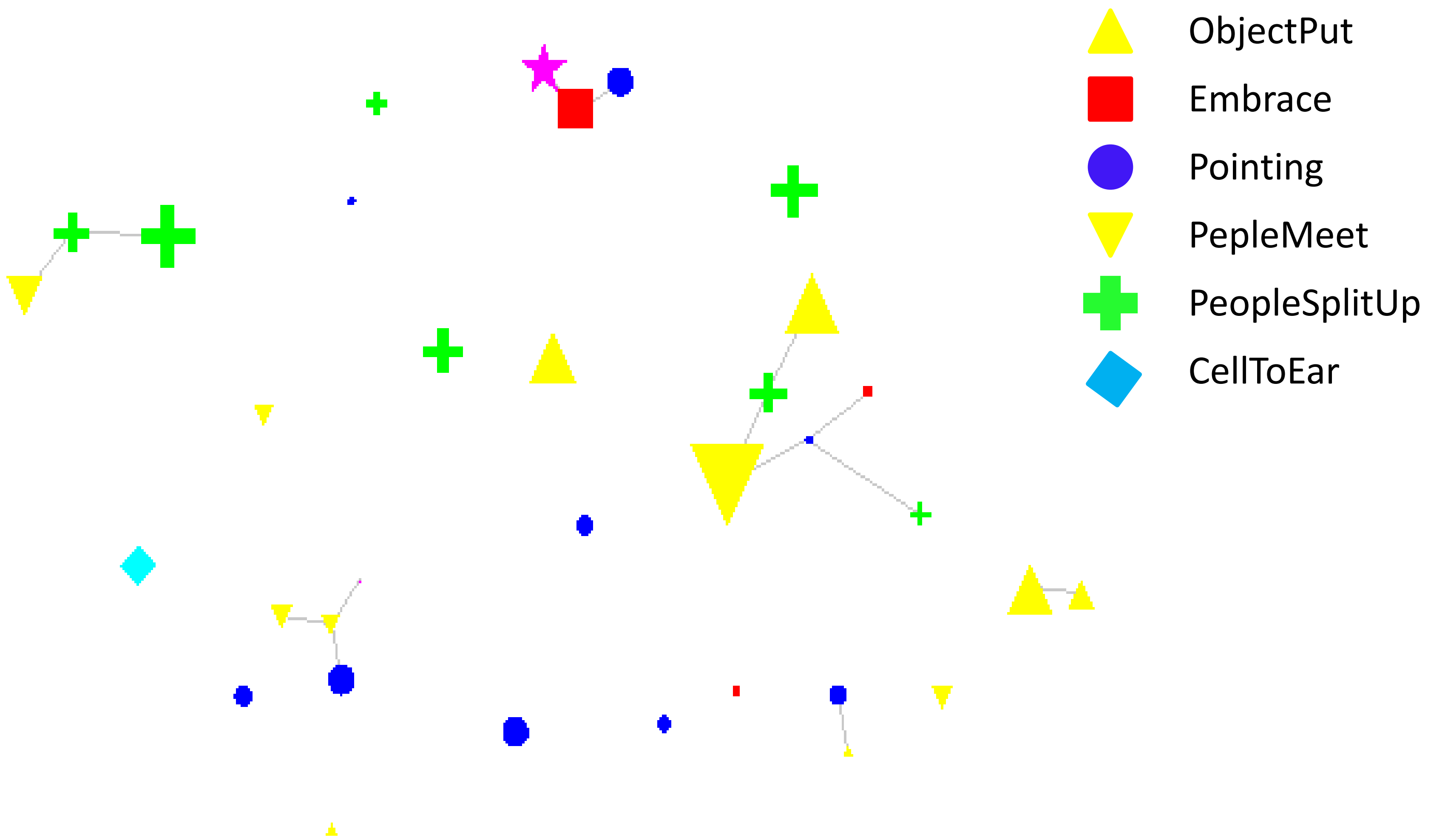
- Retrospective Event Detection
  - System Overview
  - Temporal Modeling for Event Detection
  - Performance Evaluation

- Interactive Event Detection
  - Interactive Visualization
  - Risk Ranking
  - Performance Evaluation

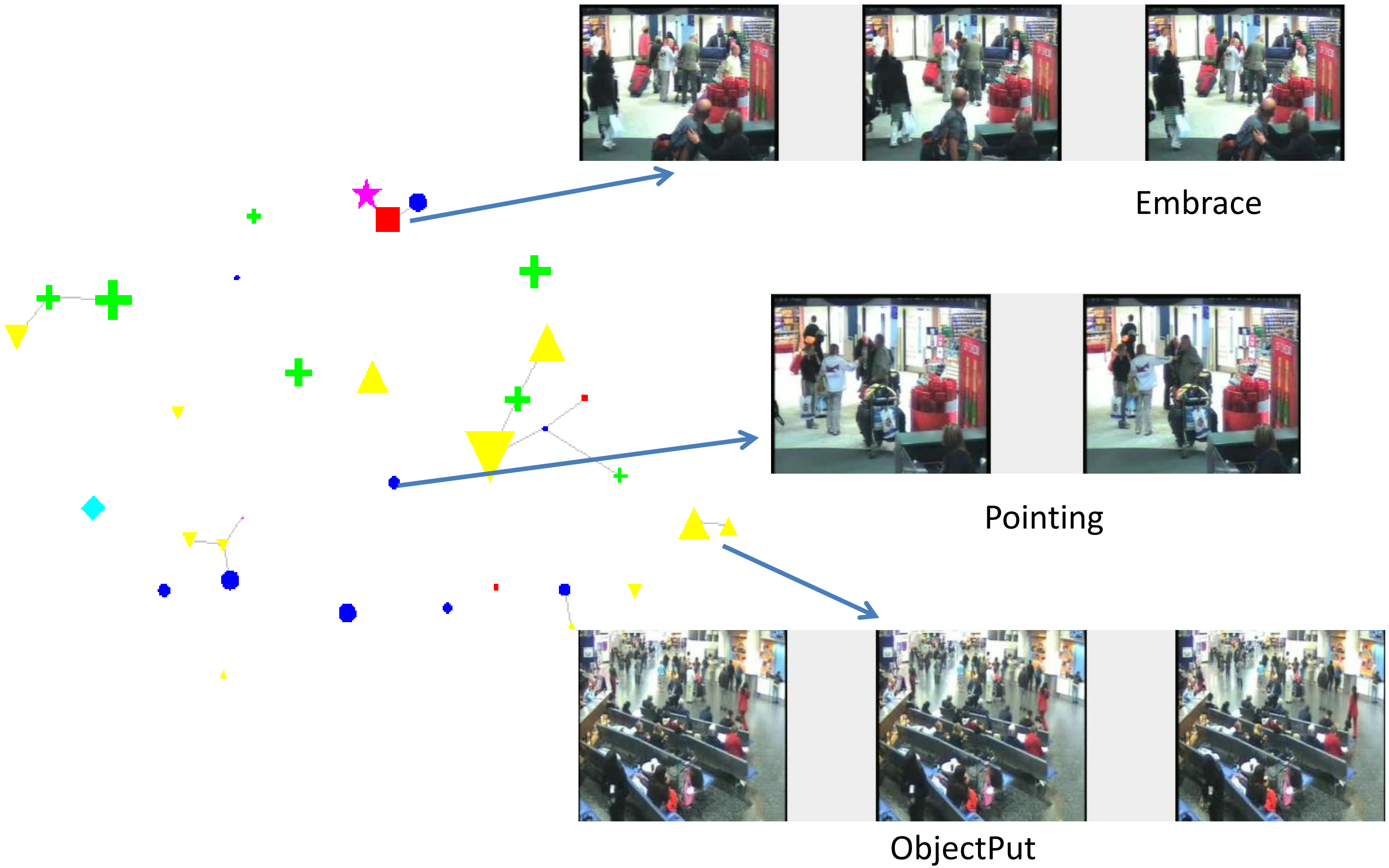
Interactive Visualization

- Motivations:
  - How can we present events to the users more effectively?
    - E.g. two events “peoplemeet” and “pointing” may exist successively. Looking at them together are more beneficial than checking one at each time alone.
  - How can we present more informative events to the users for correction/verification?
    - E.g. correcting mis-detected events is more rewarding. for example, “embrace” → “peoplemeet” vs. “pointing” → “nonevent”.

Event-specific Detection Visualization



Event-specific Detection Visualization

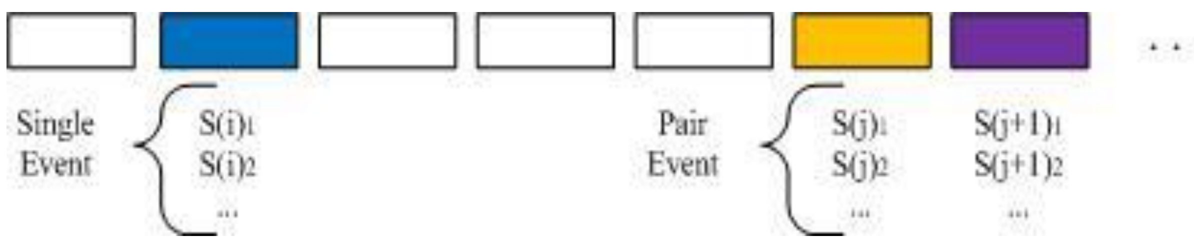


Risk Ranking of Detected Events

- Approach
  - To measure the adjudication risk of each event detected by considering: 1) the margin of the top two candidates in classification; 2) temporal relations and 3) potential gain of DCR
  - Ranking events by their risk scores
  - Checking and re-labeling events from high risk to low risk.

Risk Ranking of Detected Events

– Considering our classification results: for each segmentation S_i we have its top two candidates $\varphi^k(S_i)$ and $\varphi^{k'}(S_i)$, and their priors $p(k)$ and $p(k')$

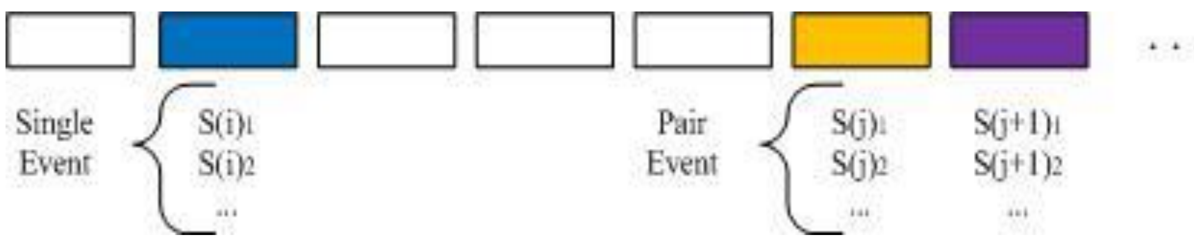


$$R(S_i) = \frac{1 - (\varphi^k(S_i)p(k) - \varphi^{k'}(S_i)p(k'))}{||S_i||} \cdot \begin{cases} w_m \\ w_f \\ w_m + w_f \end{cases}$$

w_m is the cost of a mis-detection and w_f is the cost of a false alarm, $\sum$ is the normalizer. ($w_m = 1, w_f = 0.005$ were set based on DCR)

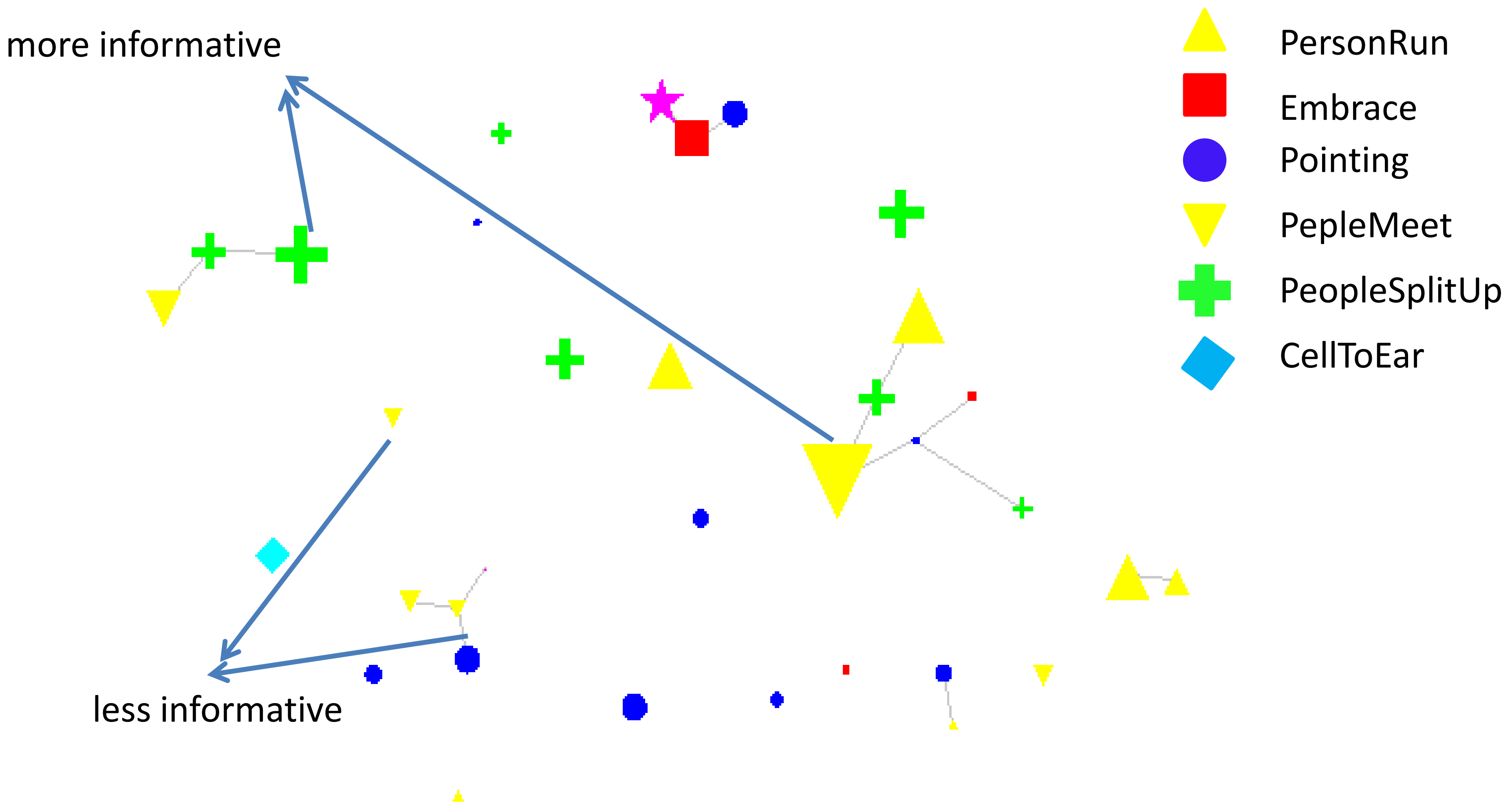
Risk Ranking of Detected Events

– Pair-wise events : for S_i and S_{i+1}, we have $\varphi^{k_j}(S_i)\varphi^{k_{j+1}}(S_{i+1})$ $\varphi^{k'_j}(S_i)\,\varphi^{k'_{j+1}}(S_{i+1})$ and their priors $p(k_j, k_{j+1})$ and $p(k'_j, k'_{j+1})$



$$R(S_i, S_{i+1}) = \frac{1 - ((\varphi^k(S_i) + \varphi^k(S_{i+1}))p(k_j, k_{j+1}) - (\varphi^{k'}(S_i) + \varphi^k(S_{i+1}))p(k'_j, k'_{j+1})))}{\|S_i \bigcup S_{i+1}\|} \cdot \begin{cases} 2 \cdot w_m \\ 2 \cdot w_f \\ 2 \cdot (w_m + w_f) \\ \ldots \end{cases}$$

Risk Ranking of Detected Events



Performance Evaluation

Actual DCR	Evaluation Set (25min * 7)			
	Retro	Risk-1 (primary)	Risk-2	Risk-3
CellToEar	0.9985	0.9956	0.994	1.0013
Embrace	0.7873	0.7337	0.6551	0.6705
ObjectPut	1.0046	0.9928	0.987	1.0053
PeopleMeet	1.0267	0.9584	0.9145	0.9684
PeopleSplitUp	0.8364	0.8489	0.8304	0.8924
PersonRuns	0.7887	0.7188	0.6865	0.7588
Pointing	1.0045	0.9781	0.974	0.9877

- **Retro**: retrospective event detection
- **Risk-1**: independent evaluation by risk ranking (25 mins for each event type)
- **Risk-2**: joint evaluation by risk ranking (a total of 175 mins)
- **Risk-3**: independent evaluation using classification scores

Risk-2 > Risk-1 > Risk-3 >Retro

Discussions

- A few thoughts
  - ground truth (automatic, crowdsourcing,…)?
  - Independent and/or dependent evaluation?

Conclusions

- Retrospective System:
  - Joint-segmentation-classification provides a promising schema for surveillance event detection.
  - Modeling temporal relations between events can boost the detection performance.
- Interactive System:
  - Event visualization with strong temporal patterns is a more efficient presentation for an interactive system.
  - Risk-based ranking demonstrates its effectiveness in relabeling events.

References:

- [1] M. yu Chen and A. Hauptmann. Mosift: Recognizing human actions in surveillance videos. In CMU-CS-09-161, 2009.

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

- [3] K. Crammer and Y. Singer. On the Algorithmic Implementation of Multi-class SVMs, JMLR, 2001.

- [4] M. Hoai, Z.-Z. Lan, and F. De la Torre. Joint segmentation and classification of human actions in video. In CVPR, 2011.

Future Works

- Retrospective System:
  - Exploiting long distance temporal relations into this joint-segmentation-detection framework.
  - Exploring the performance trade-offs between localization and categorization.

- Interactive System:
  - Better visualization layout need to be developed, E.g. time layout.
  - Various risk ranking methods need to be tried.
  - User feedback utilization methods need to be incorporated. E.g. interactive learning.

Multiple Detections Visualization

- Objective:
  - To find visualization methods that enable multiple events representation.
- Solution:
  - Visualize the events in a graph-based layout: each node is an individual event and the edge between them representing the temporal relation.

Outline

- Retrospective Event Detection
  - System Overview
  - Temporal Modeling for Event Detection
  - Performance Evaluation
- Interactive Event Detection
  - Interactive Visualization
  - Risk Ranking
  - Performance Evaluation