CMU-Informedia @ TRECVID 2011 Surveillance Event Detection

Longfei Zhang , Lu Jiang , Lei Bao, Shohei Takahashi, Yuanpeng Li,

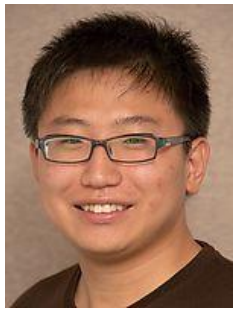
Alexander Hauptmann

Carnegie Mellon University

SED11 Team

- Team members:

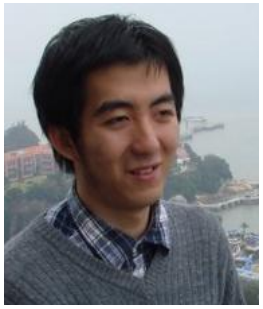
Longfei



Lu

Lei

Shohei



Yuanpeng

Alex

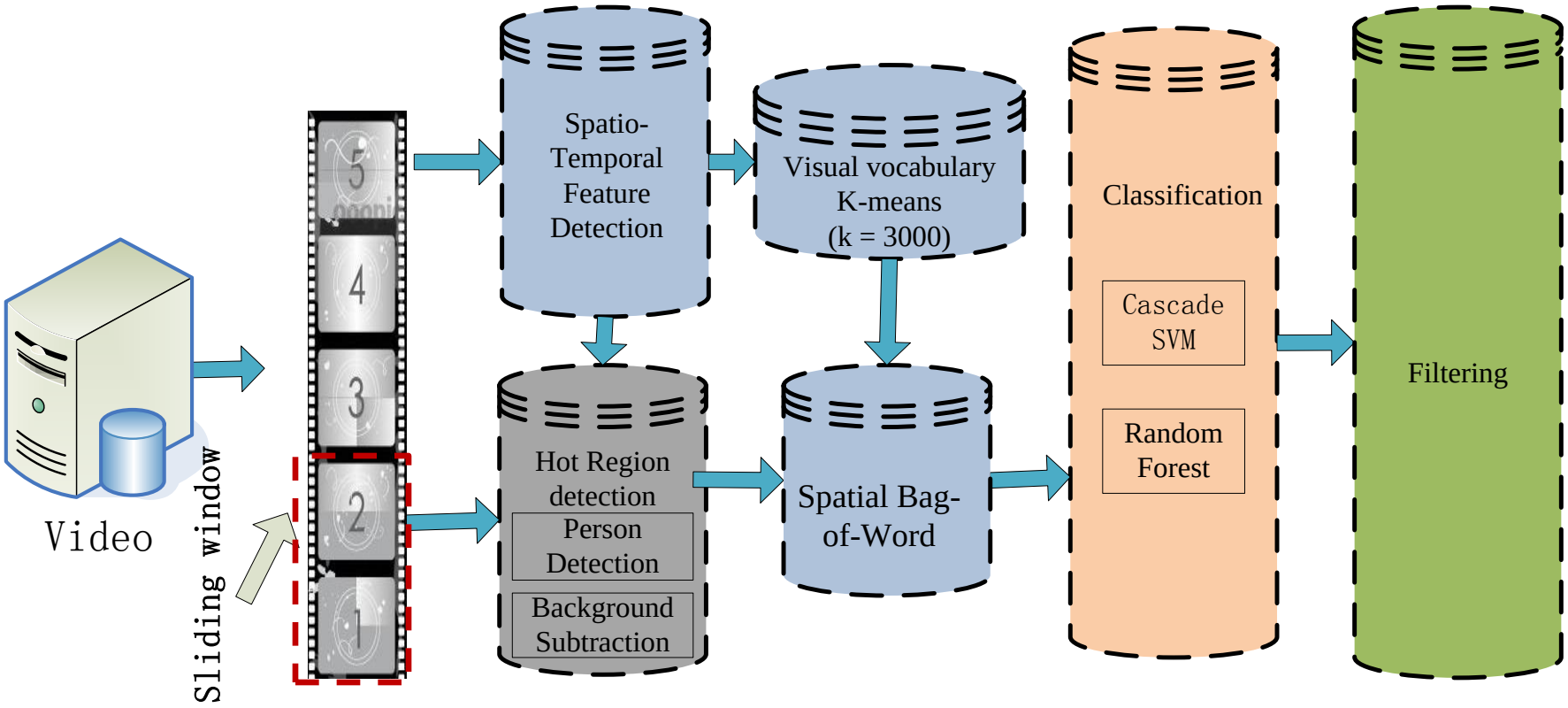
Carnegie Mellon University

Outline

- Framework
- MoSIFT based Action Recognition
  - MoSIFT feature
  - Spatial Bag of Word
  - Tackling highly imbalanced datasets
- Experiment Results

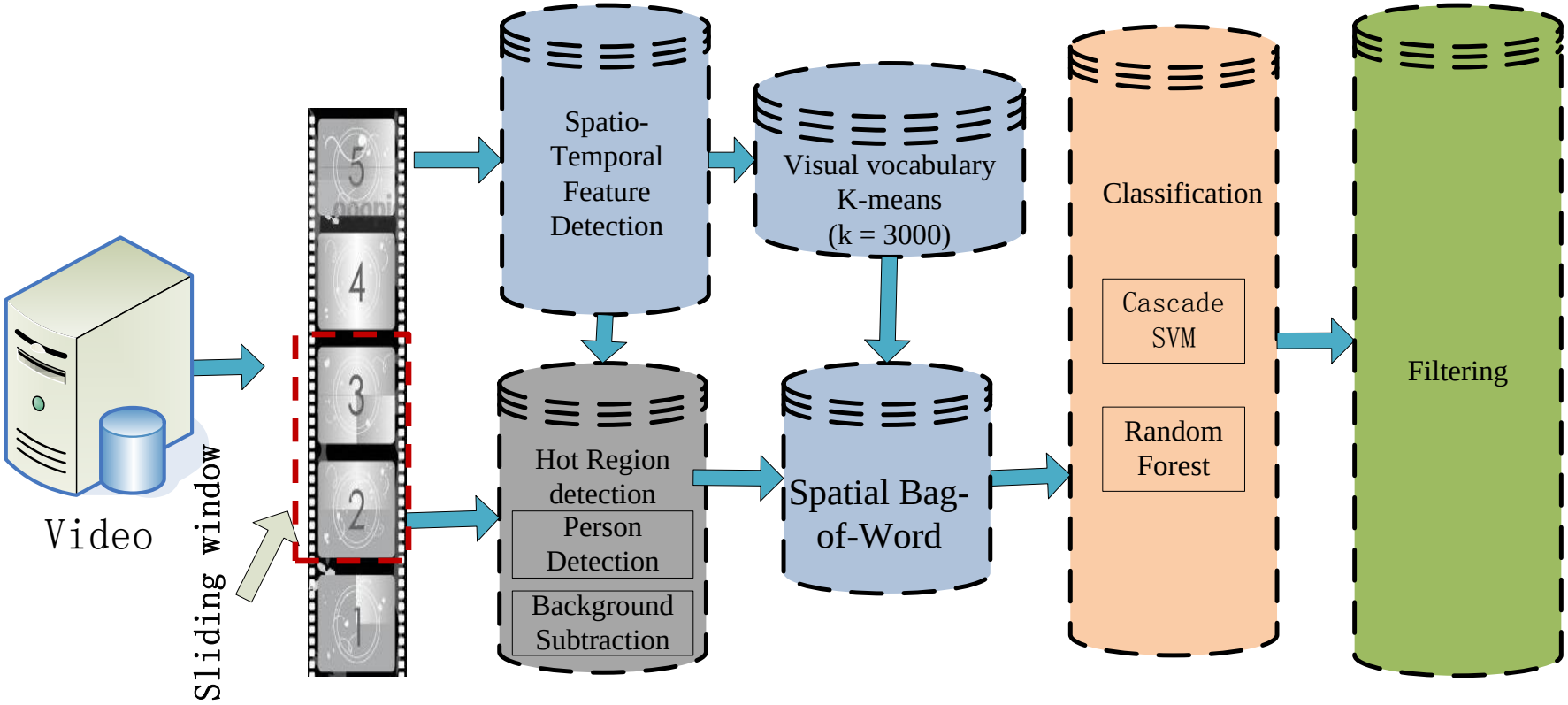
Framework

- **Augmented Boosted Cascade**



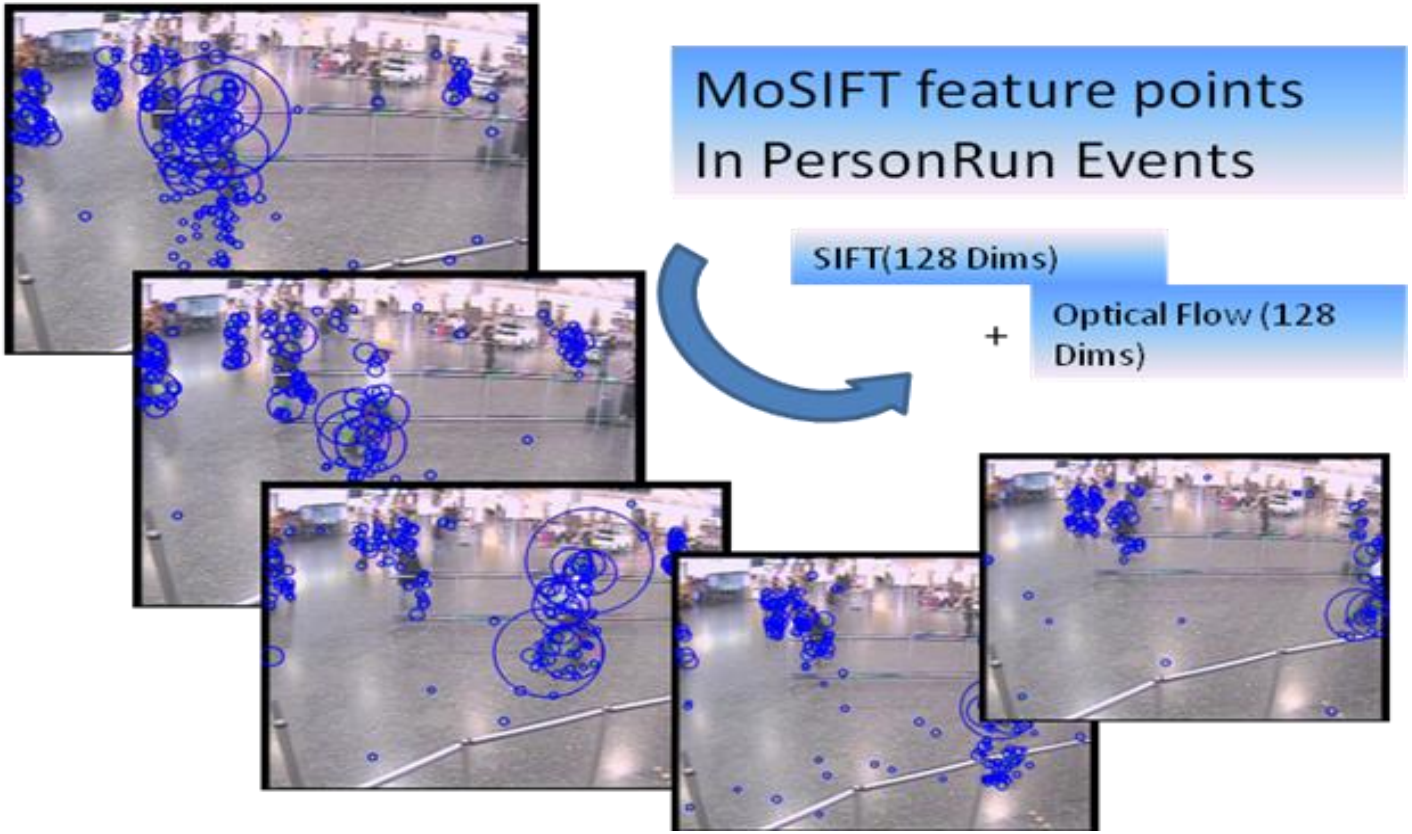
Framework

- **Augmented Boosted Cascade**



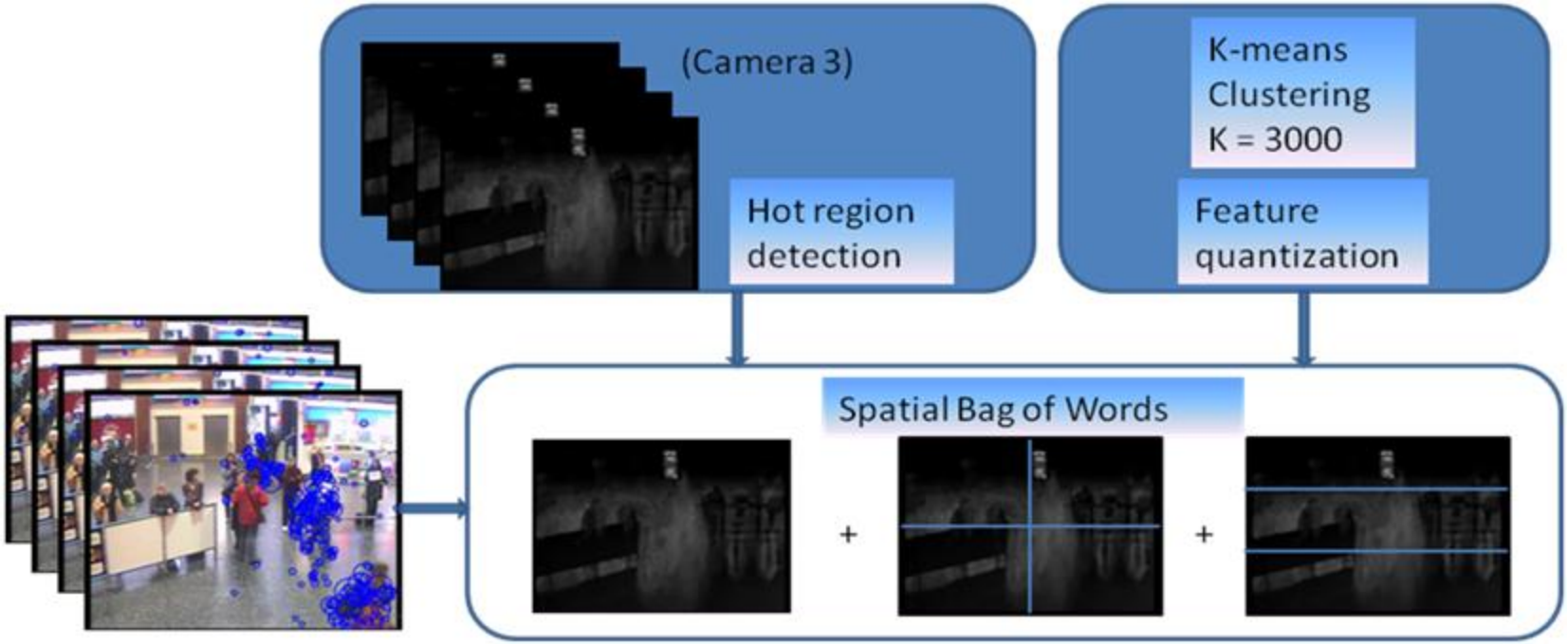
MoSIFT

- Given pairs of video frames, detect spatio-temporal interest points at multiple scales.
  - SIFT point detection with sufficient optical flow.
  - Describing SIFT points through SIFT descriptor and optical flow.



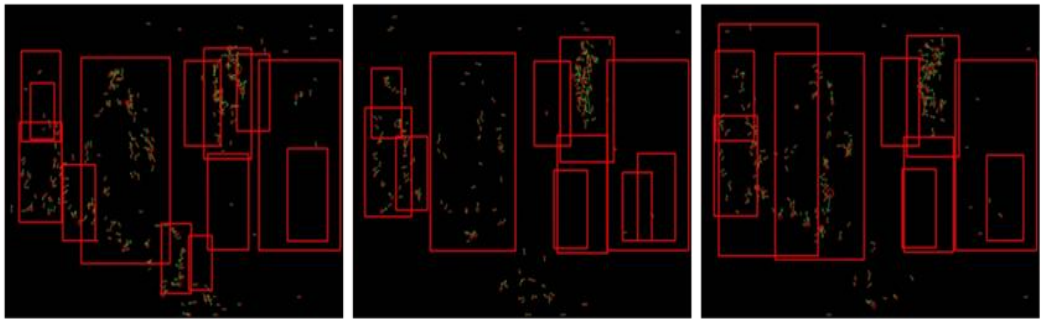
Spatial Bag of Words

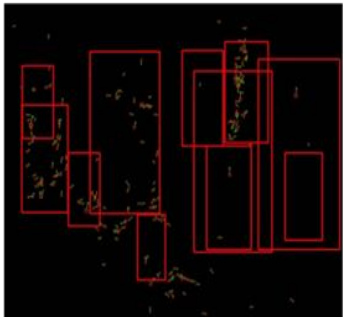
- Each frame is divided into a set of non-overlapping rectangular tiles.
- The resulting BoW features are derived by concatenating the BoW features captured in each tile.
- Encode the spatial (tile) information in BoW.



Hot Region Detection

- Person Detection: Person detection based on Histogram of Oriented Gradient (HOG) features.
- Background subtraction.

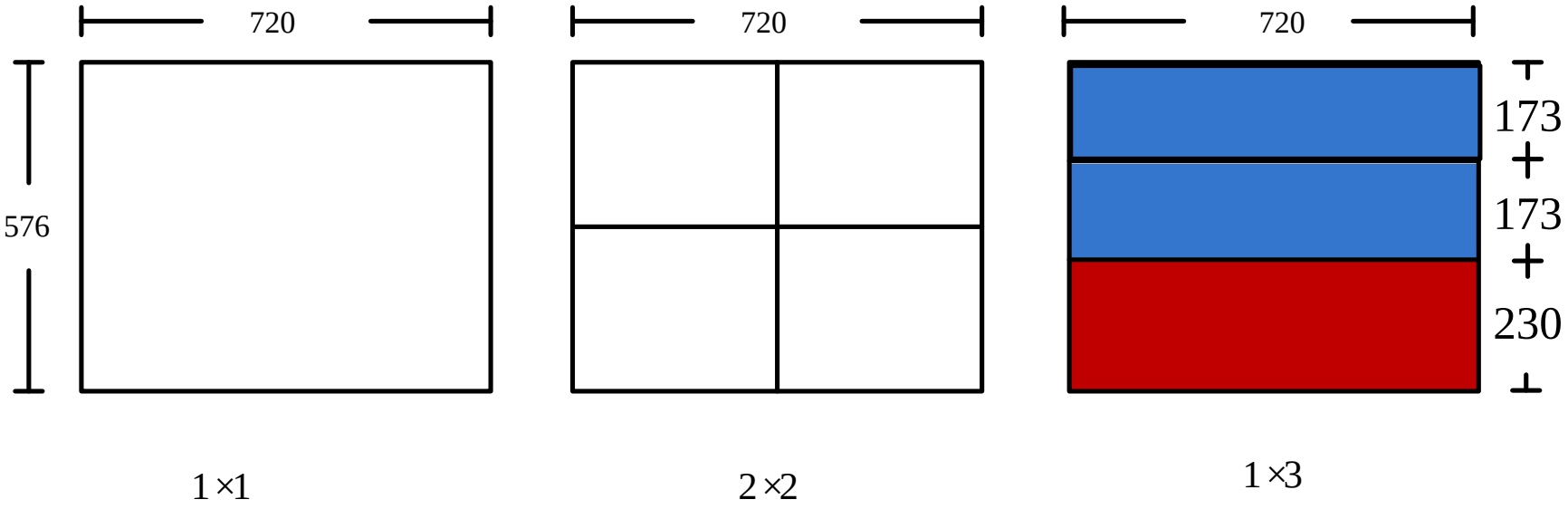




Over generated Person detection results for tracking and feature selection

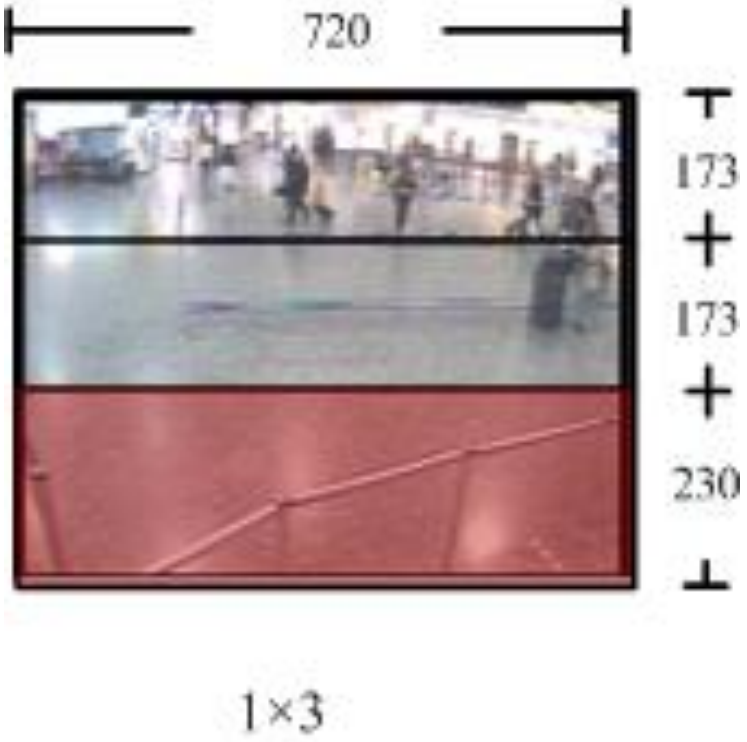
Spatial Bag of Features

- Each frame is divided into a set of rectangular tiles or grids.
- The resulting Bow features are derived by concatenating the BoW features captured in each grid.
- Encode the adjusted spatial information in BoW.



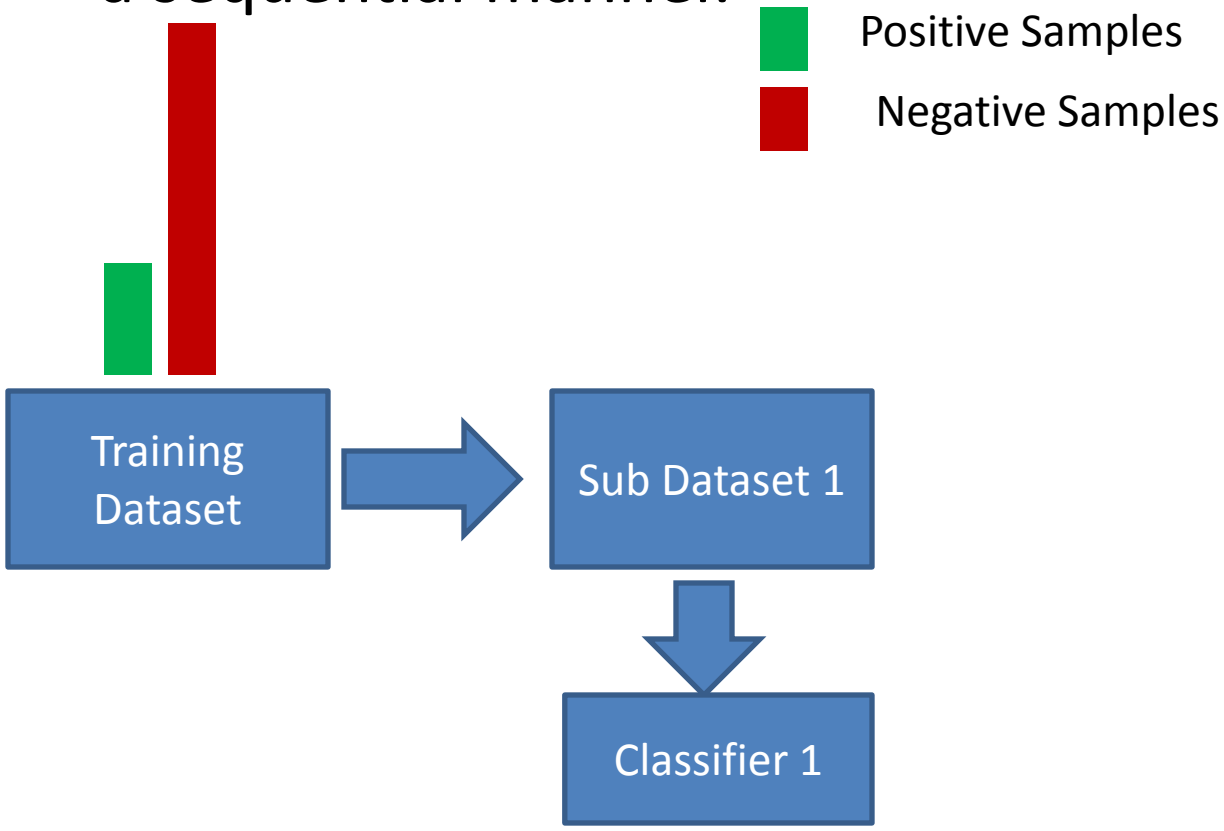
Spatial Bag of Features

- Each frame is divided into a set of rectangular tiles or grids.
- The resulting Bow features are derived by concatenating the BoW features captured in each grid.
- Encode the adjusted spatial information in BoW.



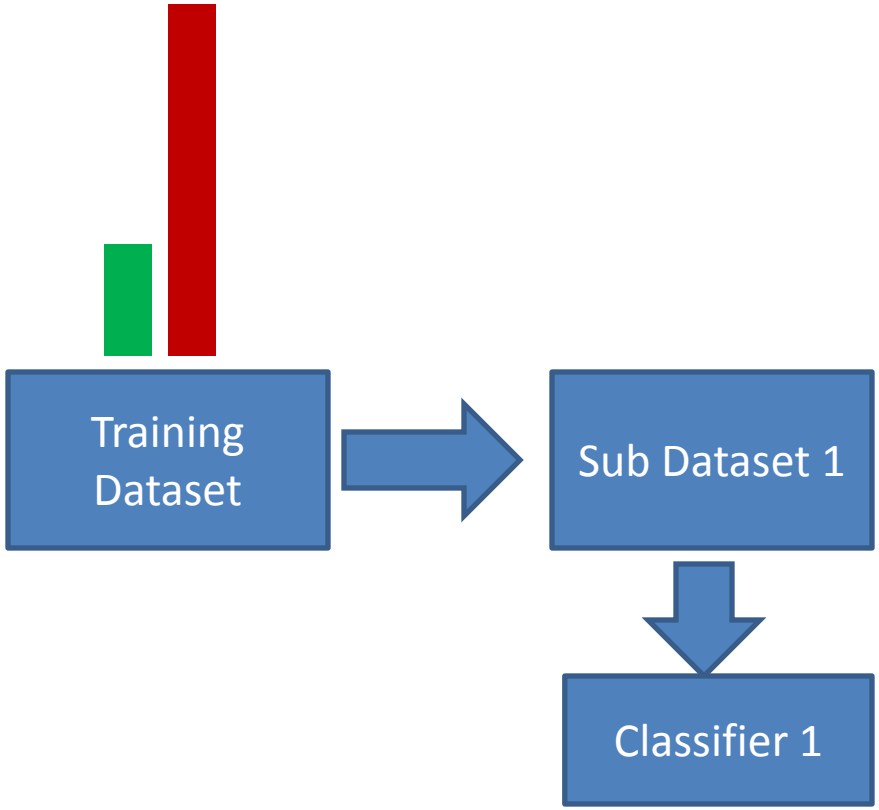
Tackling the highly imbalanced data

- Augmented Cascade SVM.
- Bagging classification method except it adopts probabilistic sampling to select negative samples in a sequential manner.



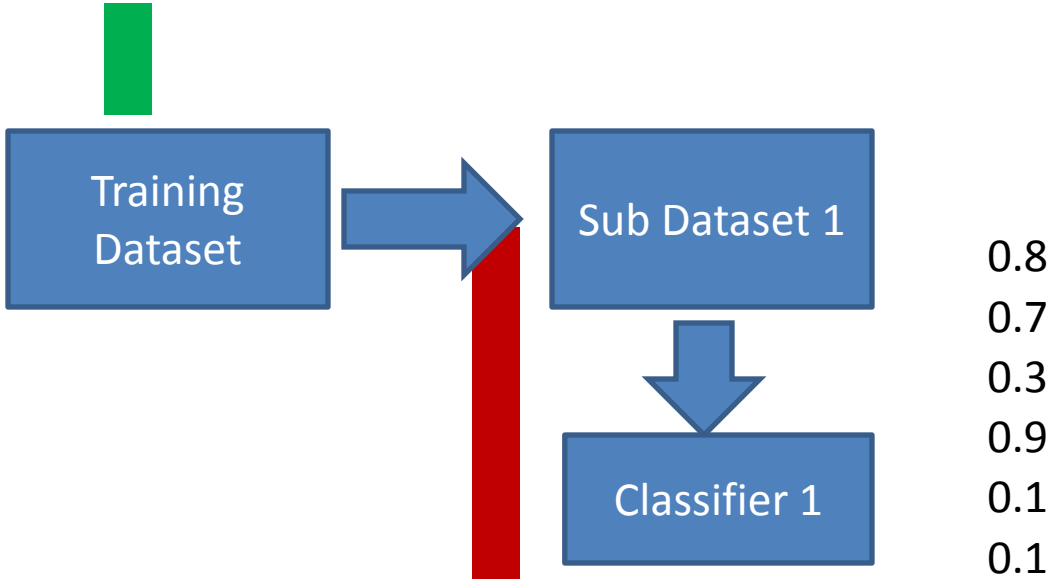
Tackling the highly imbalanced data

- Augmented Cascade SVM.
- Bagging classification method except it adopts probabilistic sampling to select negative samples in a sequential manner.



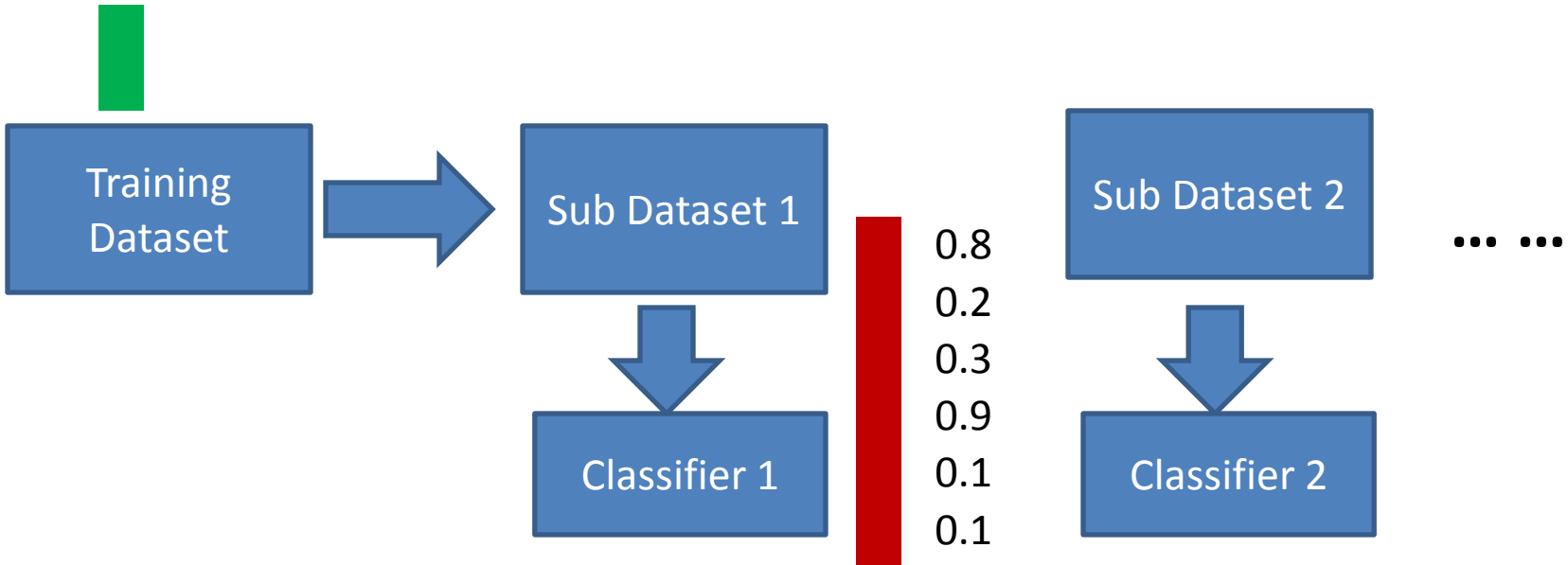
Tackling the highly imbalanced data

- Augmented Cascade SVM.
- Bagging classification method except it adopts probabilistic sampling to select negative samples in a sequential manner.



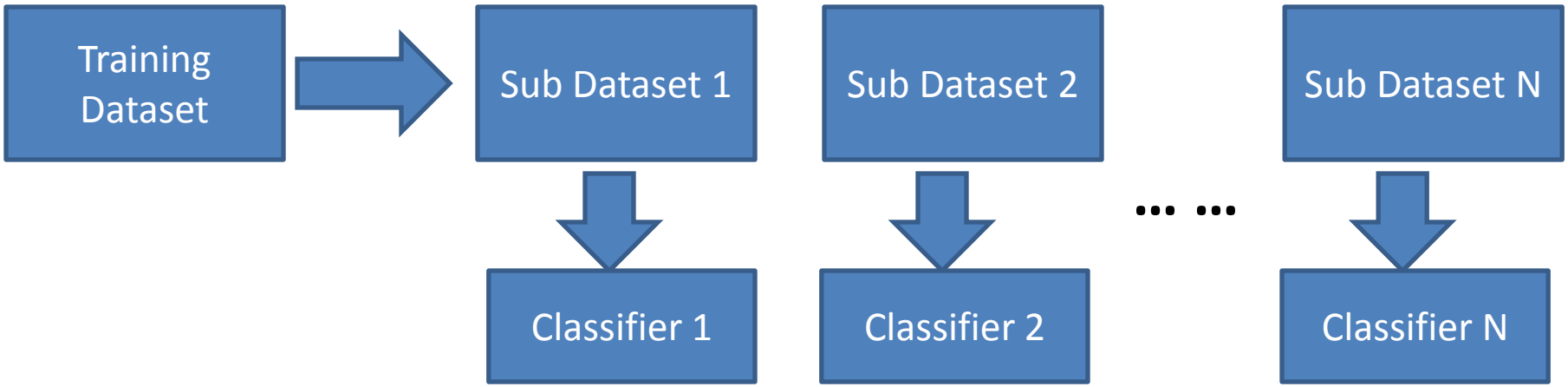
Tackling the highly imbalanced data

- Augmented Cascade SVM.
- Bagging classification method except it adopts probabilistic sampling to select negative samples in a sequential manner.



Tackling the highly imbalanced data

- Augmented Cascade SVM.
- Bagging classification method except it adopts probabilistic sampling to select negative samples in a sequential manner. N = 10 layers.



Carnegie Mellon University

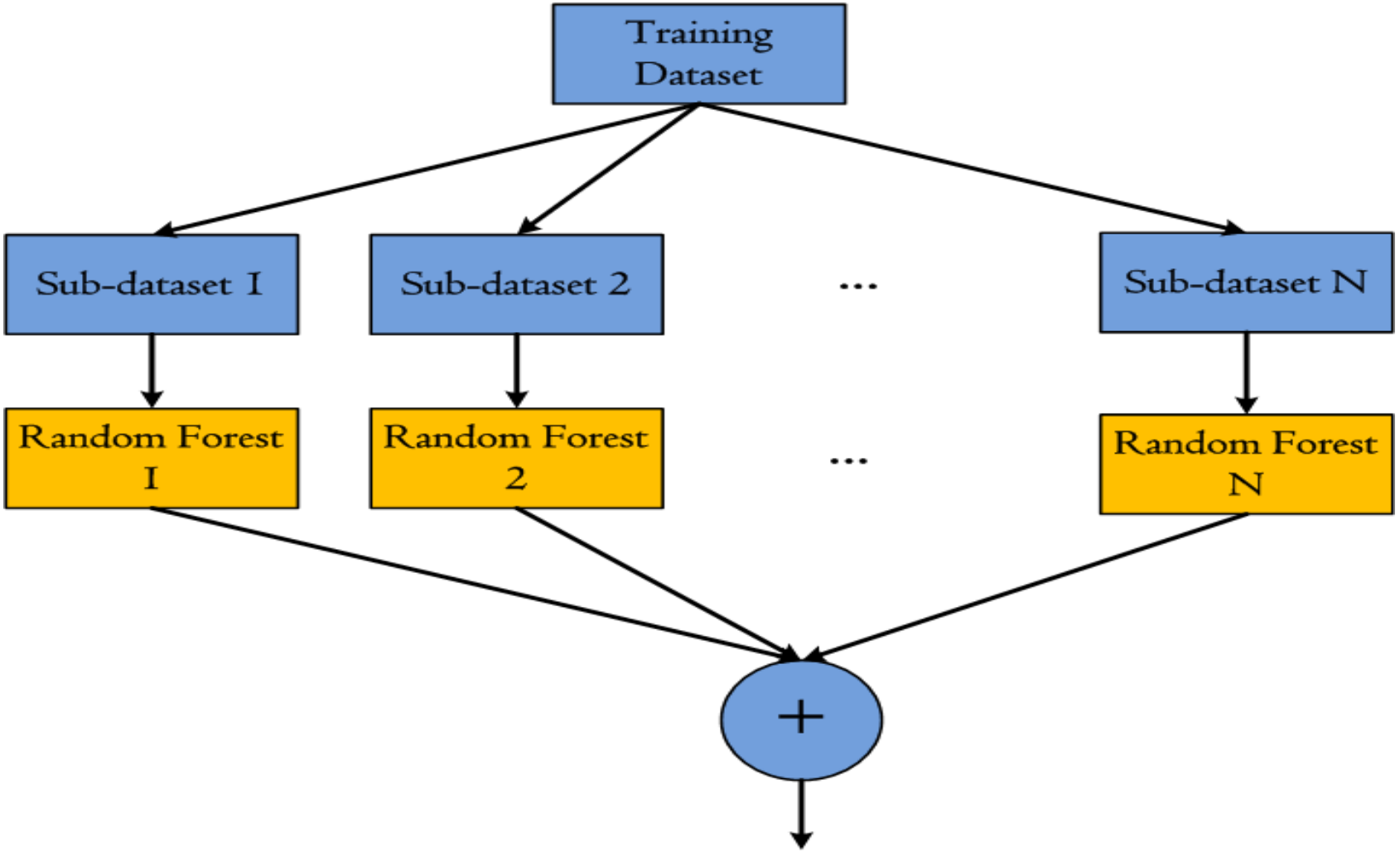
Tackling highly imbalanced data

Bagging Ensemble of Random Forests

- Random Forest is a forest of decision trees.
- Two parameters:
  - n is the number of trees in the forest.
  - m the number of features in each decision tree.
- Build each decision tree by randomly selecting m features and use C4.5.
- Each tree is grown without pruning.

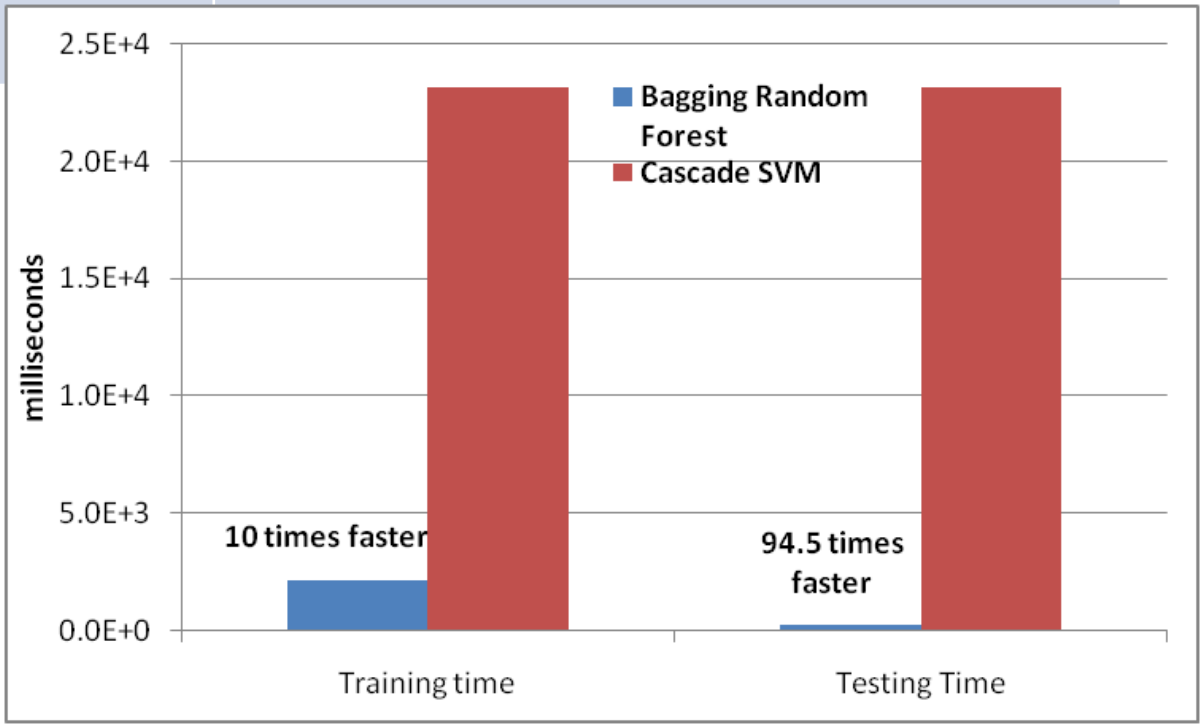
Tackling highly imbalanced data

Bagging Random Forest: Ensemble of Random Forests



Cascade SVM vs. Bagging Random Forest

	Cascade SVM (chi² kernel)	Bagging Random Forest
Effectiveness	Most Effective	Usually 3-8% less in Average Precision
Efficiency	Time consuming	Usually tens to hundreds of times faster
Sensitive to Parameter settings	Sensitive	Relatively insensitive



Carnegie Mellon University

Results

- 8 Submissions:
  - The first 6 runs use cascade SVM with different sliding window sizes and parameter sets.
  - Last 2 runs use bagging random forest method.

Carnegie Mellon University

Results

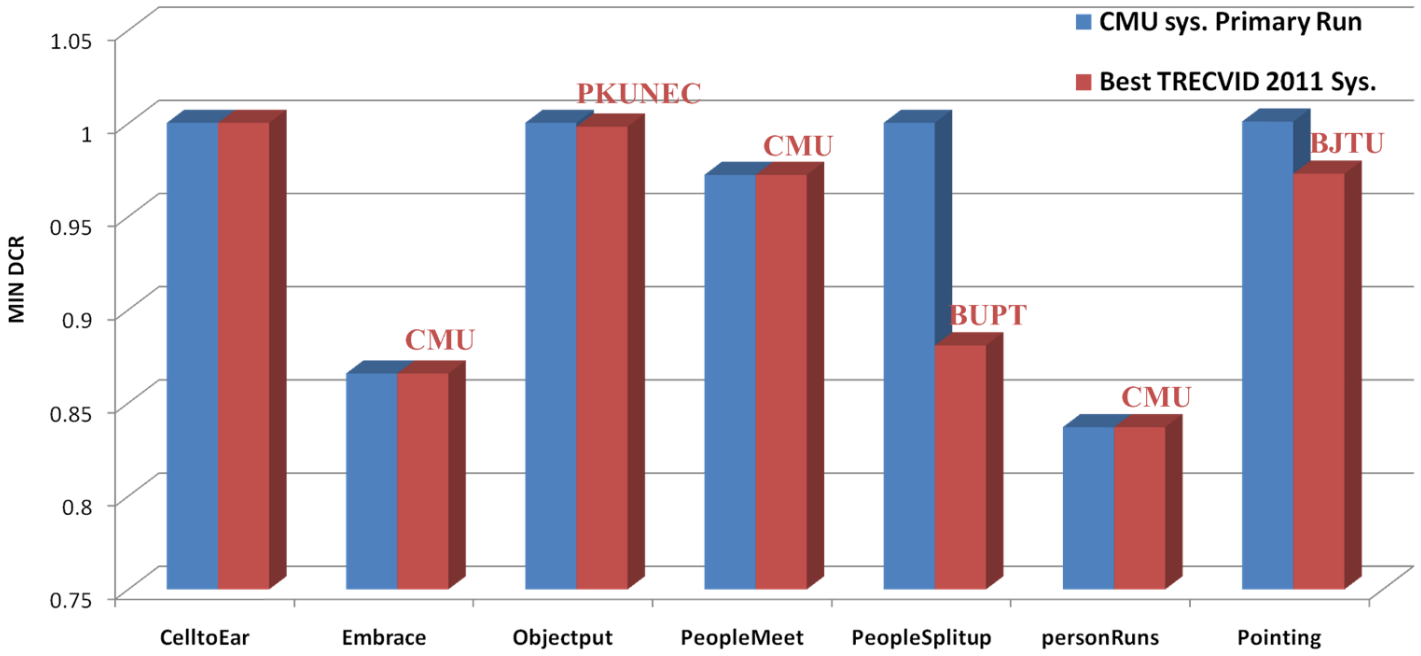
- Results for **Primary** run:

	Inputs			Actual DCR					Minimum DCR
	#Targ	#NTarg	#Sys	#CorDet	#CorDet	#FA	#Miss	DCR	DCR
CellToEar	194	127	128	1	0	127	193	1.0365	1.0003
Embrace	175	657	715	58	0	657	117	0.8840	**0.8658**
ObjectPut	621	57	58	1	0	57	620	1.0171	1.0003
PeopleMeet	449	336	381	45	0	336	404	1.0100	**0.9724**
PeopleSplitUp	187	115	118	3	0	115	184	1.0217	1.0003
PersonRuns	107	413	439	26	0	413	81	0.8924	**0.8370**
Pointing	1063	1960	2092	132	0	1960	931	1.5186	1.0001

Results

Compared with our primary run with those of other teams.

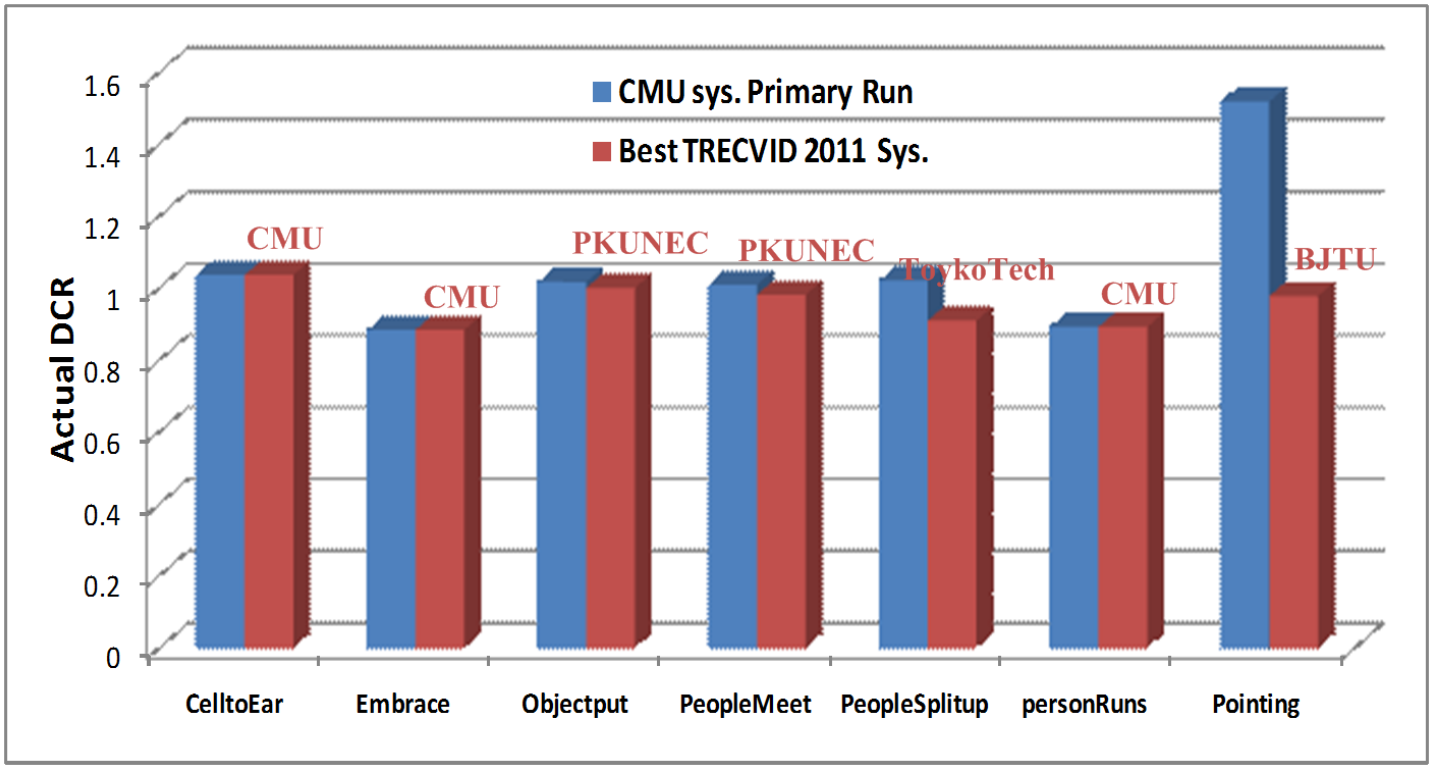
We have the best Min DCR in 3 out of 6 events.



Results

Compared with our primary run with those of other teams.

We have the best Actual DCR in 3 out of 7 events.



Results

Compared with our last year's result, we get improvement in terms of MIN DCR in 5 events "Embrace", "People Meet", "People Slit up", "Person Runs" and "Pointing".

- Best event results over all CMU runs

Min DCR	Cell To Ear	Embrace	Object Put	People Meet	People Split Up	Person Runs	Pointing
2010 CMU	1.0003	0.9838	1.0003	0.9793	0.9889	0.9477	1.0003
2010 Overall Best Event	1	0.9663	0.9971	0.9787	0.9889	0.6818	0.996
2011 CMU	1.0003	0.8658	1.0003	0.9684	0.7838	0.837	0.9996

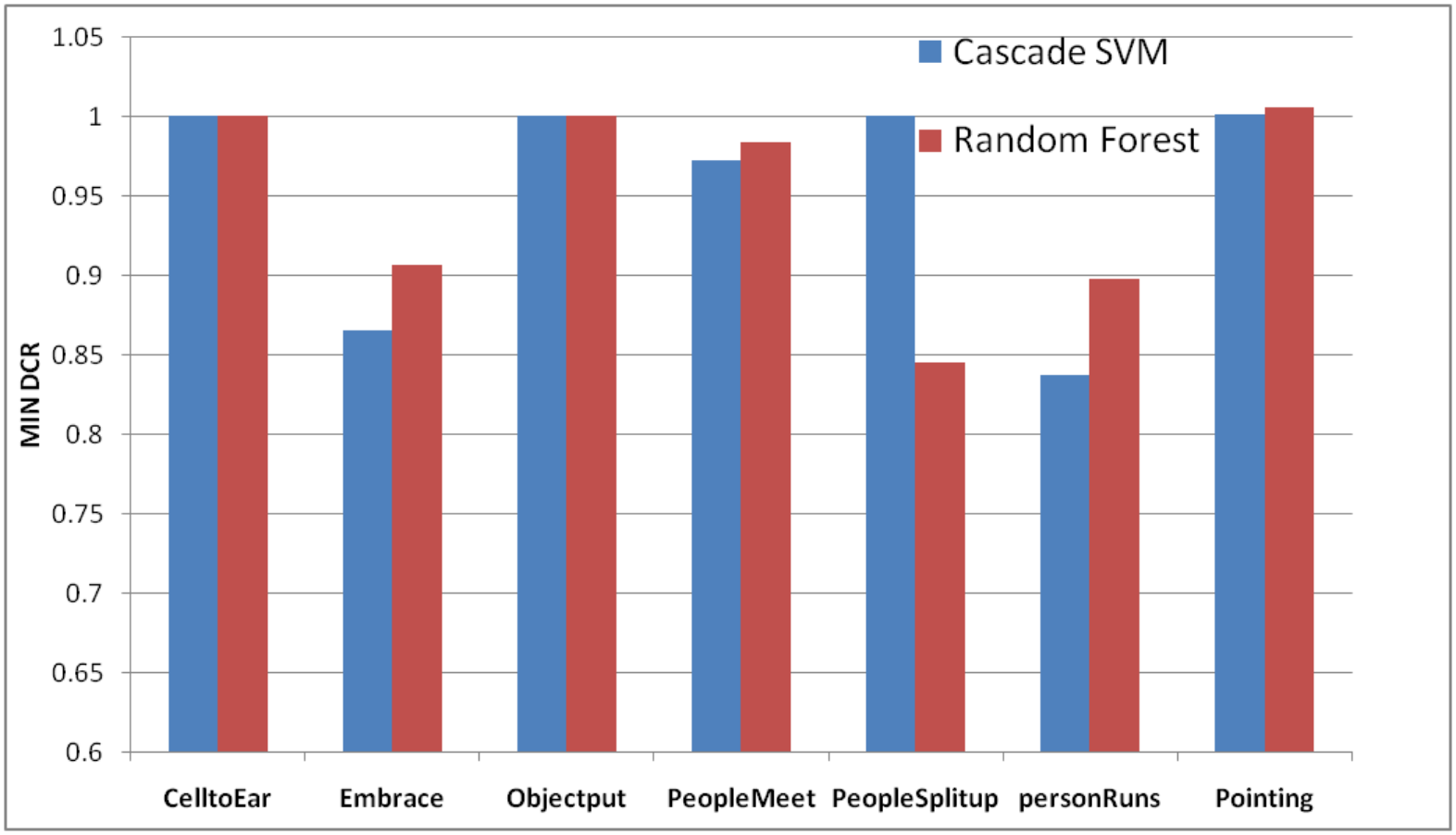
Results

Compared with the best event results in TRECVID 2010, for event "Embrace", "PeopleMeet" and "People Split Up" ours are the best system.

Min DCR	Cell To Ear	Embrace	Object Put	People Meet	People Split Up	Person Runs	Pointing
2010 CMU	1.0003	0.9838	1.0003	0.9793	0.9889	0.9477	1.0003
2010 Overall Best Event	1	0.9663	0.9971	0.9787	0.9889	0.6818	0.996
2011 CMU	1.0003	0.8658	1.0003	0.9684	0.7838	0.837	0.9996

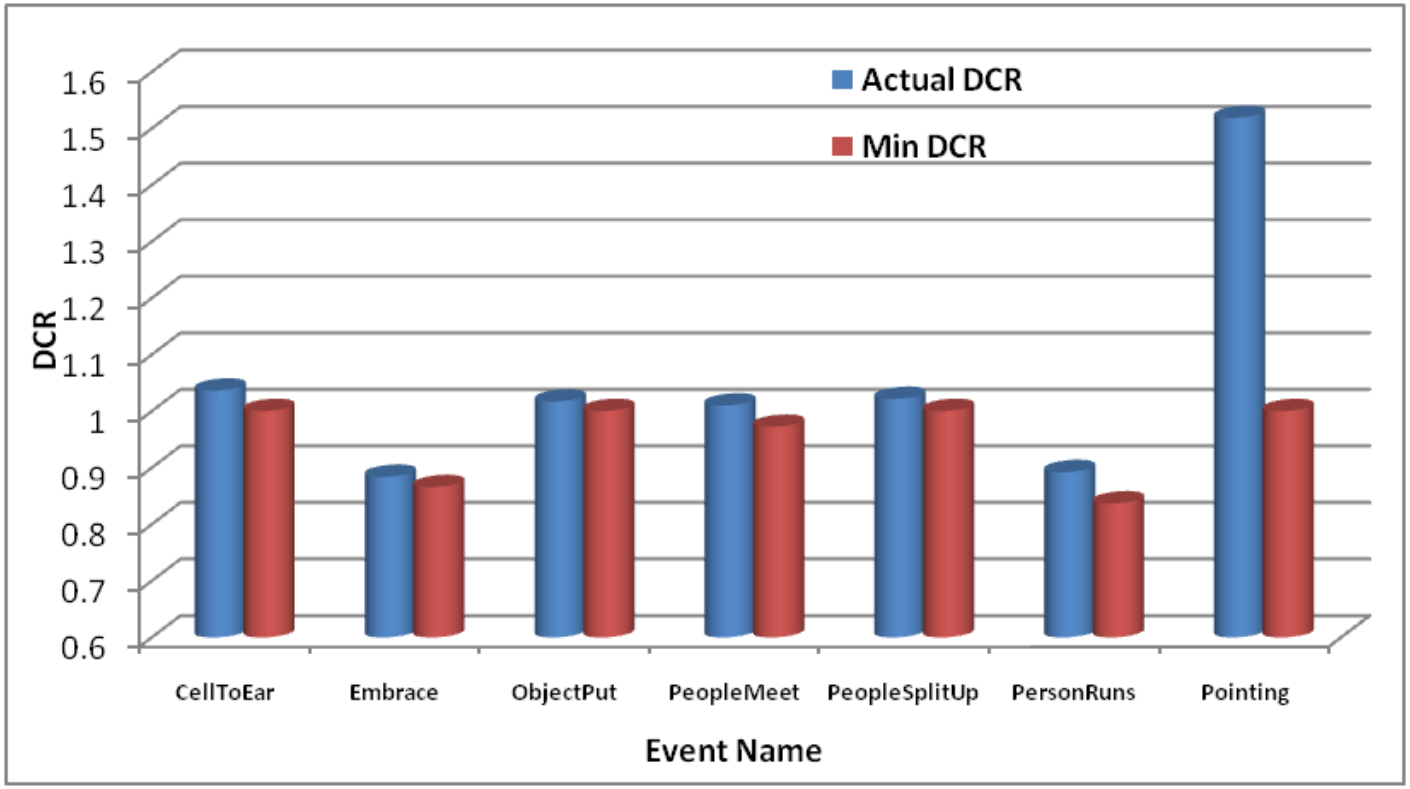
Cascade SVM vs. Random Forest

- Comparison between Run 1 (Cascade SVM) and Run 7 (Random Forest) in terms of Min DCR.



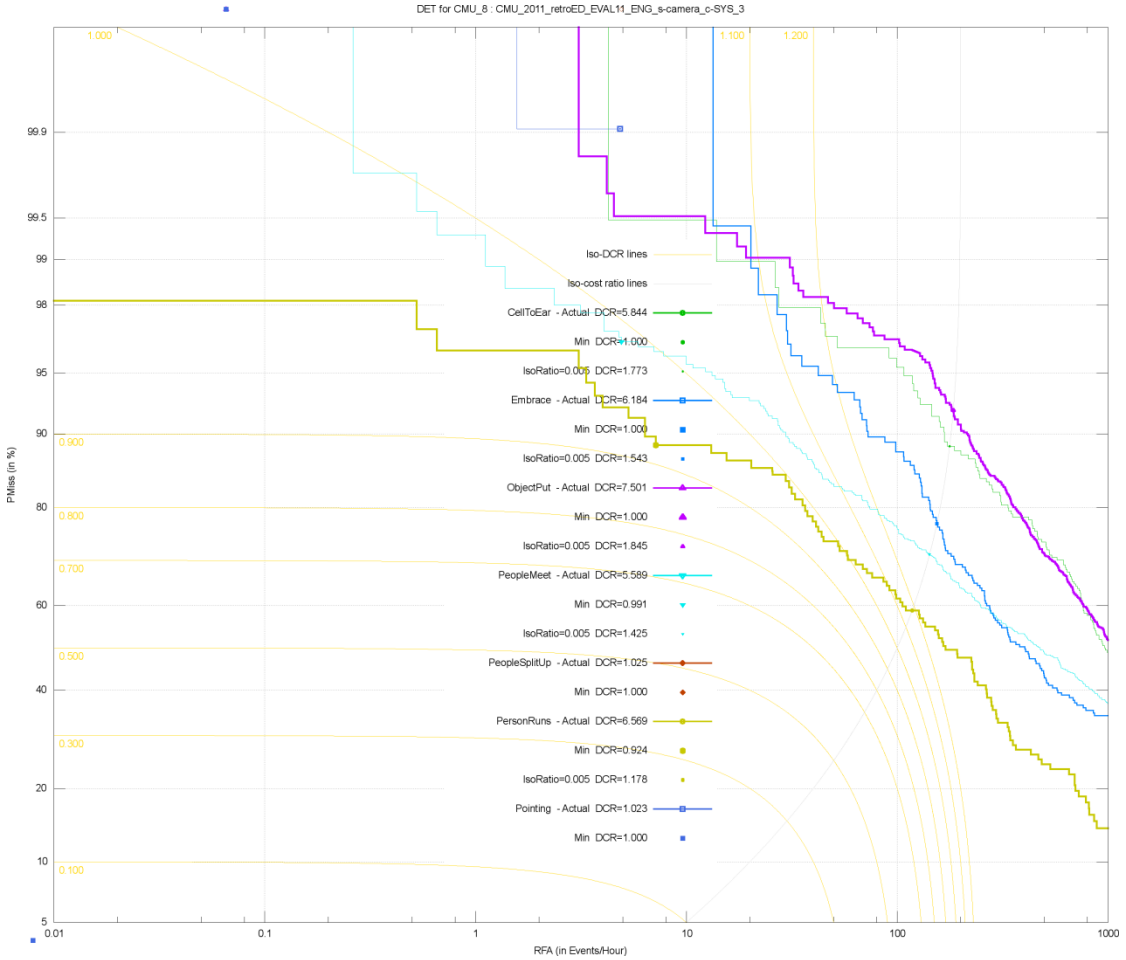
Threshold Search

- Searching for Min DCR using cross validation.
- Actual DCR provides reasonable estimates of Min DCR on all runs.



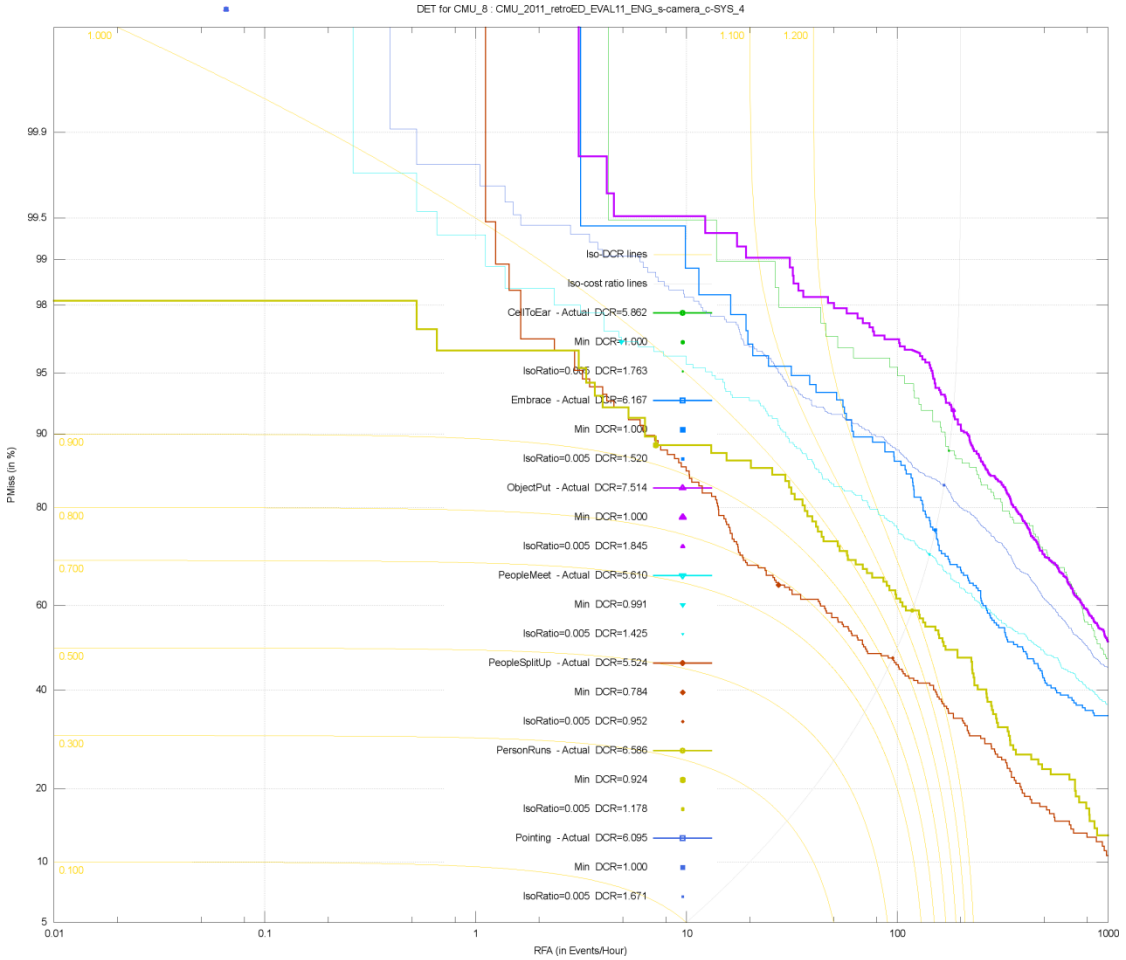
Impact of sliding window size

- Results for all events with sliding window size 25 frames (Run 3).



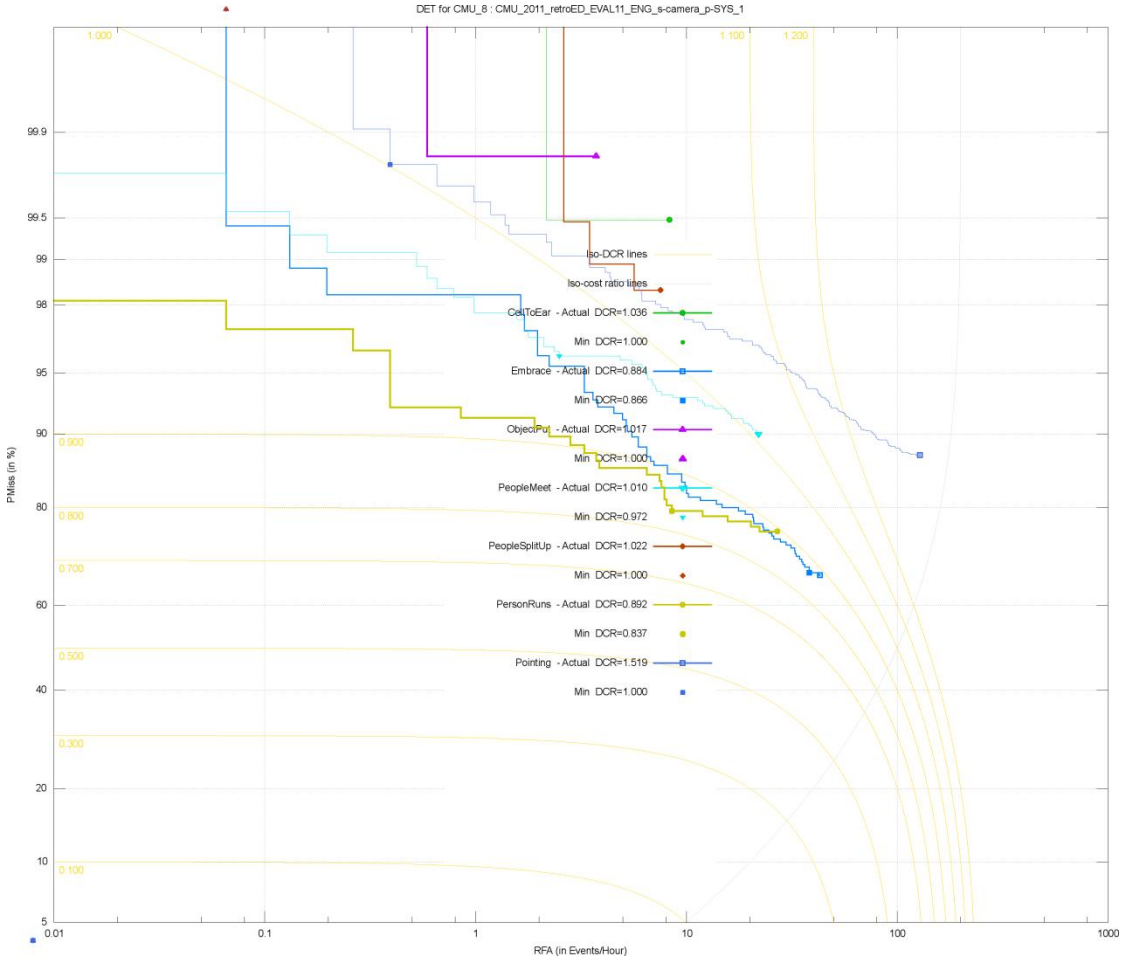
Impact of sliding window size

- Results for all events with sliding window size 60 (Run 5).



Event-specific sliding window size

- For PersonRuns, CellToEar, Embrace and Pointing a good sliding window is small.
- For Embrace, ObjectPut and PeopleMeet a good sliding window size is larger.



Carnegie Mellon University

Conclusions

- Observations:
  - MoSIFT feature captures salient motions in videos.
  - Spatial Bag of Words can boost the performance over last year's result.
  - Event-specific sliding window size impacts the final result.
  - Both cascade SVM and bagging random forest can handle highly imbalanced data sets. Random forest is much faster.

THANK YOU.
Q&A?