Columbia-UCF MED2010: Combining Multiple Modalities, Contextual Concepts, and Temporal Matching

Yu-Gang Jiang[1], Xiaohong Zeng[1], Guangnan Ye[1], Subh Bhattacharya[2], Dan Ellis[1], Mubarak Shah[2], Shih-Fu Chang[1]

[1] Department of EE, Columbia University

[2] Department of EECS, University of Central Florida

TRECVID 2010 workshop, NIST, Gaithersburg, MD

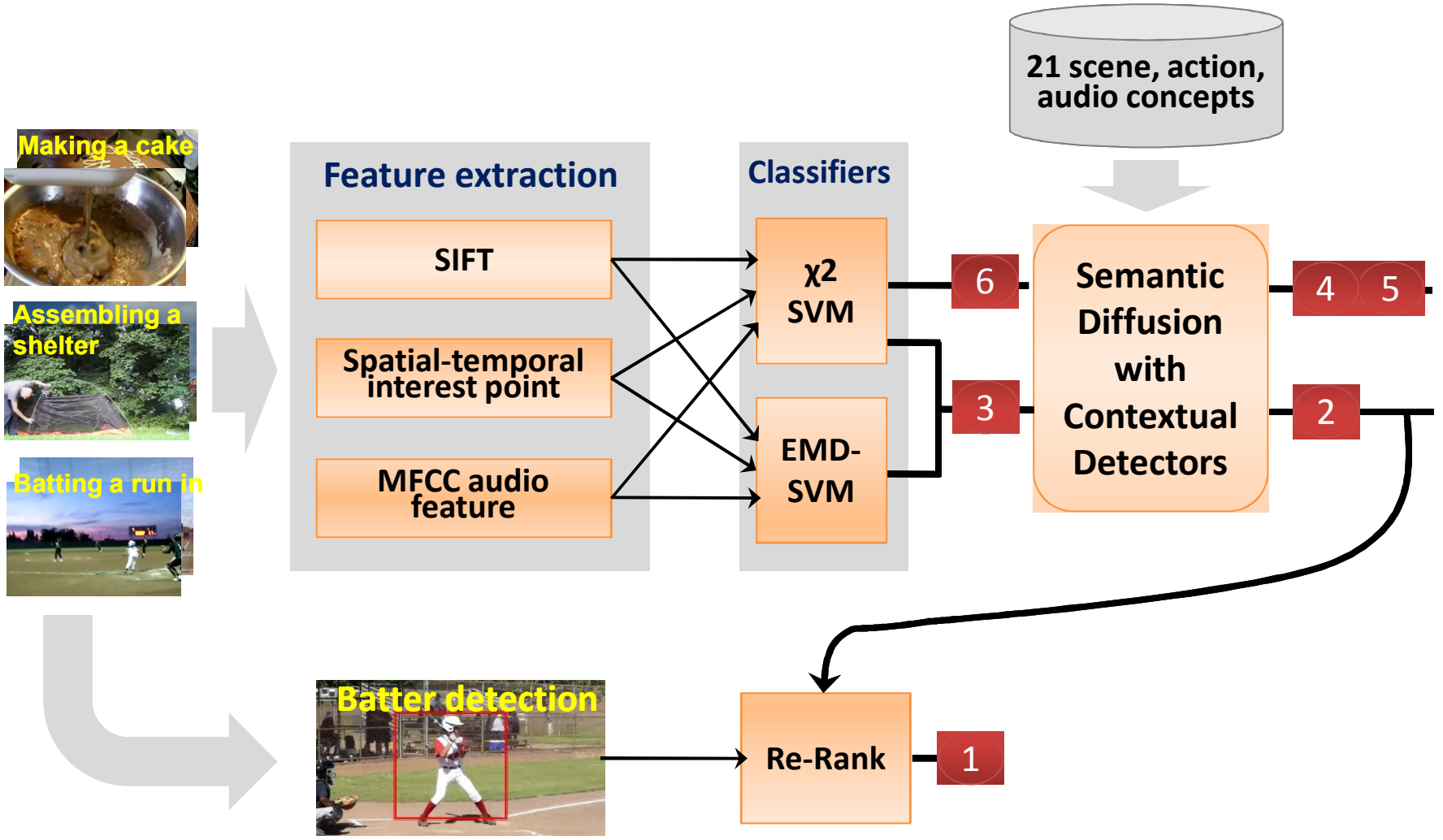
The target…

Making a cake

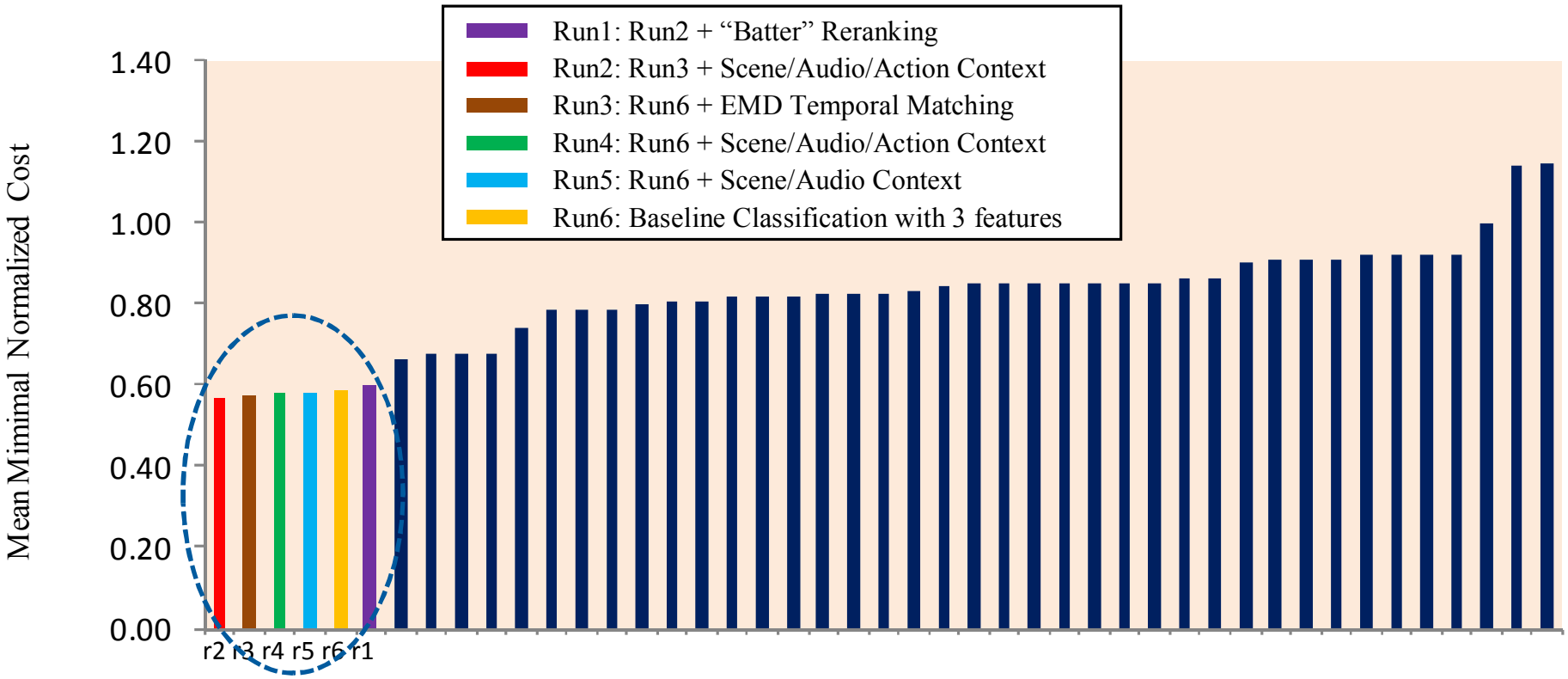
Assembling a shelter

Batting a run in

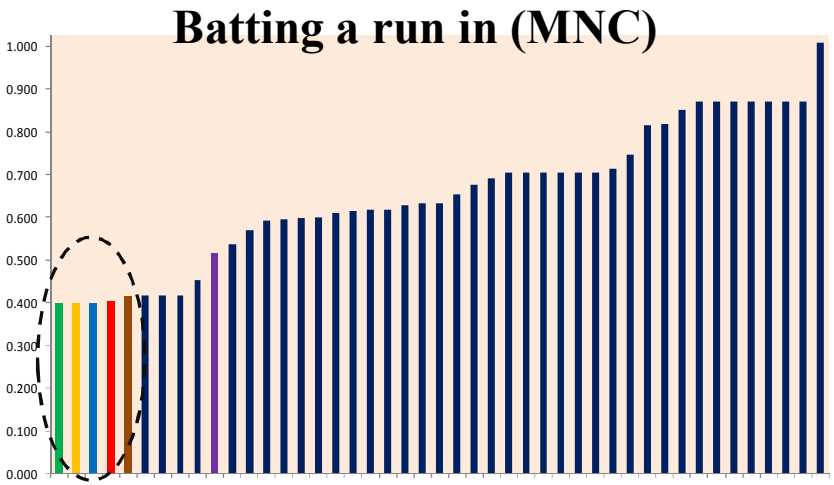
Overview: 4 major components & 6 runs

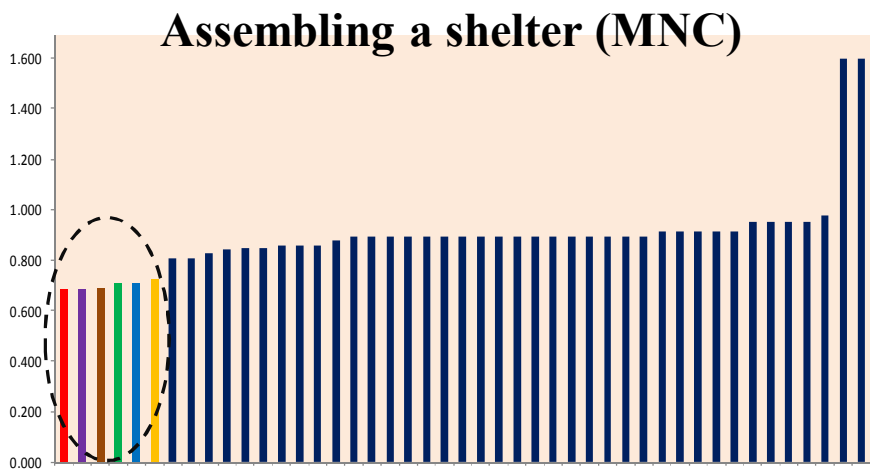


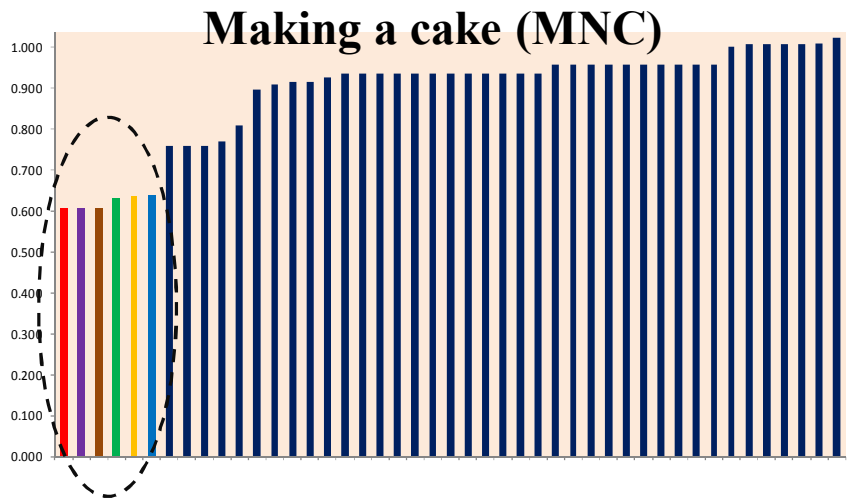
Overview: overall performance



Overview: per-event performance

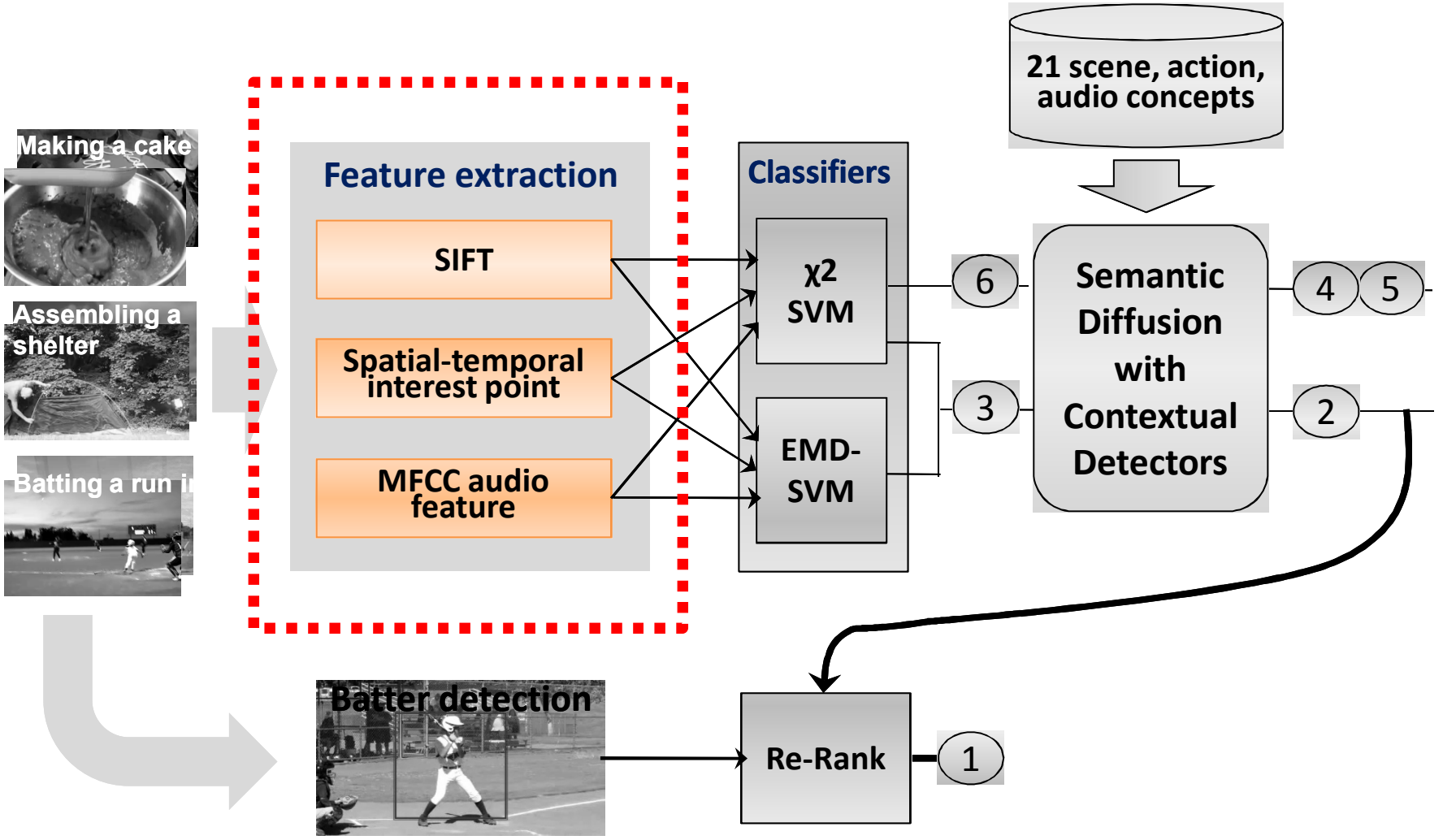






- Run1: Run2 + "Batter" Reranking
- Run2: Run3 + Scene/Audio/Action Context
- Run3: Run6 + EMD Temporal Matching
- Run4: Run6 + Scene/Audio/Action Context
- Run5: Run6 + Scene/Audio Context
- Run6: Baseline Classification with 3 features

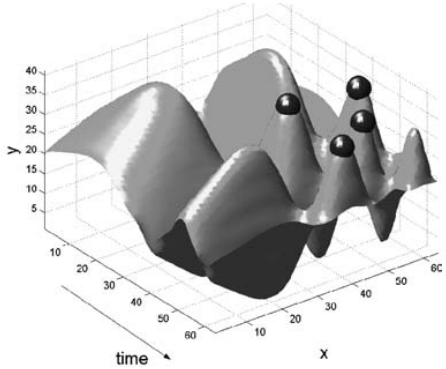
Roadmap > multiple modalities



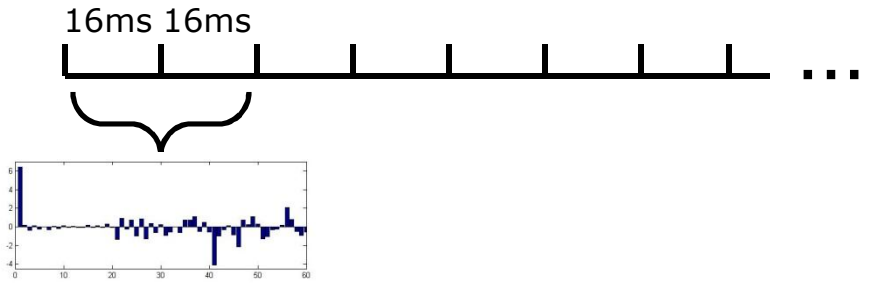
Three Feature Modalities…

- *SIFT (visual)*
  - *D. Lowe, IJCV 04.*

- *STIP (visual)*
  - *I. Laptev, IJCV 05.*

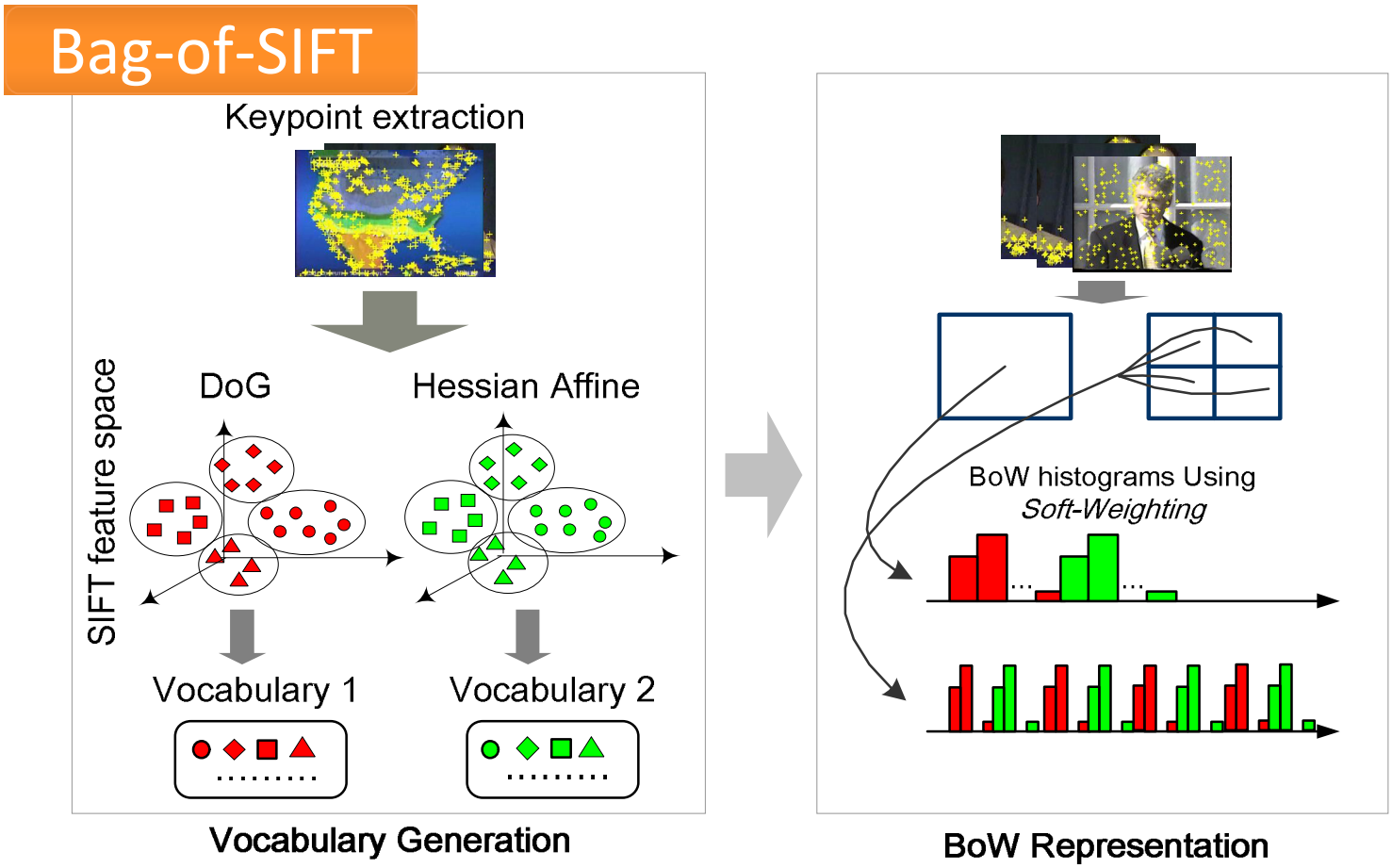


- *MFCC (audio)*



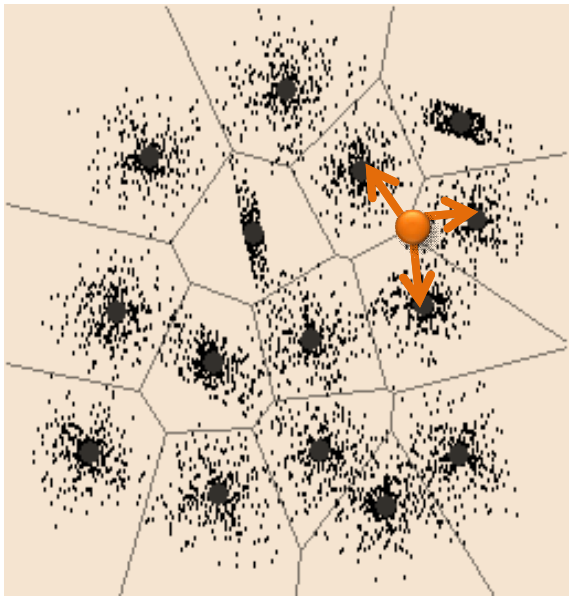
Bag-of-X Representation

- **X = SIFT or STIP or MFCC**
- **Soft weighting** (Jiang, Ngo and Yang, ACM CIVR 2007)



Soft-weighting in Bag-of-X

- Soft weighting is used for all the three Bag-of-X representations



-- Assign a feature to multiple visual words

-- weights are determined by feature-to-word similarity

Details in: Jiang, Ngo and Yang, ACM CIVR 2007.

Image source: http://www.cs.joensuu.fi/pages/franti/vq/lkm15.gif

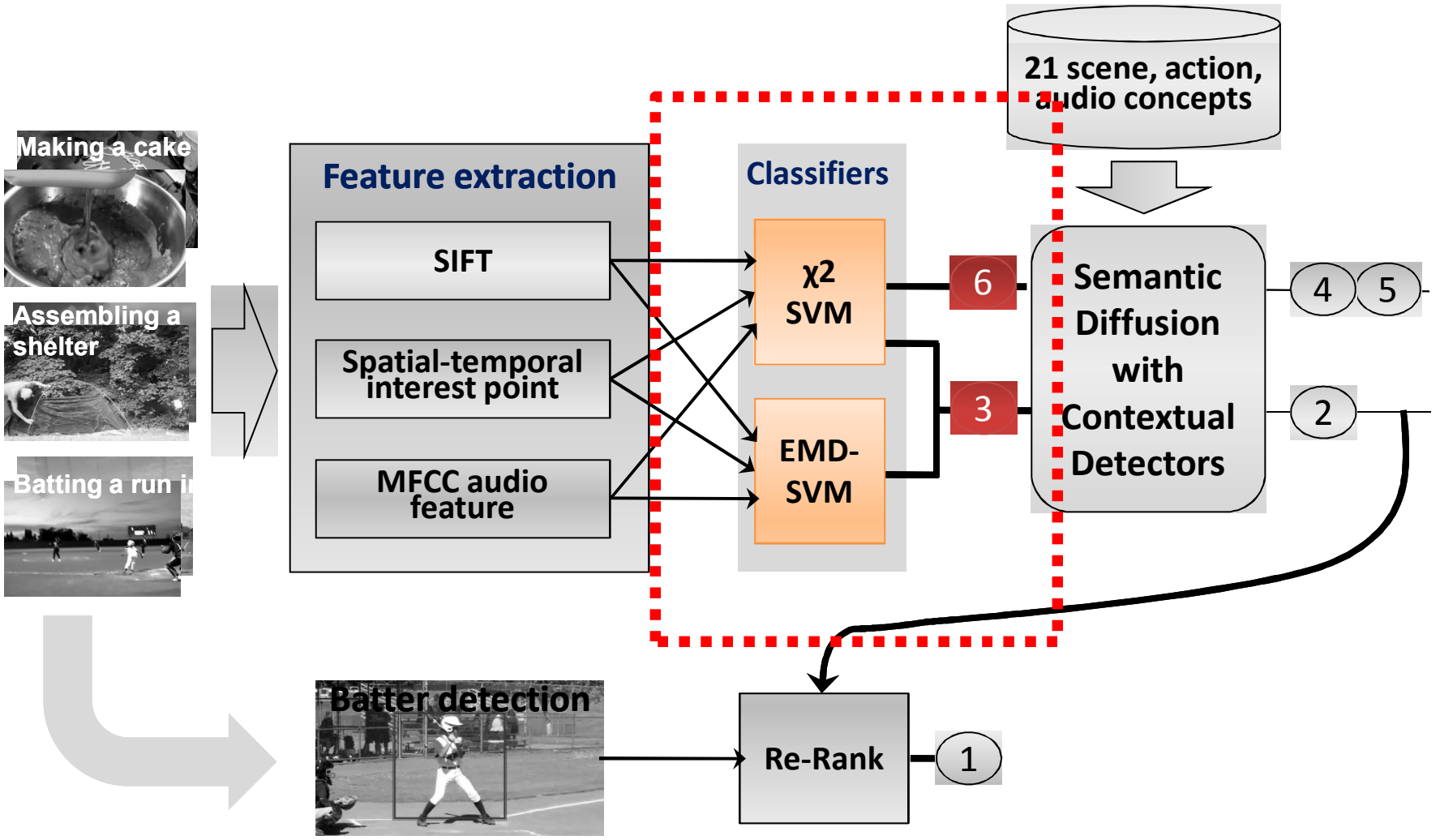
Results on Dry-run Validation Set

- Measured by Average Precision (AP)

	Assembling a shelter	Batting a run in	Making a cake	*Mean AP*
Visual STIP	0.468	0.719	0.476	0.554
Visual SIFT	0.353	0.787	0.396	0.512
Audio MFCC	0.249	0.692	0.270	0.404
STIP+SIFT	0.508	0.796	0.476	0.593
STIP+SIFT+MFCC	**0.533**	**0.873**	**0.493**	**0.633**

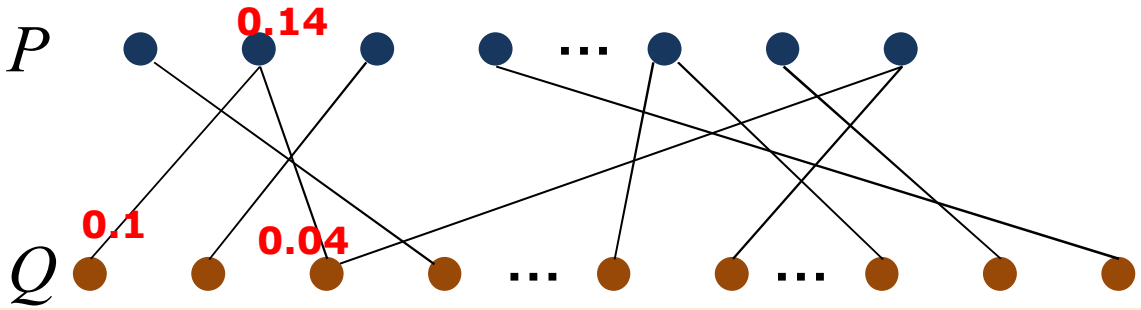
- STIP works best for event detection
- The 3 features are highly complementary!
  - Should be jointly used for multimedia event detection

Roadmap > temporal matching



Temporal Matching With EMD Kernel

- Earth Mover's Distance (EMD)



Given two frame sets $P = \{(p_1, w_{p1}), \ldots, (p_m, w_{pm})\}$ and $Q = \{(q_1, w_{q1}), \ldots, (q_n, w_{qn})\}$, the EMD is computed as

$$\text{EMD}(P, Q) = \Sigma_i \Sigma_j f_{ij} d_{ij} / \Sigma_i \Sigma_j f_{ij}$$

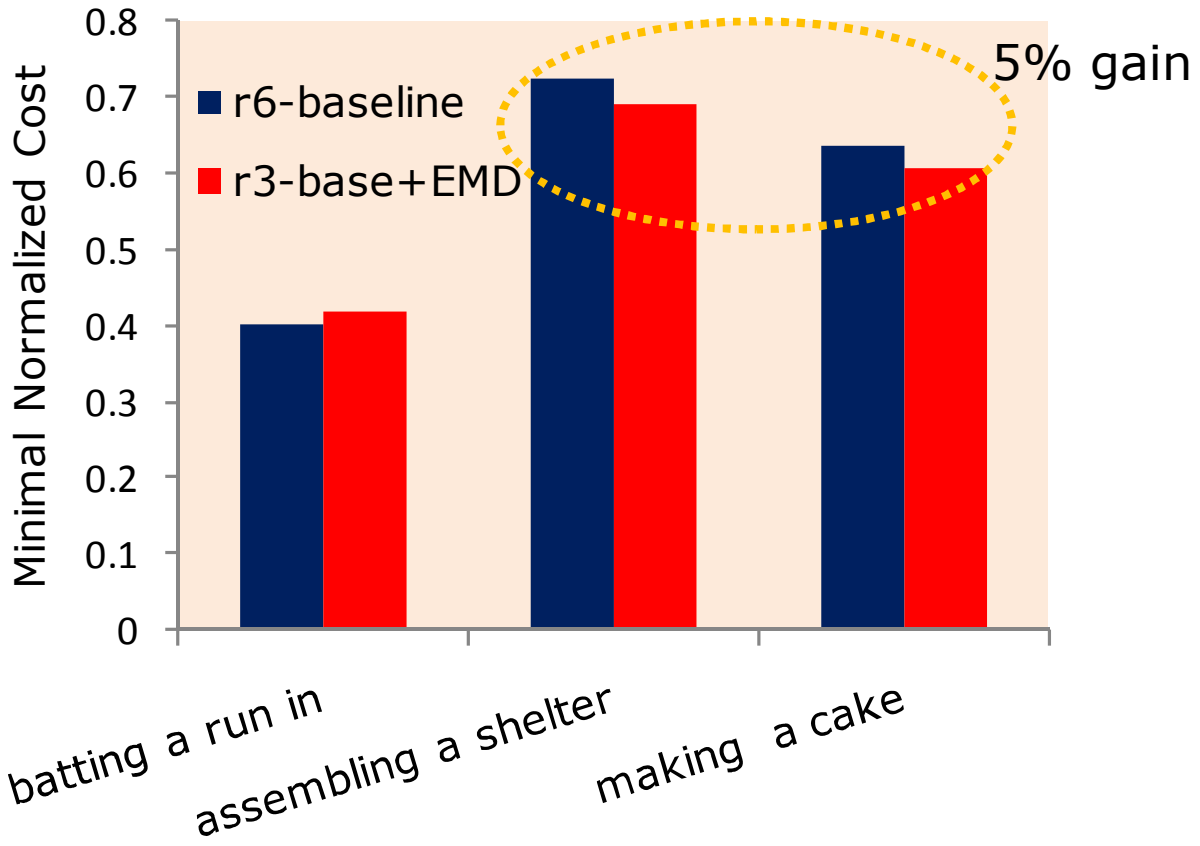
d_{ij} is the χ^2 visual feature distance of frames p_i and q_j. f_{ij} (weight transferred from p_i and q_j) is optimized by minimizing the overall transportation workload $\Sigma_i \Sigma_j f_{ij} d_{ij}$

- EMD Kernel: $K(P,Q) = \exp^{-\rho \text{EMD}(P,Q)}$

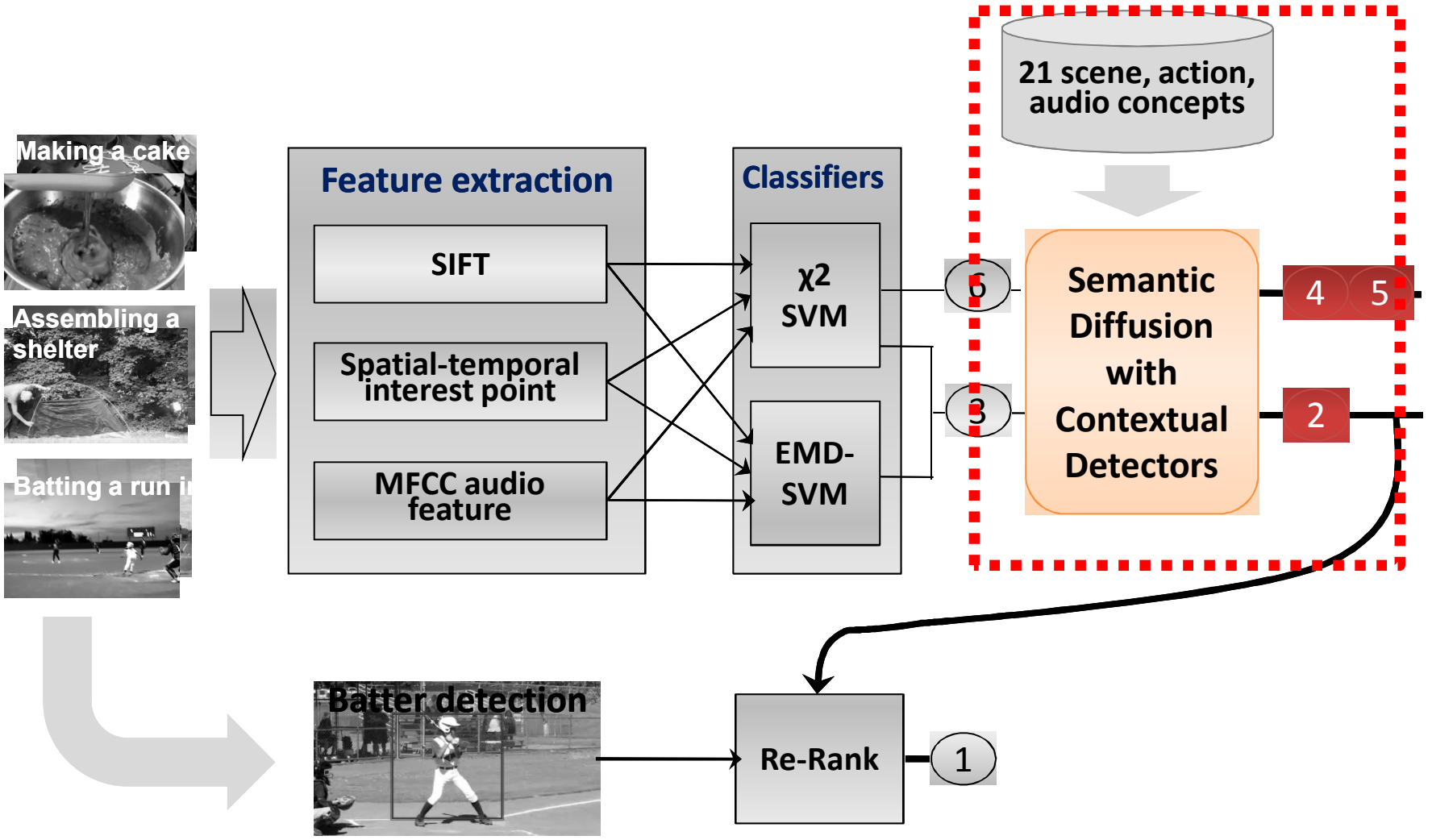
Y. Rubner, C. Tomasi, L. J. Guibas, "A metric for distributions with applications to image databases", ICCV, 1998.
D. Xu, S.-F. Chang, "Video event recognition using kernel methods with multi-level temporal alignment", PAMI, 2008.

Temporal Matching Results

- EMD is helpful for two events
  - results measured by minimal normalized cost (lower is better)

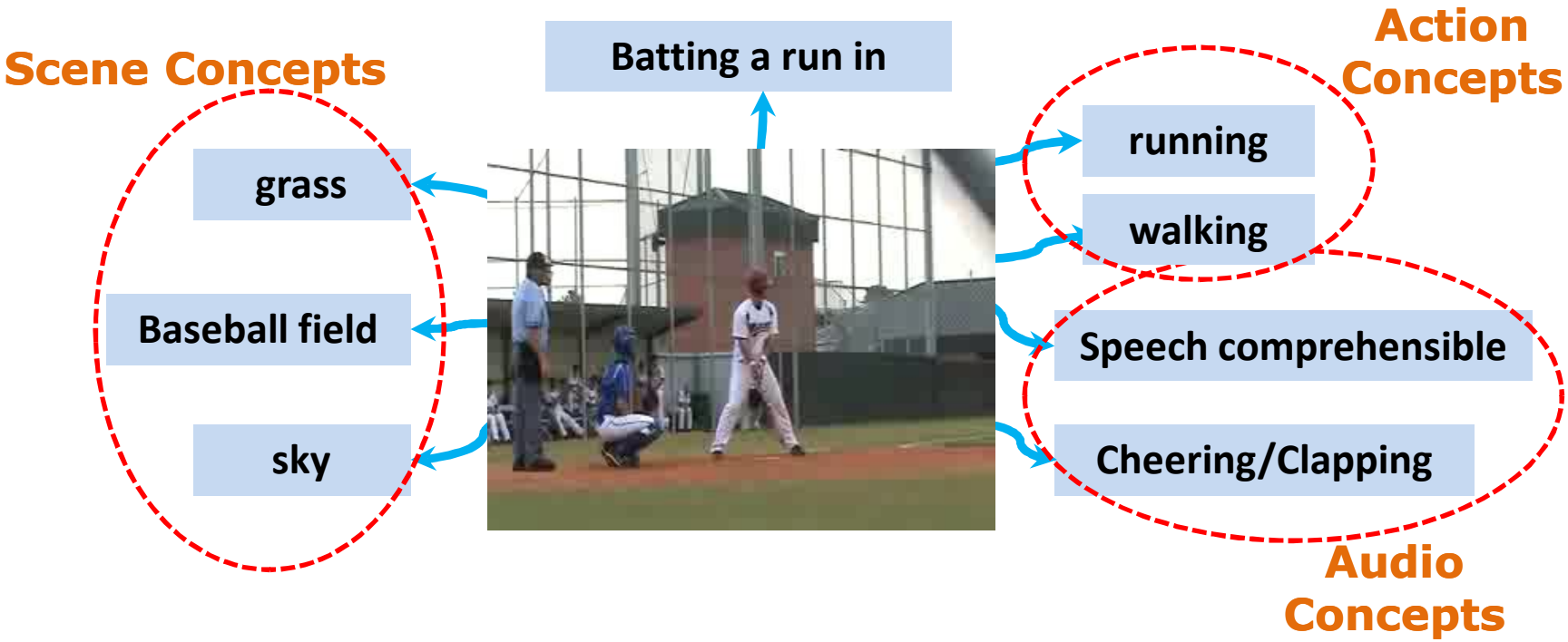


Roadmap > contextual diffusion



Event Context

- Events generally occur under particular scene settings with certain audio sounds!
  - Understanding contexts may be helpful for event detection



Contextual Concepts

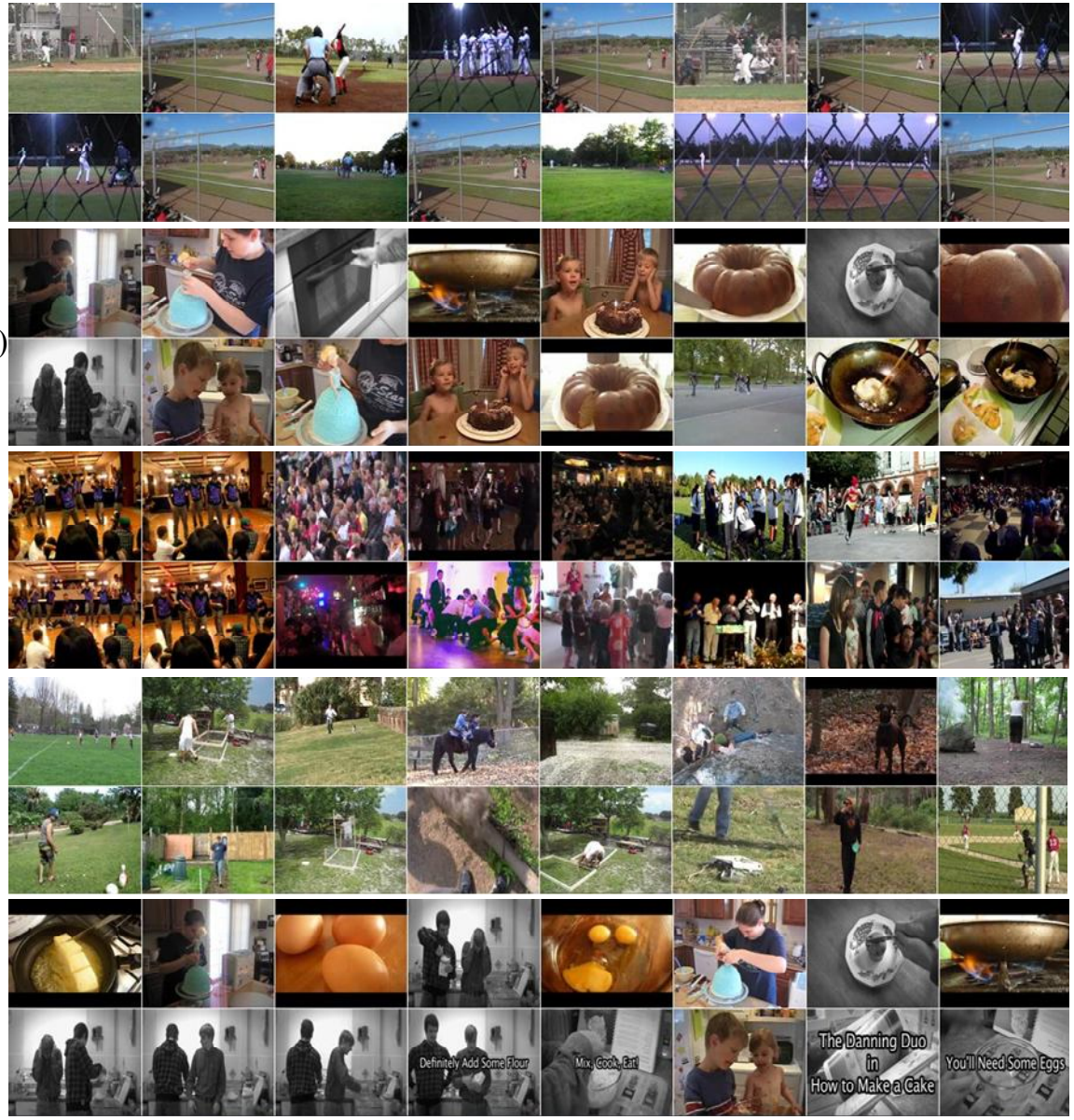
- 21 concepts are defined and annotated over MED development set.

Human Action Concepts	Scene Concepts	Audio Concepts
▪ Person walking ▪ Person running ▪ Person squatting ▪ Person standing up ▪ Person making/assembling stuffs with hands (hands visible) ▪ Person batting baseball	▪ Indoor kitchen ▪ Outdoor with grass/trees visible ▪ Baseball field ▪ Crowd (a group of 3+ people) ▪ Cakes (close-up view)	▪ Outdoor rural ▪ Outdoor urban ▪ Indoor quiet ▪ Indoor noisy ▪ Original audio ▪ Dubbed audio ▪ Speech comprehensible ▪ Music ▪ Cheering ▪ Clapping

- SVM classifier for concept detection
  - **STIP for action concepts, SIFT for scene concepts, and MFCC for audio concepts**

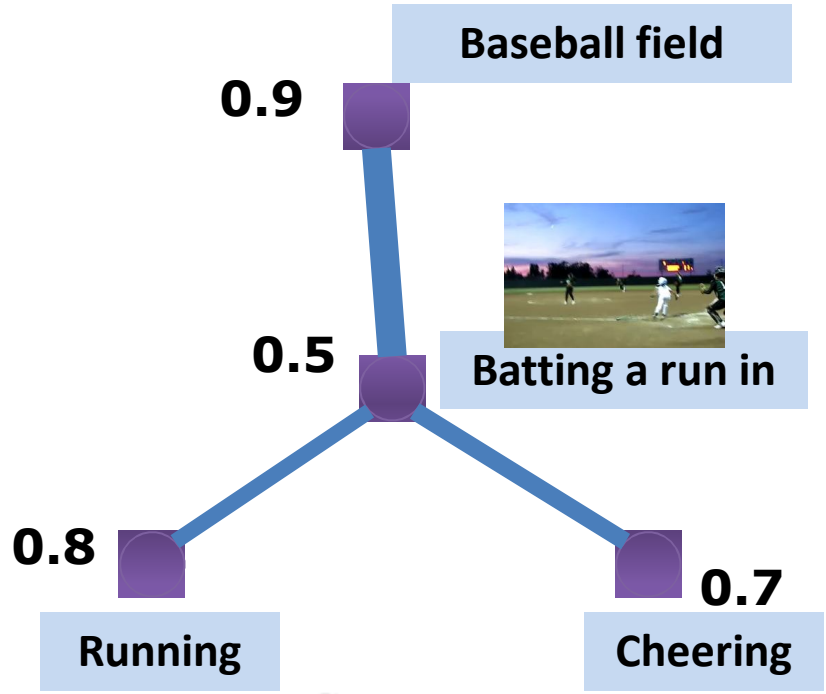
Jingen Liu, Jiebo Luo & Mubarak Shah, Recognizing Realistic Actions from Videos "in the Wild", CVPR 2009
Shih-Fu Chang et al. Columbia University/VIREO-CityU/IRIT TRECVID2008 High-Level Feature Extraction and Interactive Video Search. TRECVID Workshop, 2008

Concept Detection: example result



Contextual Diffusion Model

- Semantic Diffusion
  [Jiang, Wang, Chang & Ngo, ICCV 2009]
  - Semantic graph
    - Nodes are concepts/events
    - Edges represent concept/event correlation
  - Graph diffusion
    - Smooth detection scores w.r.t. the correlation

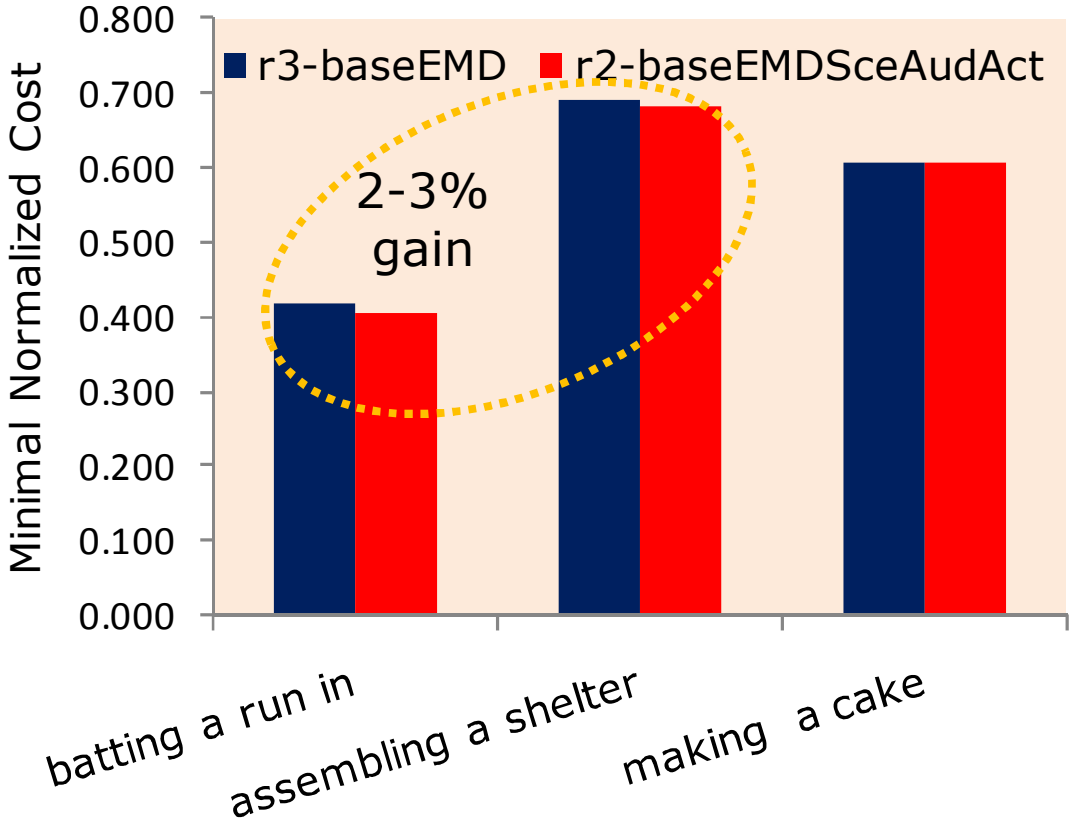


$$\mathcal{E}(\hat{g}) = \frac{1}{2}\sum_{i=1}^{C}\sum_{j=1}^{C} W_{ij} \| \hat{g}(c_i) - \hat{g}(c_j) \|^2 + \mu \sum_{i=1}^{C} \| \hat{g}(c_i) - g(c_i) \|^2$$

Project page and source code:
http://www.ee.columbia.edu/ln/dvmm/researchProjects/MultimediaIndexing/DASD/dasd.htm

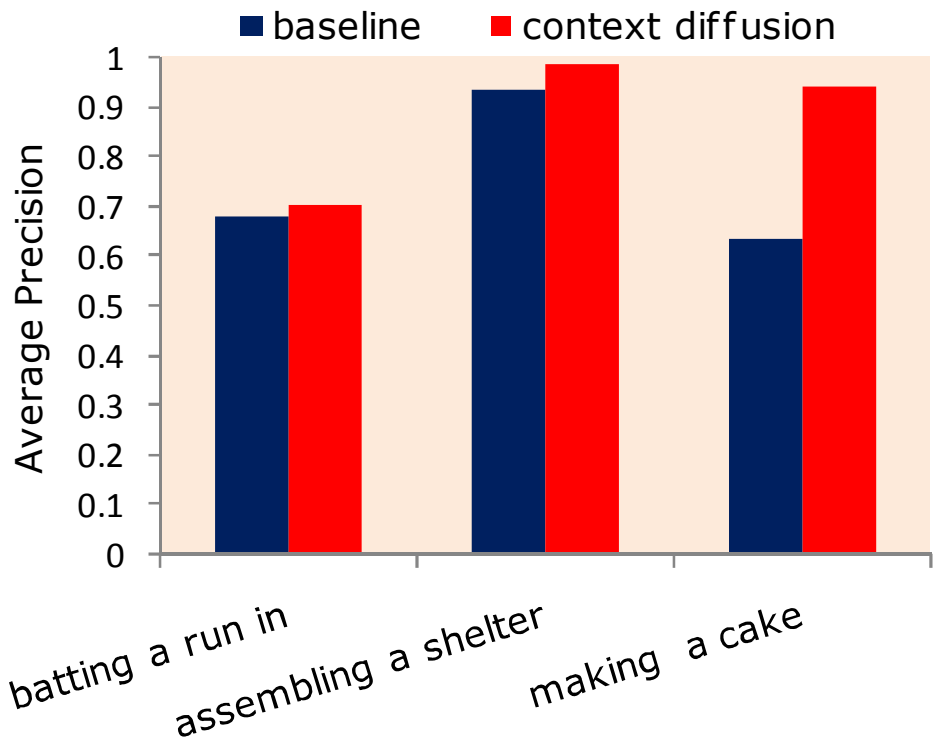
Contextual Diffusion Results

- Context is *slightly* helpful for two events
  - results measured by minimal normalized cost (lower is better)

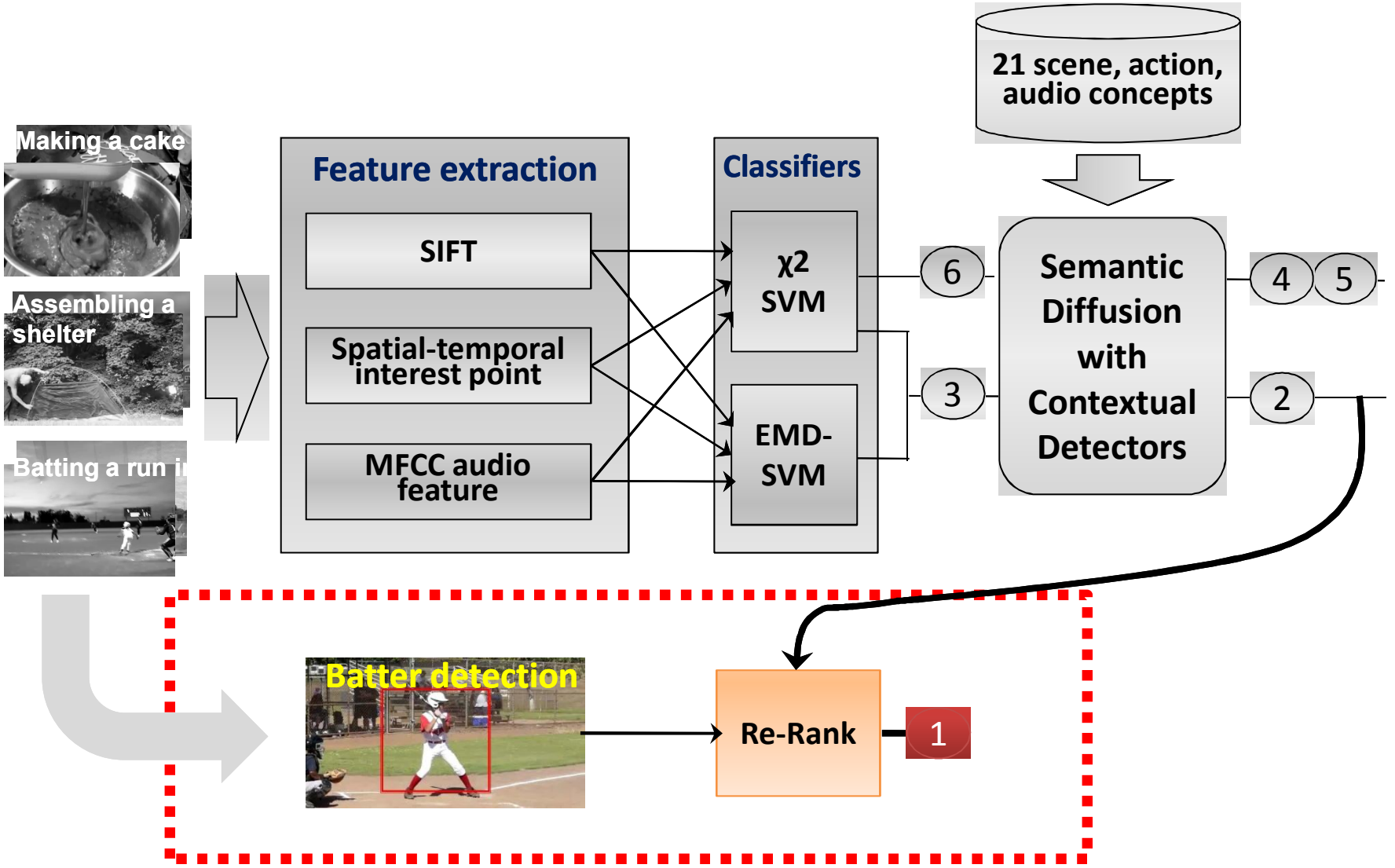


Contextual Diffusion Results

- … but the improvement is much higher when context is perfect (on a validation set)
  - results measured by average precision (higher is better)



Roadmap > reranking with event-specific object detector

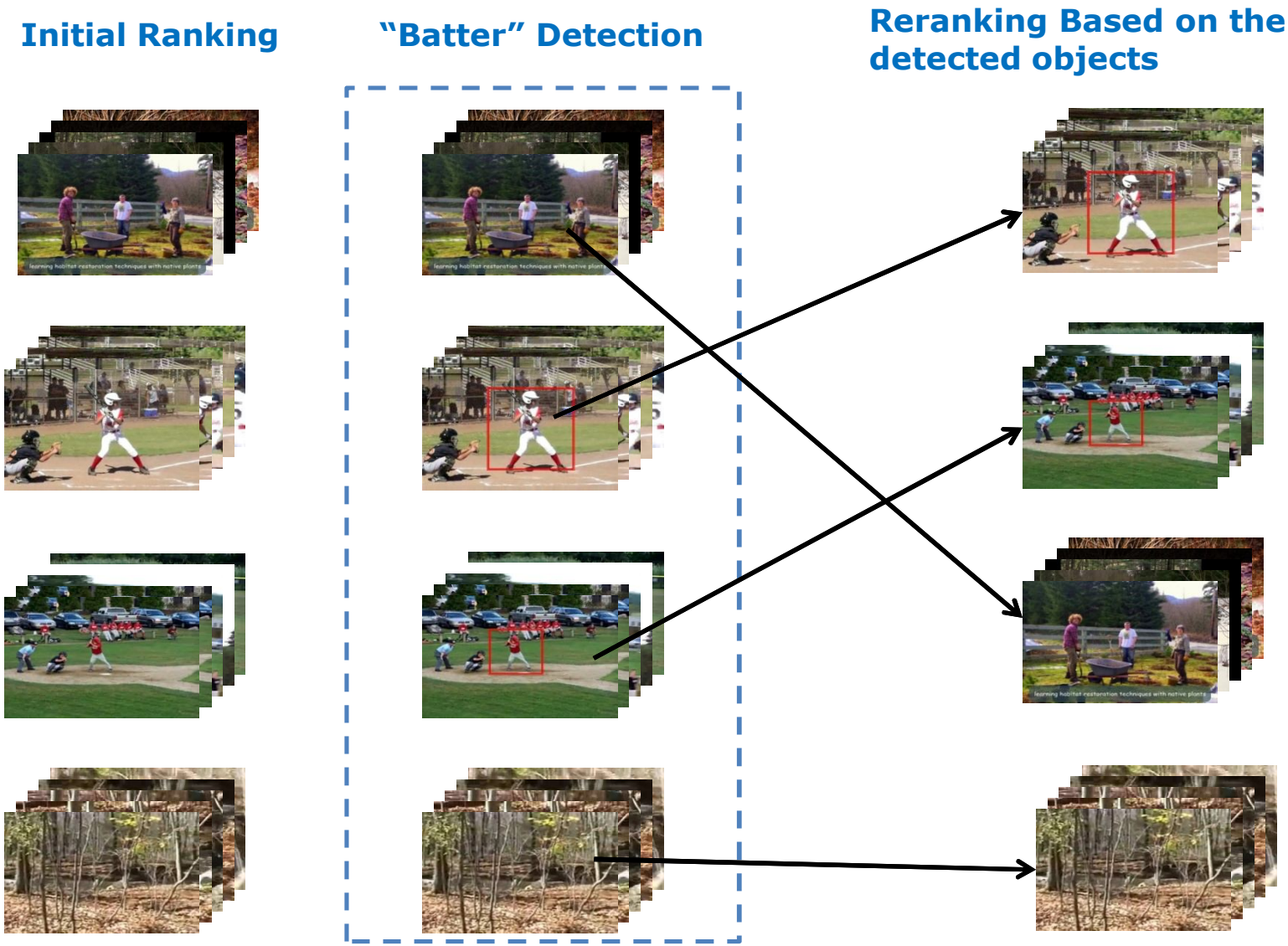


Reranking with Event-Specific Object Detector

- "Batter" detector is trained by AdaBoost framework

Reranking with Event-Specific Object Detector

- "Batter" detector is trained by AdaBoost framework



Lessons learned

1. STIP is powerful for event detection.
2. Combining multiple audio-visual features is very effective!
3. Temporal Matching with EMD is useful for some events
4. Diffusion with Contextual Concepts is promising, and deserves deeper research

Future Work

1. Explore deep joint audio-visual representation, e.g., Audio-Visual Atoms [Jiang et al, ACMMM09]
2. Another interesting research direction is to investigate an adaptive method to find the best components for each event

COLUMBIA UNIVERSITY

UCF

THANK YOU!

More information at:

http://www.ee.columbia.edu/dvmm/