Real-World Applications of Activity Recognition

Sangmin Oh
Kitware

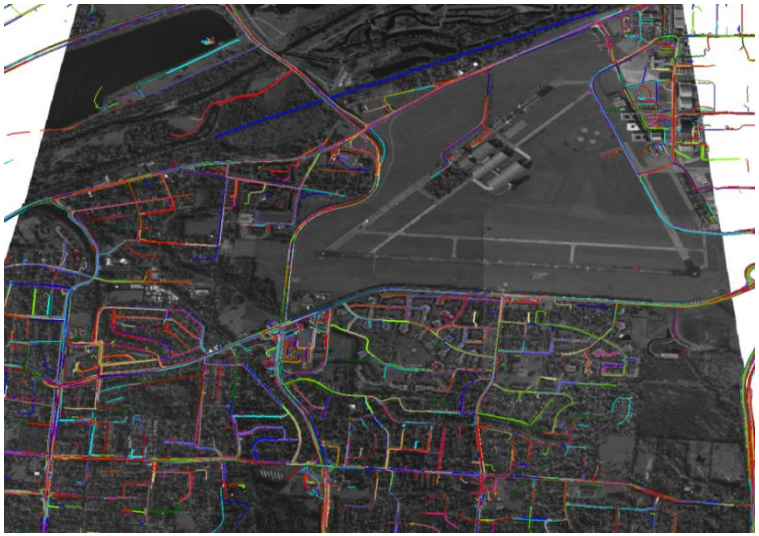
CVPR tutorial on 2014/06/23

Emerging Applications

Unconstrained Video Search

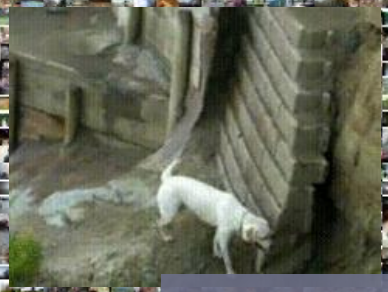
Sports Video Analysis

Aerial Video Analysis



Unconstrained Video Search

Lots of unconstrained video…
…find me activities I want



Task: Retrieve clips with activities of interest (e.g. “flash mob” or “birthday”)

Challenges:

Content variation across archive is huge

Content variation within activity is large

Metadata variations (frame size, clip length, bitrates, …)

Archive size is large (150K+ clips)

Interaction

Unconstrained Video Search Datasets

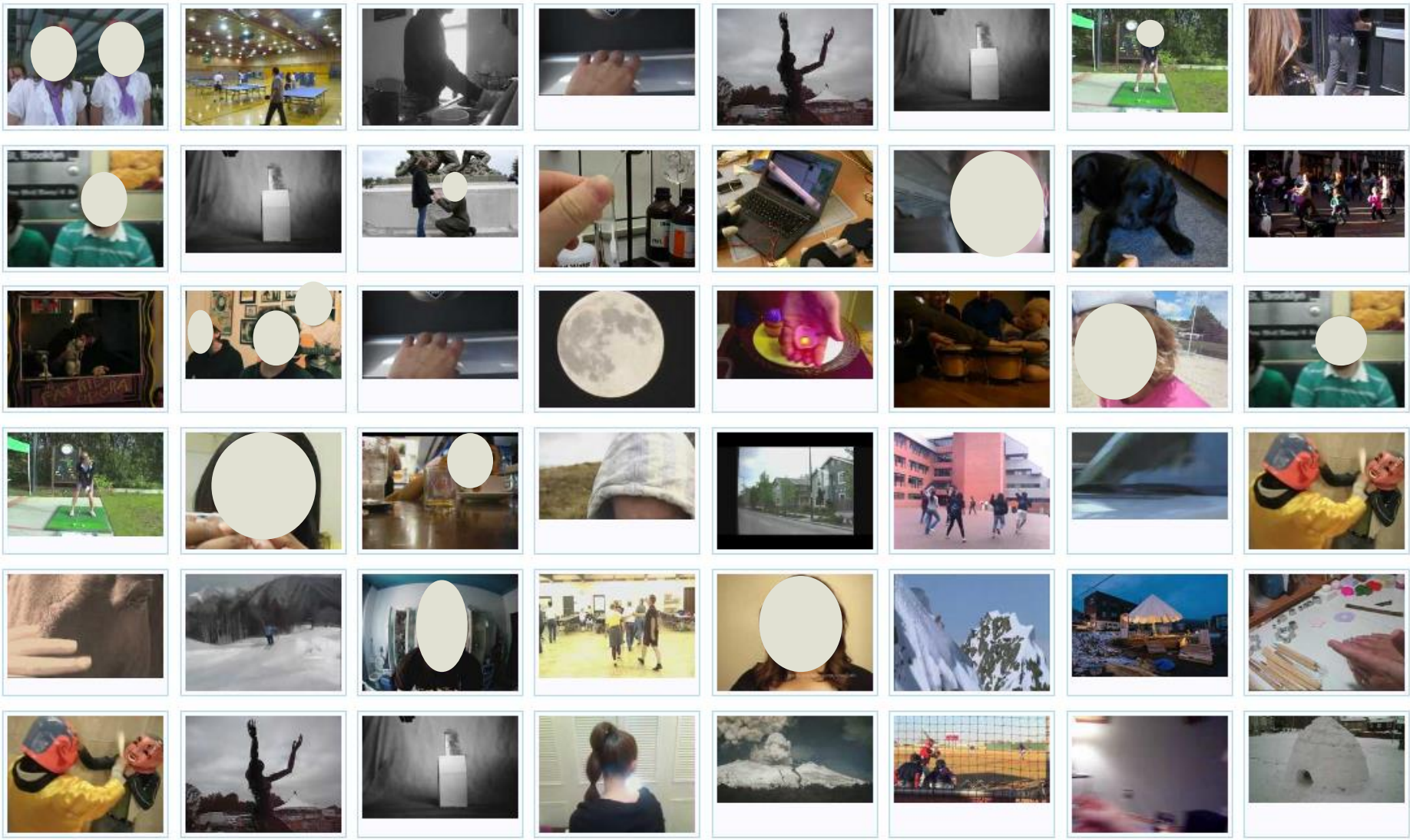
TRECVID Multimedia Event Detection (MED) Dataset

- Evaluation data: Very large collection of web videos and detection of known event types.
- Available from a webpage (pending TRECVID participation): trecvid.nist.gov
- Complex events
  - 25 Test events (as of 2012, and increasing):
    - Wedding, changing a tire, woodworking project, parkour, townhall meeting, marriage proposal etc.
  - Full clips: Includes stitching, severe camera motion, temporal and spatial clutter, e.g., 1 Hour long.

Columbia Consumer Video (CCV) dataset

- Total 9317 videos (210 hours in total)
- Average length: 80 secs
- Complex events
  - 20 events
    - Wedding ceremony, wedding reception, biking, graduation, baseball, birthday, bird, playground etc
- ***Consumer Video Understanding: A Benchmark Database and an Evaluation of Human and Machine Performance***, by Jiang, Ye, Chang, Ellis, Loui, in ICMR 2011

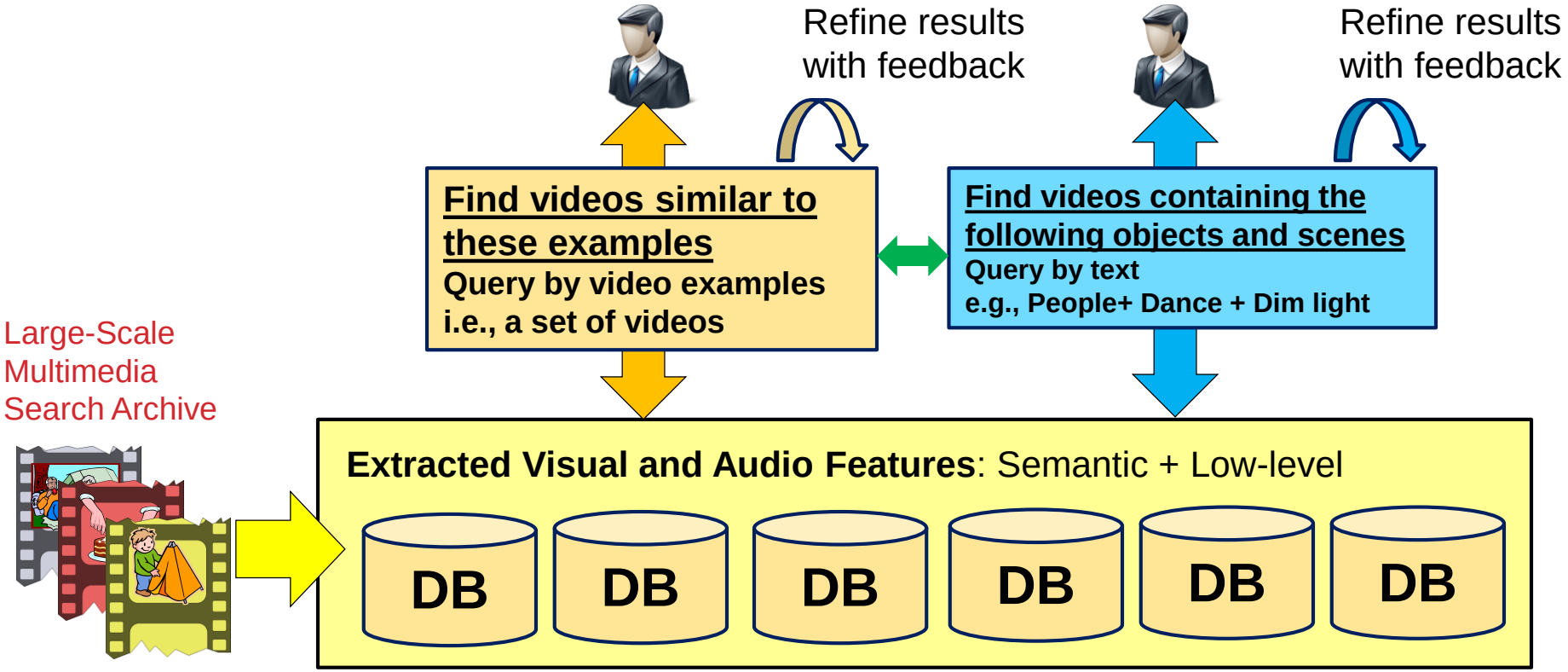
How does Random Look?



Random images from typical unconstrained videos

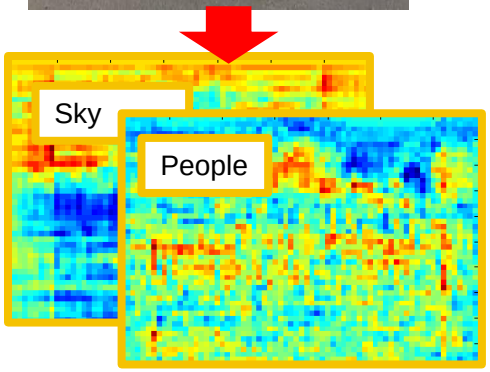
Flash Mob

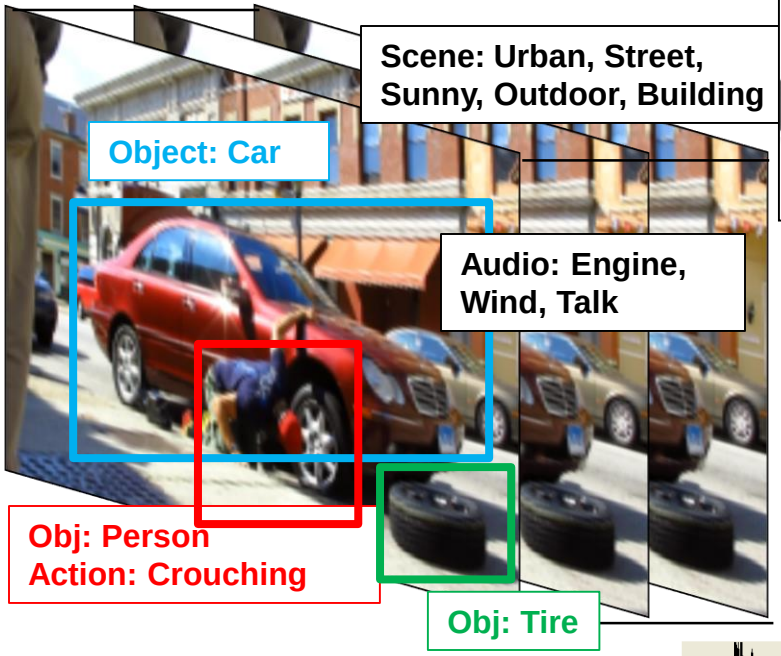
Multiple Search Modes



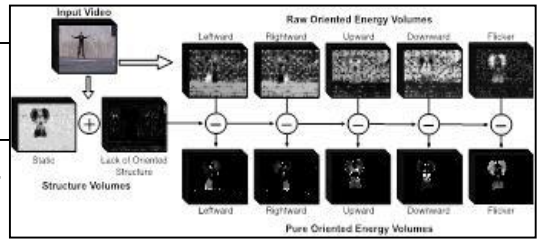
Examples of Video Features: Visual & Audio

Objects





Actions



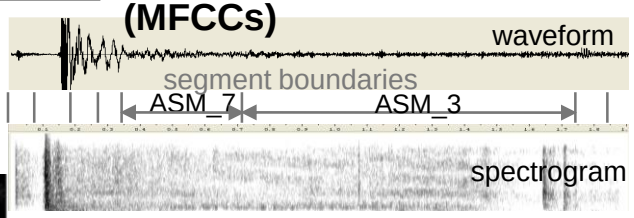
Videography

Pan / Tile / Zoom

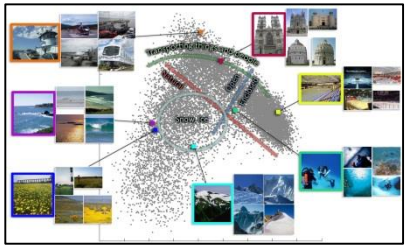
Size of people

Correlation
of camera and FG motion

Low-level Audio Signatures (MFCCs)



Scene Attributes



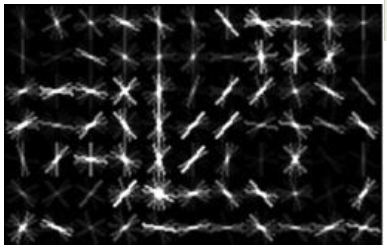
Low-level Visual Features

Indoor/outdoor lighting

Emotion Functions ….

Materials Viewpoint

Histogram of Oriented Gradients, Texture



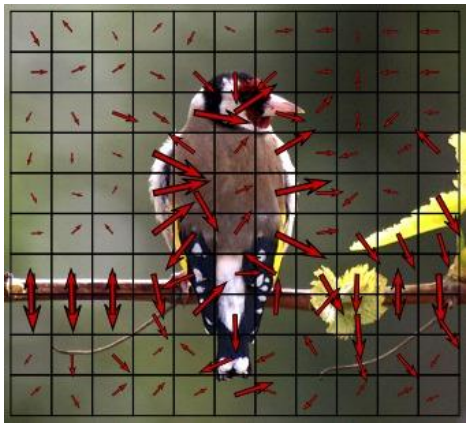
Audio Events

Engine Human Chat Outdoor

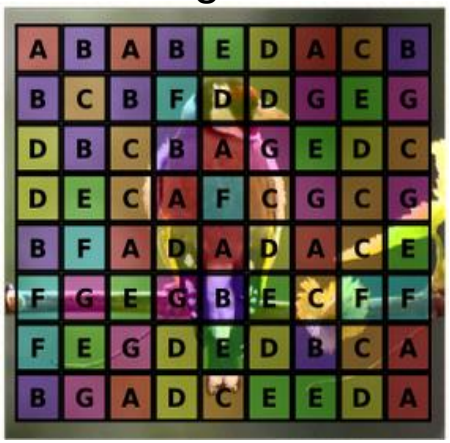
Explosion Animal Water ….

Low-level Feature & Encoding

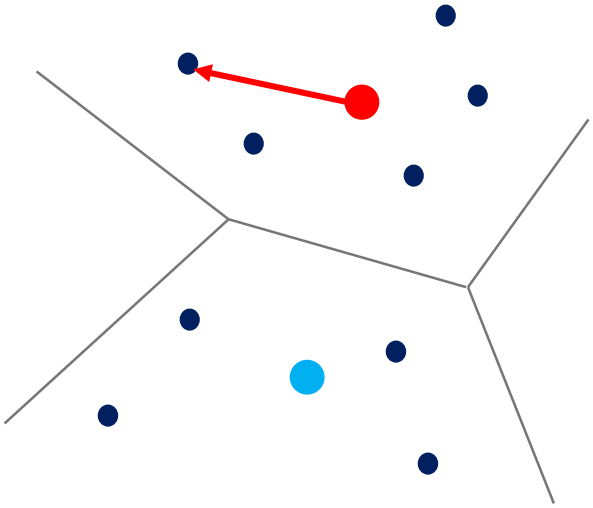
Local feature extraction



Quantization using
Clustering Codebook



BoW Histogram



Concatenate

|sum of diffs to c(1)|…. |sum of diffs to c(n)|

Normalize

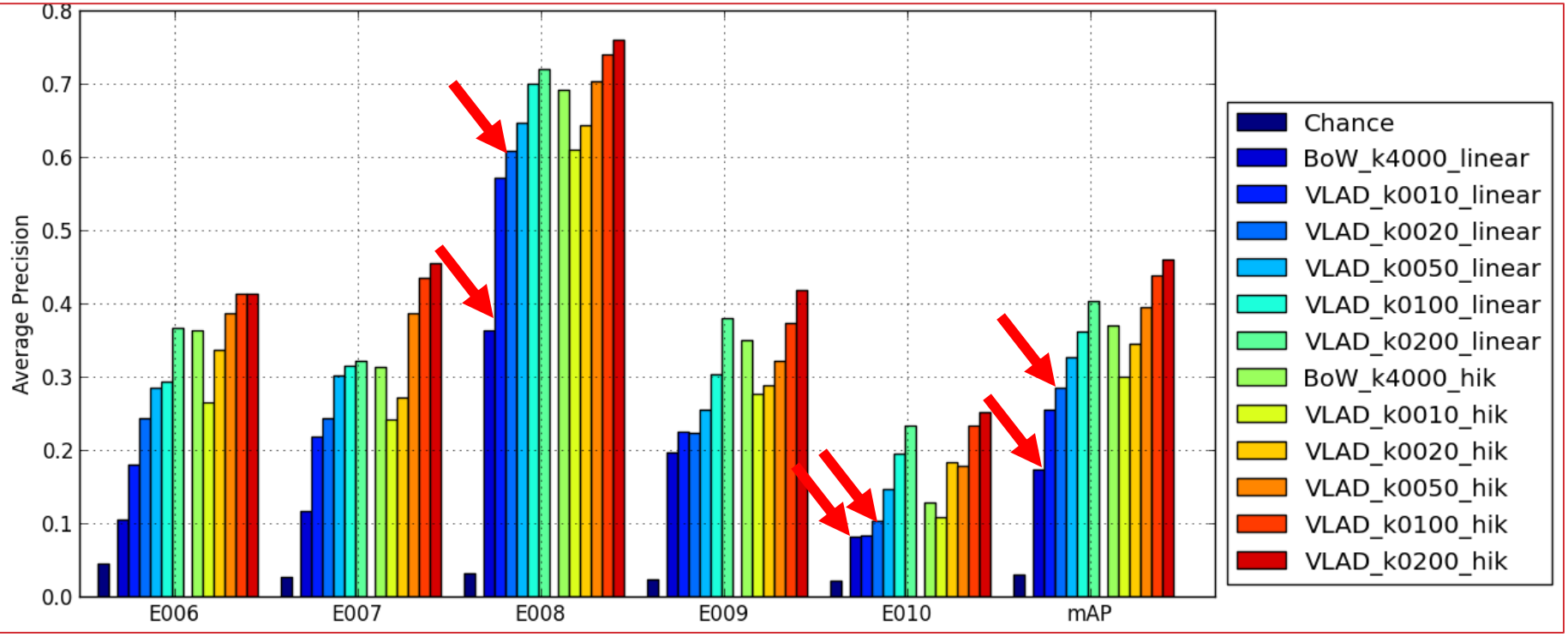
Difference Coding Vector

Dimension
= K*D

Aggregating local descriptors into a compact image representation Jegou, Douze, Schimid, Perez, CVPR 2010.
Fisher Vectors for Fine-Grained Visual Categorization Perronnin, Sanchez, Akata, CVPR 2011
Large-scale Web Video Event Classification by use of Fisher Vectors, Sun, Navatia in WACV 2013

VLAD vs BoW

Color SIFT



- Difference coding method can achieve higher accuracy with lower computational demand. Most expensive step is quantization, and difference coding may require less number of quantizations are required due to reduced cluster centers.
- Cost is potentially larger memory footprint.

Activities and Objects

Average Object detector responses on Wedding Videos

(TRECVID MED dataset)

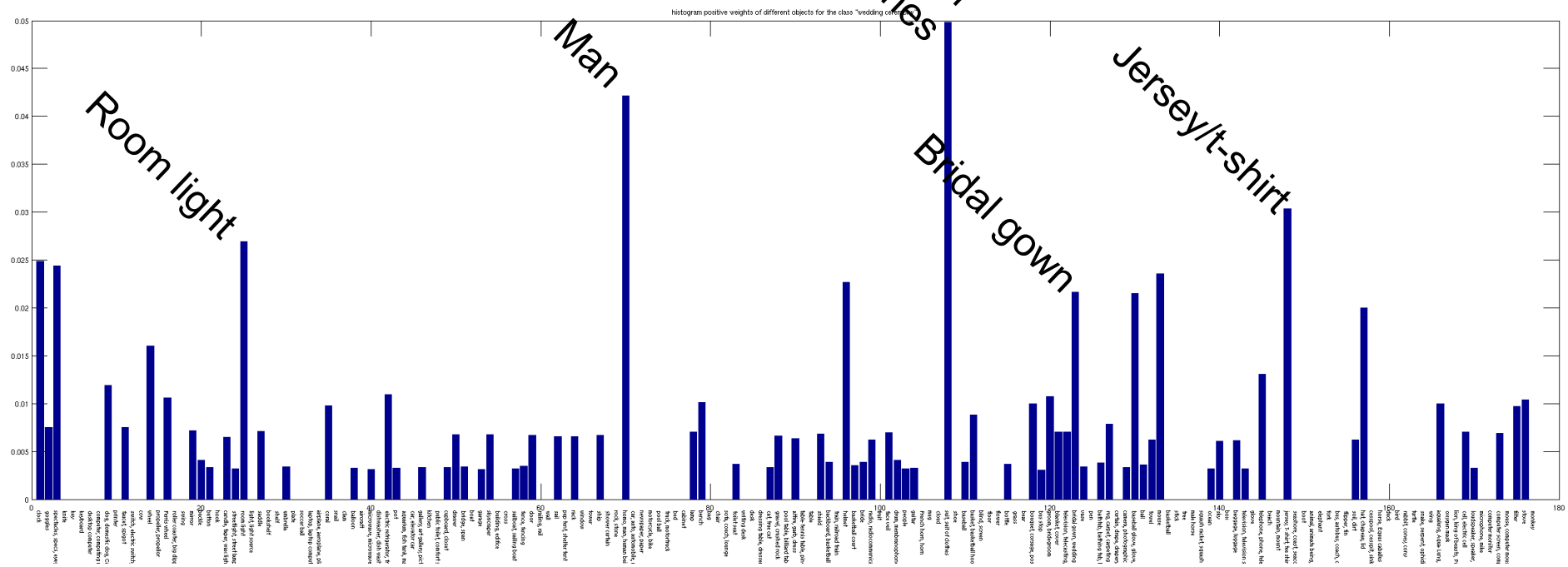


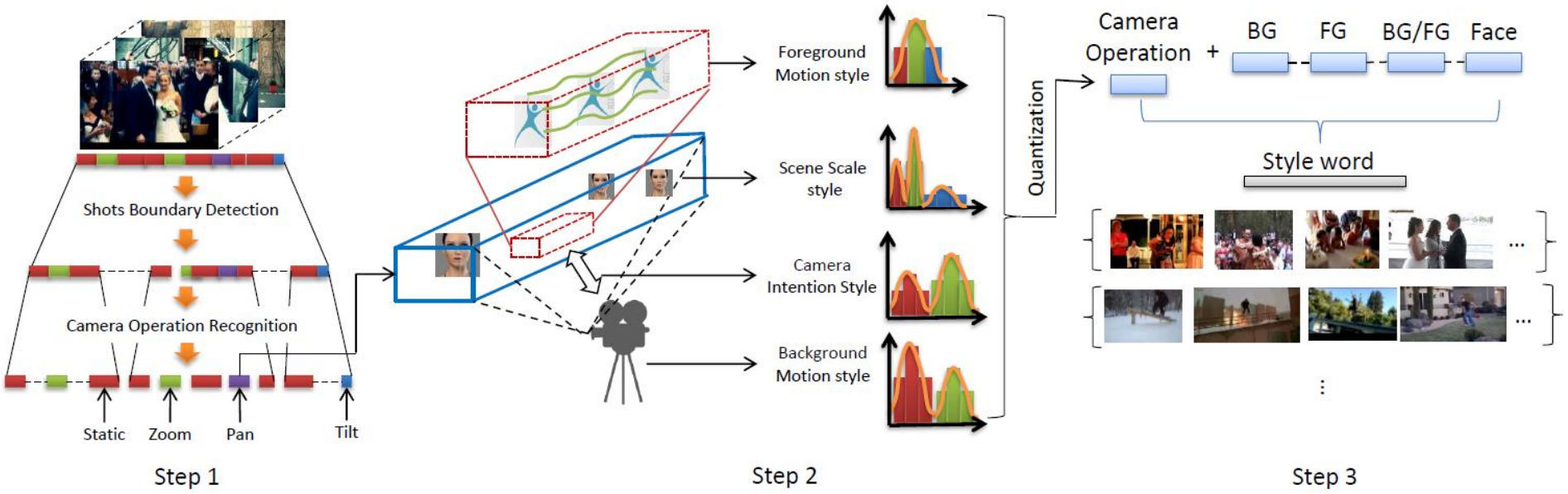
Image Courtesy of Greg Mori's group at Simon Fraser Univ.

Object Bank: A High-Level Image Representation for Scene Classification and Semantic Feature Sparsification Li, Su, Xing, Fei-Fei, NIPS 2010

Videography Style Analysis

Combine a set of camera motion and related features into a “videography style descriptor”

Idea is for the style descriptor to capture some semantically meaningful things about how the video was taken



A Videography Analysis Framework for Video Retrieval and Summarization Oh, Li, Perera, Fu, BMVC 2012

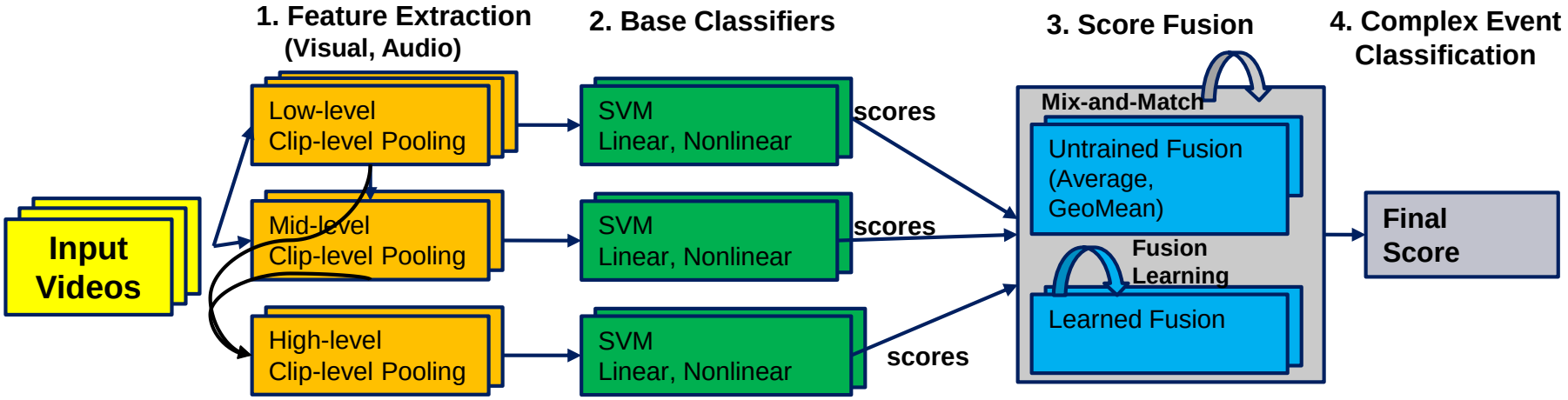
Videography Styles

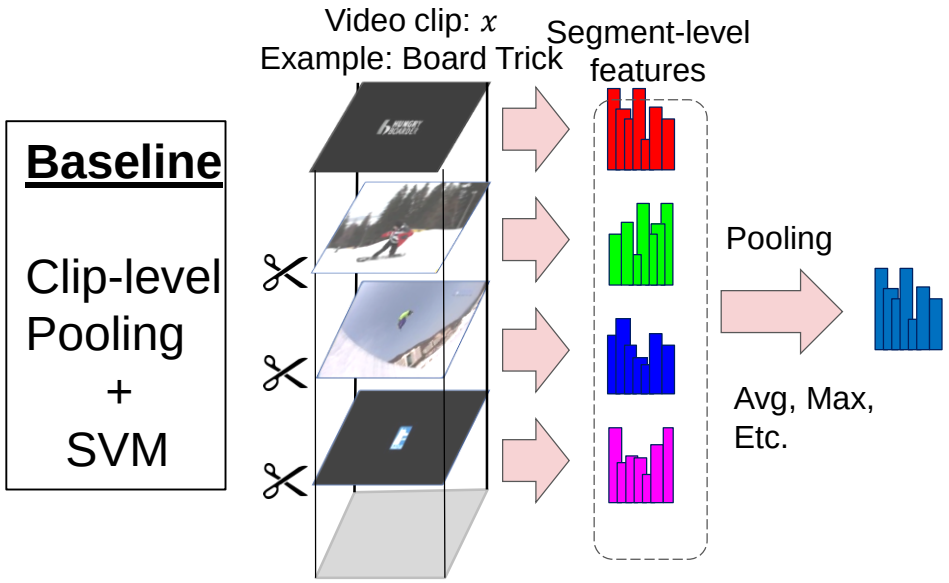
Captures
- Background Motion
- Foreground Motion
- Correlation BG/FG
- Scale

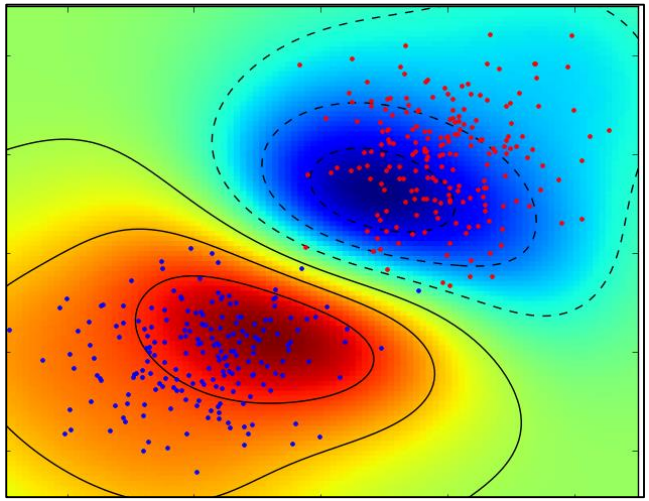
- **Red**: background
- **Green**: foreground
- **White arrows**: Camera

Example on Parkour video

Classifier Baseline Architecture







Support Vector Machine (SVMs)

Multimedia event detection with multimodal feature fusion and temporal concept localization Oh, McCloskey, Kim, at al. Machine Vision and Applications 25(1), 2014.

Multimodal feature fusion for robust event detection in web videos Natarjan et al. CVPR 2012.

Single Feature and Fusion Results

Different Events, (see TRECVID MED dataset for details)

Feature & Classifier Combinations

Base Classifier	E006	E007	E008	E009	E010	E011	E012	E013	E014	E015
GIST HIK	0.667	0.685	0.326	0.558	0.736	0.593	0.524	0.544	0.551	0.691
HoG3D Linear	0.645	0.775	0.341	0.484	0.724	0.700	0.452	0.519	0.590	0.728
HoG3D HIK	0.457	0.550	0.273	0.463	0.586	0.500	0.448	**0.317**	0.359	0.531
HoG3D NGD	0.462	0.468	0.258	0.474	0.529	0.479	**0.352**	0.327	0.321	0.531
ISA HIK	0.516	0.514	0.288	0.453	0.540	0.536	0.424	**0.317**	0.333	**0.506**
CSIFT Linear	0.704	0.649	0.364	0.453	0.747	0.679	0.463	0.683	0.526	0.654
CSIFT HIK	0.468	0.414	**0.197**	0.368	0.540	0.493	0.359	0.413	0.372	**0.506**
CSIFT NGD	0.495	0.405	0.205	0.379	0.540	0.479	0.368	0.423	0.359	0.519
TCH Linear	0.774	0.676	0.447	0.474	0.805	0.793	0.489	0.712	0.564	0.753
TCH HIK	0.532	0.532	0.273	0.389	0.575	0.543	0.398	0.404	0.397	0.593
TCH NGD	0.532	0.477	0.250	0.400	0.621	0.536	0.420	0.404	0.423	0.580
OB AVG, Linear (L0)	0.645	0.550	0.311	0.432	0.644	0.600	0.519	0.625	0.590	0.617
OB MAX, Linear (L0)	0.597	0.541	0.379	0.442	0.632	0.500	0.554	0.558	0.474	0.630
OB, LSVM (L1)	0.570	0.577	0.280	0.568	0.621	0.586	0.442	0.538	0.500	0.605
OB Avg, HIK	0.532	0.505	0.250	0.411	0.575	0.550	0.442	0.375	0.436	0.568
OB Max, HIK	0.516	0.477	0.250	**0.337**	0.529	**0.457**	0.429	0.442	0.385	0.519
SUN09 MKL	0.441	**0.351**	0.205	**0.337**	**0.483**	0.507	0.355	0.337	0.321	**0.506**
MFCCs Linear	0.548	0.782	0.545	0.681	0.814	0.761	0.645	0.709	0.346	0.667
MFCCs HIK	0.446	0.618	0.424	0.564	0.686	0.739	0.584	0.583	0.295	0.654
MFCCs NGD	0.462	0.673	0.409	0.500	0.698	0.696	0.567	0.631	0.372	0.628
ASM186 HIK	**0.422**	0.561	0.470	0.553	0.744	0.618	0.607	0.602	**0.289**	0.553
ASM64 HIK	0.438	0.607	0.523	0.628	0.733	0.669	0.620	0.670	0.303	0.618
Fusion Model	E006	E007	E008	E009	E010	E011	E012	E013	E014	E015
Average	**0.265**	0.318	0.212	**0.234**	0.430	0.426	0.298	0.262	**0.184**	**0.447**
GeoMean	0.292	**0.290**	**0.189**	0.266	0.430	**0.404**	**0.281**	0.252	0.224	0.461
MFoM	0.324	0.299	0.197	0.287	0.430	0.419	0.329	0.262	0.237	0.487
LEF	**0.265**	0.318	0.197	0.245	**0.384**	0.412	0.285	**0.233**	0.211	0.461
Best Base - Best Fusion	0.157	0.061	0.008	0.103	0.099	0.053	0.071	0.084	0.105	0.059

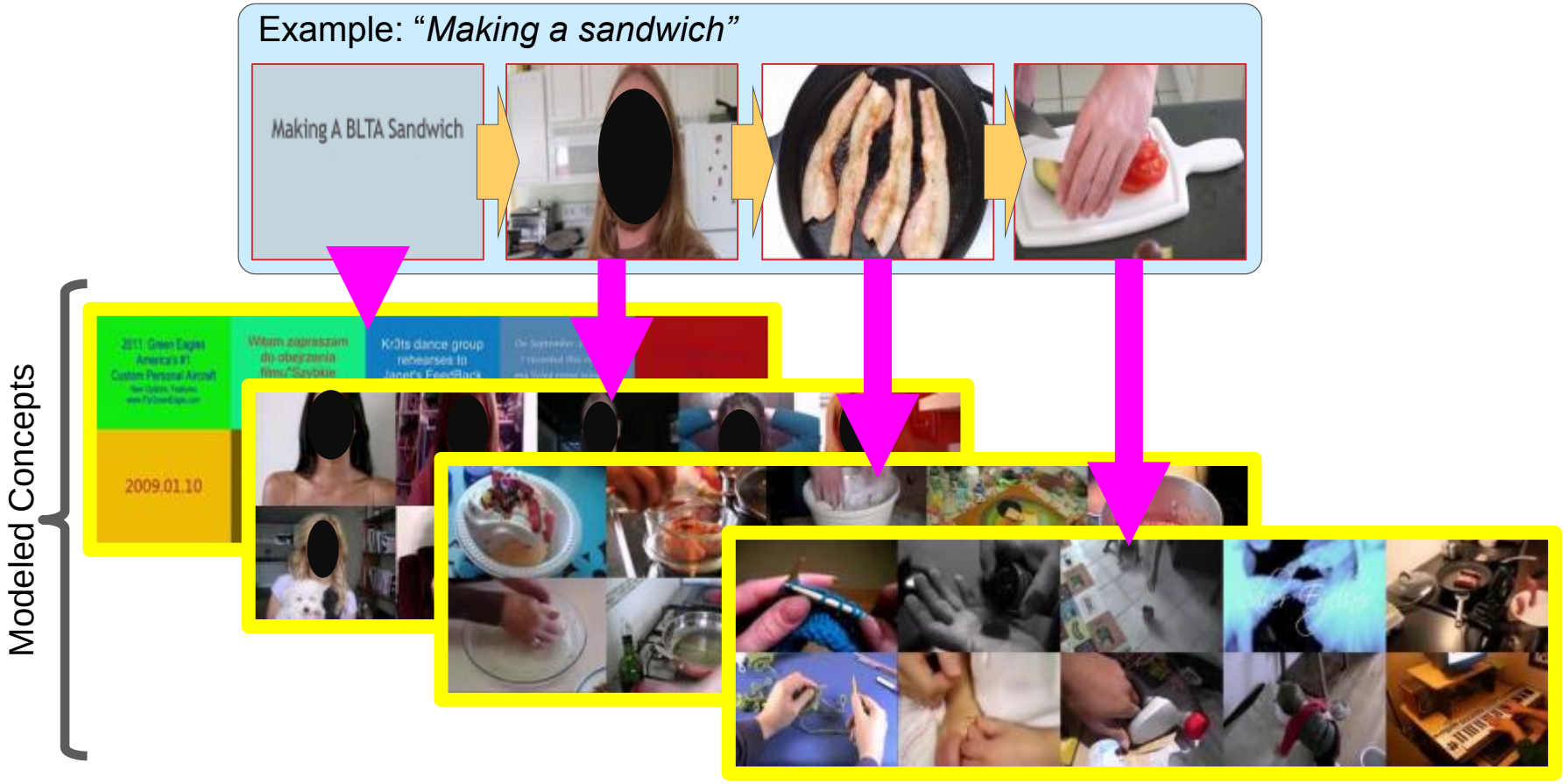
Best performance in each category marked in bold

* Lower number indicates higher accuracy

Results from ***Multimedia event detection with multimodal feature fusion and temporal concept localization*** Oh, McCloskey, Kim, at al. Machine Vision and Applications 25(1), 2014.

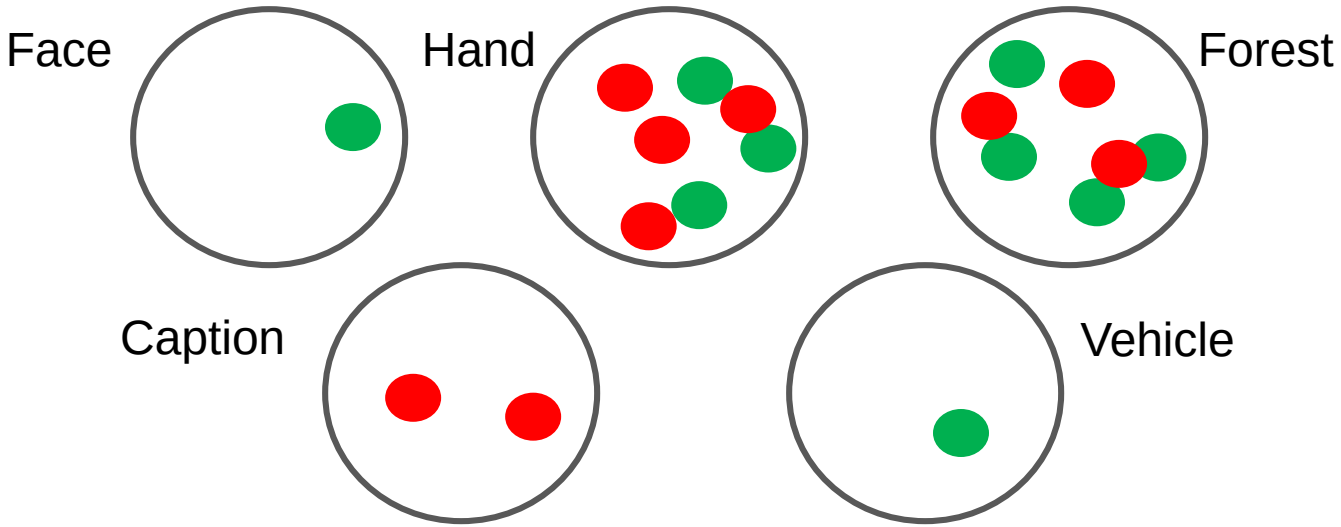
Event Structure Learning

Events have certain structures consisting of salient parts and non-important regions. How do we exploit and learn these?



Spatio-Temporal Weakly Supervised Learning

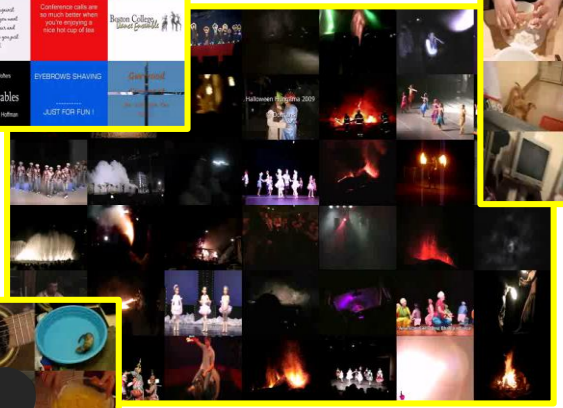
- **Weakly supervised learning formulation**
  - How do we identify important and salient segments from videos belonging to same events?
  - Can this be done implicitly or explicitly?
  - What should be the granularity in time and feature space which will work?



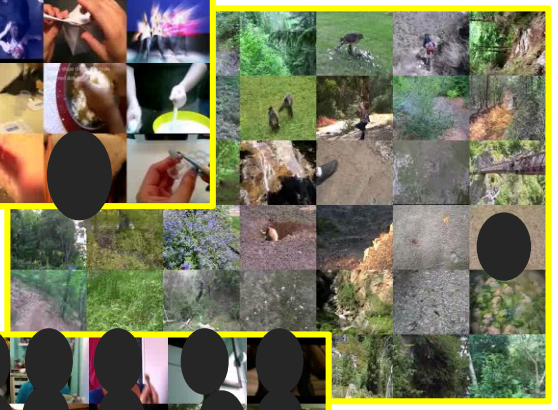
Mid-level: Frame Clusters

Title / Caption

Hands



Performance / Light Source



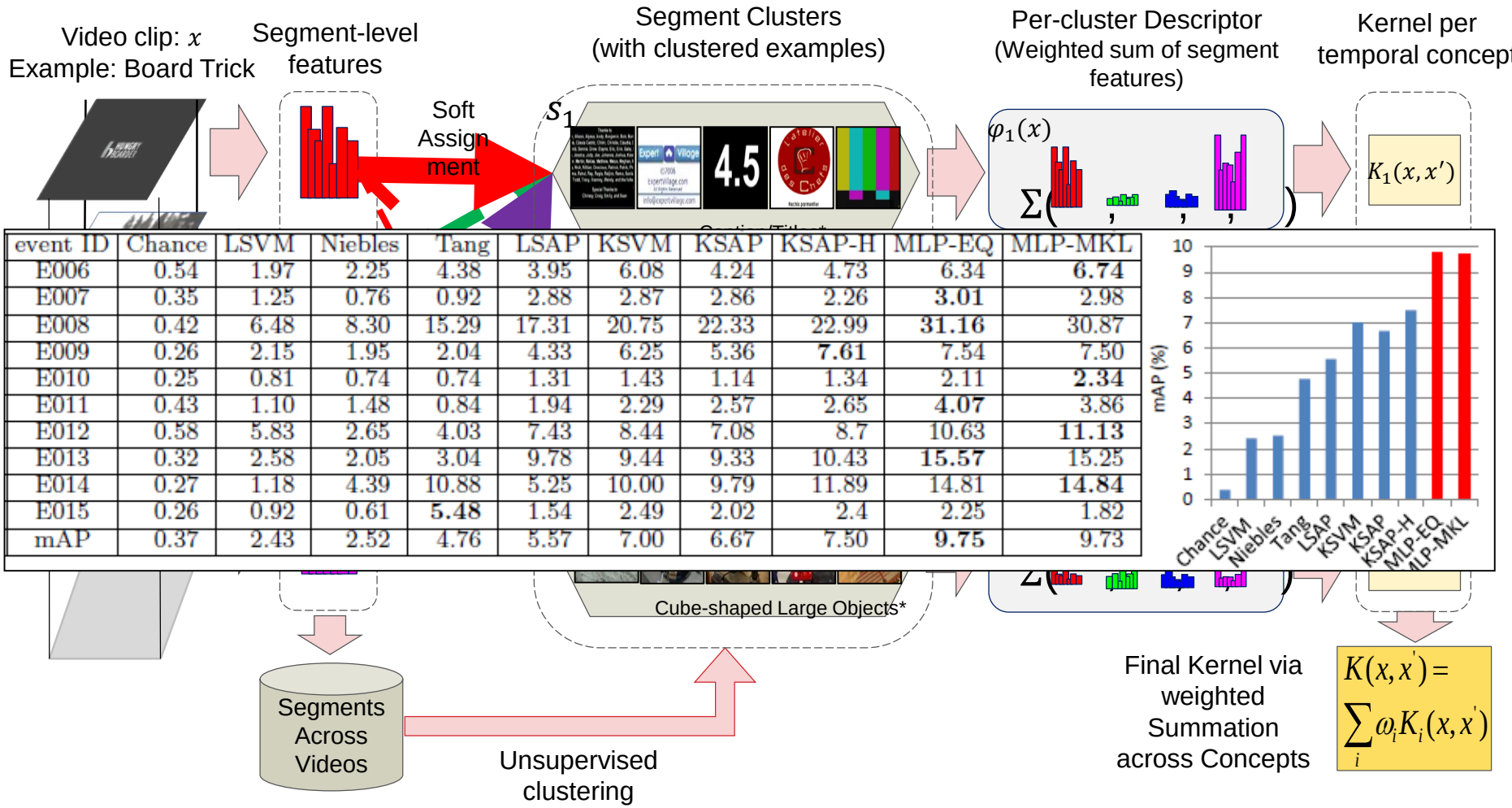
Grass / Leaves

Circular Objects

Group of People

Face Close-up

Learning Structure Implicitly using Topic-based Pooling



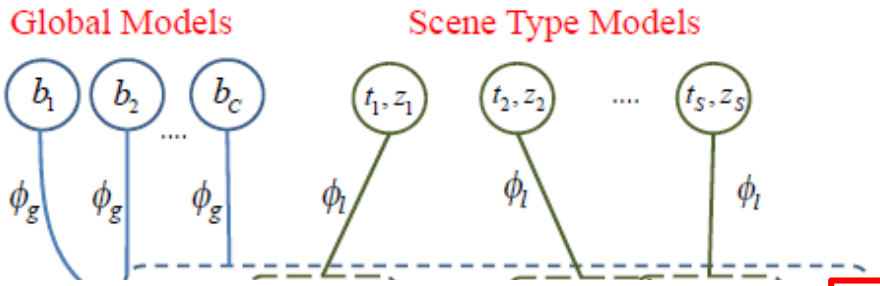
event ID	Chance	LSVM	Niebles	Tang	LSAP	KSVM	KSAP	KSAP-H	MLP-EQ	MLP-MKL
E006	0.54	1.97	2.25	4.38	3.95	6.08	4.24	4.73	6.34	**6.74**
E007	0.35	1.25	0.76	0.92	2.88	2.87	2.86	2.26	**3.01**	2.98
E008	0.42	6.48	8.30	15.29	17.31	20.75	22.33	22.99	**31.16**	30.87
E009	0.26	2.15	1.95	2.04	4.33	6.25	5.36	**7.61**	7.54	7.50
E010	0.25	0.81	0.74	0.74	1.31	1.43	1.14	1.34	2.11	**2.34**
E011	0.43	1.10	1.48	0.84	1.94	2.29	2.57	2.65	**4.07**	3.86
E012	0.58	5.83	2.65	4.03	7.43	8.44	7.08	8.7	10.63	**11.13**
E013	0.32	2.58	2.05	3.04	9.78	9.44	9.33	10.43	**15.57**	15.25
E014	0.27	1.18	4.39	10.88	5.25	10.00	9.79	11.89	14.81	**14.84**
E015	0.26	0.92	0.61	**5.48**	1.54	2.49	2.02	2.4	2.25	1.82
mAP	0.37	2.43	2.52	4.76	5.57	7.00	6.67	7.50	**9.75**	9.73

Segmental Multi-way Local Pooling for Video Recognition, Kim, Oh, Vahdat, Cannons, Mori, Perera. In ACM Multimedia 13.

Scene aligned pooling for complex video recogntion, Cao, Mu, Natsev, Chang, Hua, Smith, in ECCV 2012.

Recognition by Composition: Latent Temporal Part-based Learning

Explicitly Searches for
- Representative Segments
- Best Feature Combinations
- Best 'Hidden' segment Types



Event	Chance	Linear-SVM	Nieblas [8]	Tang [12]	Linear-SAP [1]	Linear-LSVM	KSVM	K-SAP [1]	KLSVM
E6	0.54	1.97	2.25	4.38	2.77	2.34	**6.08**	4.73	5.73
E7	0.35	1.25	0.76	0.92	2.11	1.33	2.87	2.26	**4.81**
E8	0.42	6.48	8.30	15.29	25.48	10.30	20.75	22.99	**35.82**
E9	0.26	2.15	1.95	2.04	4.14	1.79	6.25	7.61	**8.38**
E10	0.25	0.81	0.74	0.74	1.03	0.76	1.43	1.34	**2.12**
E11	0.43	1.10	1.48	0.84	1.93	1.41	2.29	2.65	**4.65**
E12	0.58	5.83	2.65	4.03	7.06	5.71	8.44	8.70	**10.99**
E13	0.32	2.58	2.05	3.04	10.38	2.57	9.44	10.43	**13.11**
E14	0.27	1.18	4.39	10.88	6.69	4.58	10.00	11.89	**23.32**
E15	0.26	0.92	0.61	**5.48**	1.21	1.09	2.49	2.4	3.29
mAP	0.37	2.43	2.52	4.77	6.28	3.19	7.00	7.50	**11.22**

Animal

Compositional Models for Video Event Detection: A Multiple Kernel Learning Latent Variable Approach, Vahdat, Cannons, Mori, Oh, Kim, in ICCV 2013

100 Ex Top 30 for "Flash Mob"

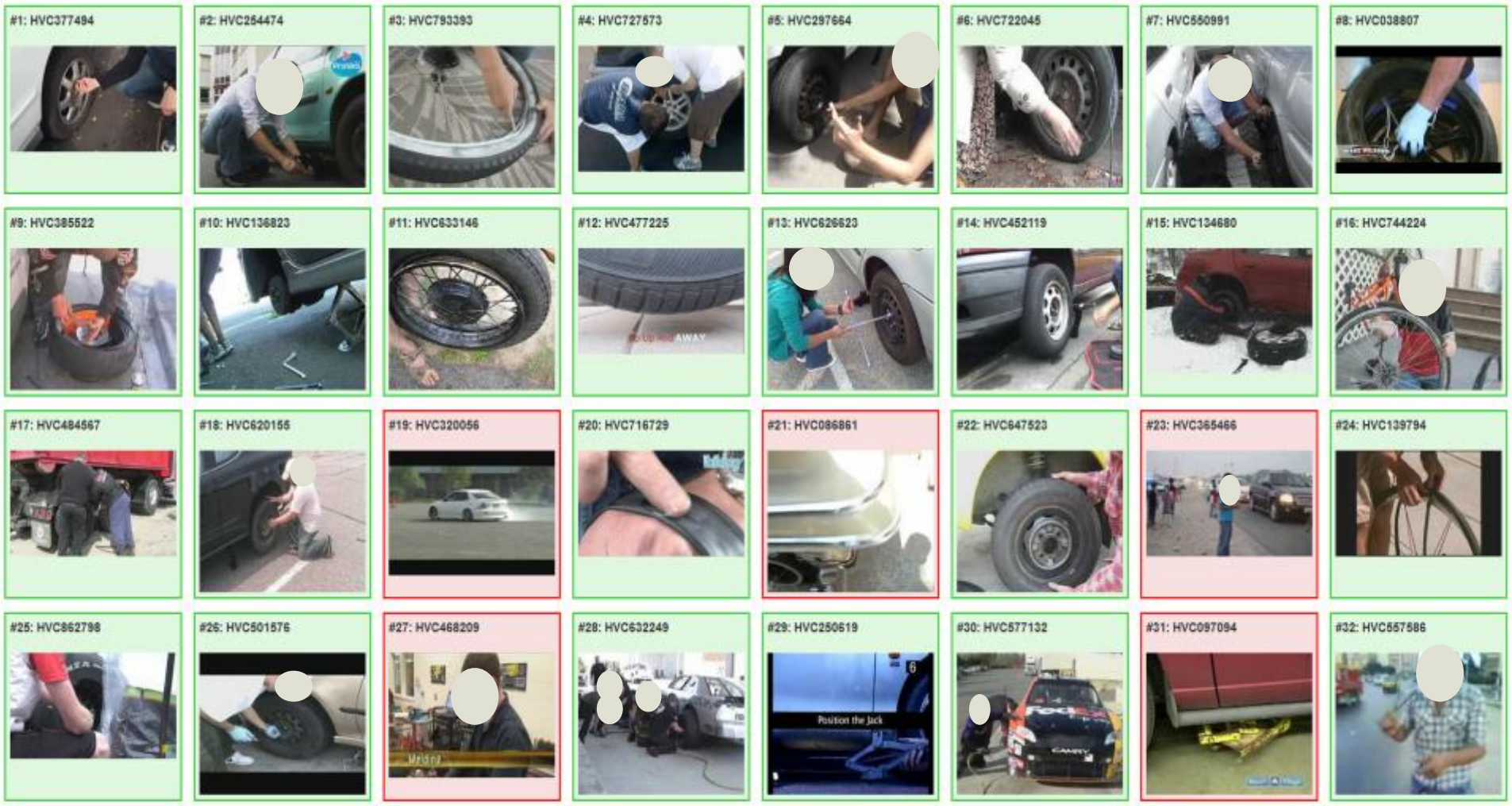
100 positive training examples used

- **It is possible to get a high quality set of matching videos from large archive**

Precision @ 32 = 97% as shown; AP = 74.3%; archive contains 26K videos including ~100 true positives

Top 30 for "Vehicle Tire Change"

100 positive training examples used

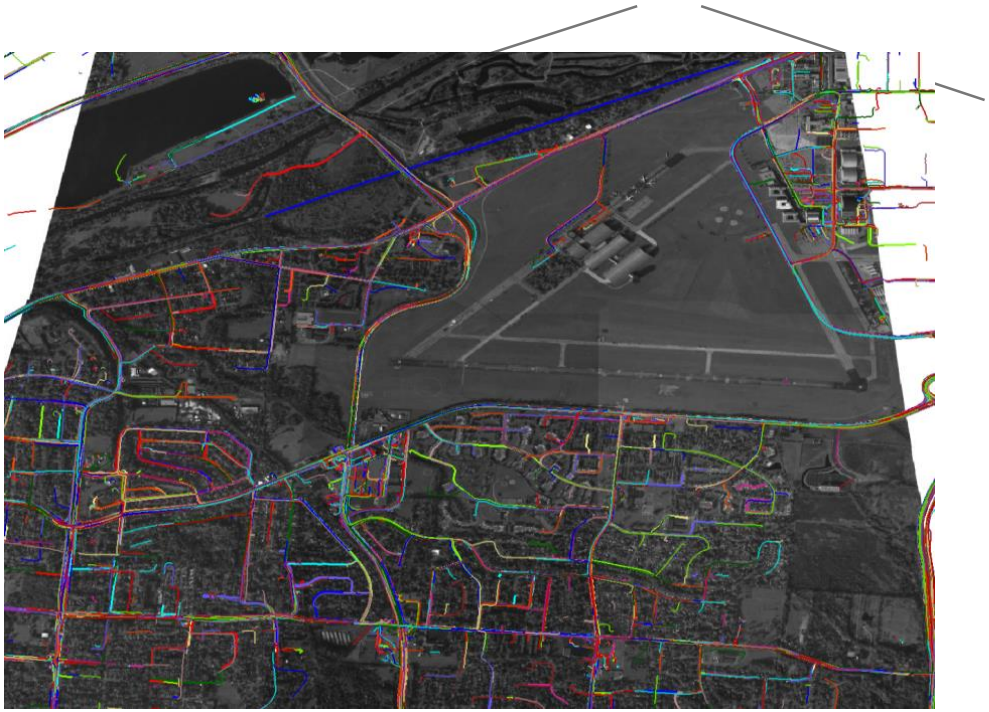


- **It is possible to get a high quality set of matching videos from large archive**

Precision @ 32 = 84% as shown; AP = 52.6%; archive contains 26K videos including ~100 true positives

Activity Recognition in Aerial Videos

Video from Sky



Characteristics

- Large Images/Videos
- Mostly vertical point of view
- Moving camera
- Small objects
- Lighting/Occlusion by nature
- Can have substantial scale changes

Application domains

- Disaster relief
- Emergency responder
- Broadcasting
- Traffic surveying/control
- Business Intelligence
- Security
- Military

Sensors: FMV and WAMI

Full Motion Video (FMV)

- Mostly single camera
- Moderate resolution
- User control
- Substantial camera motion

Wide Area Motion Imagery

- Multiple camera array
- Image stitching
- Very large image format
- Fairly good stabilization to point to certain area

Wright-Patterson Air Force Base (WPAFB) 2009 Dataset

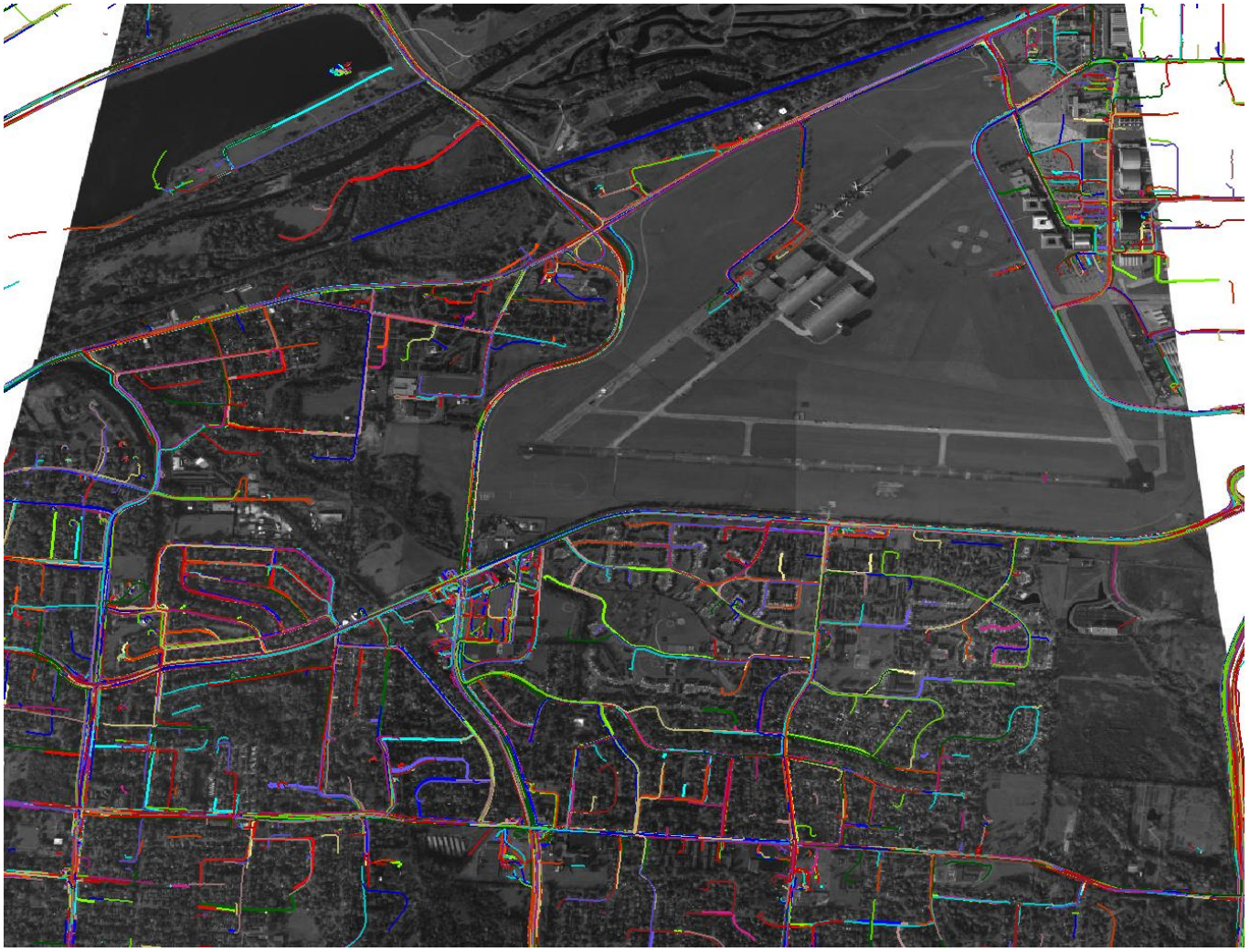
- Six cameras with ortho-rectified (stitched and geo-registered) imagery
- Image size: > 20K x 20K pixels
- GSD: 25 cm/pixel
- Frame rate: ~1.25Hz
- NITF file format with encoded sensor metadata

- 21 minutes (1,537 frames) of video
- **14 minutes (1,025 frames) with over 18K ground truth tracks**

- Publicly released by Air Force Research Lab (AFRL) SDMS

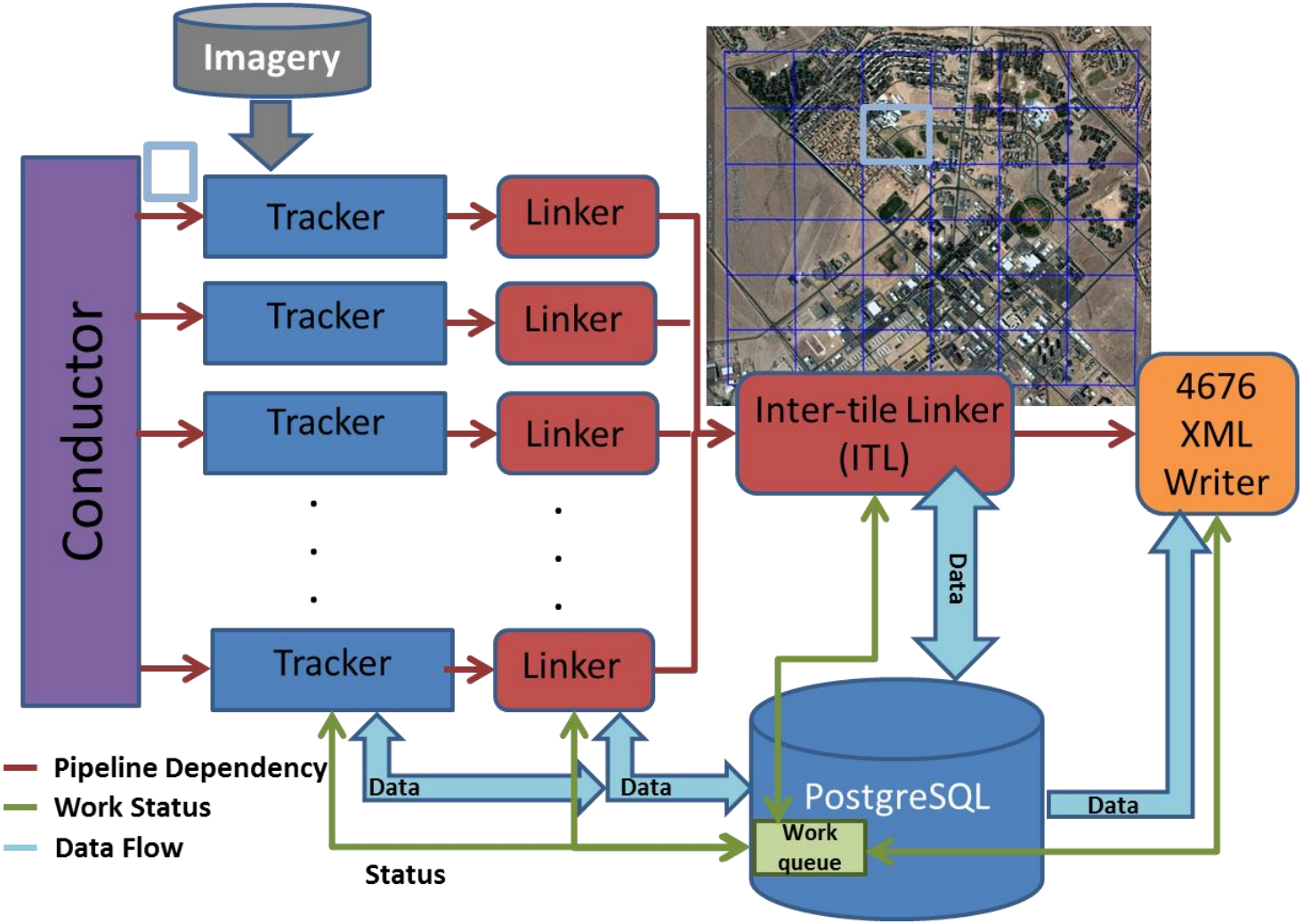
Source: https://www.sdms.afrl.af.mil/index.php?collection=wpafb2009

WPAFB Dataset: Track Ground Truth



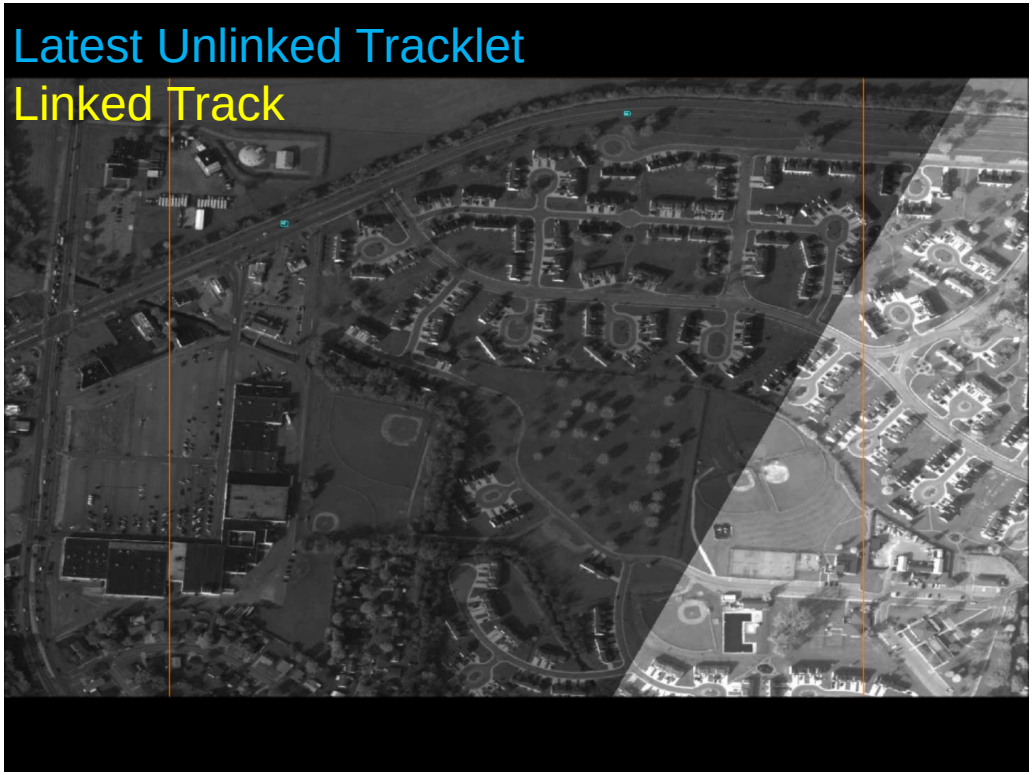
6,500+ ground-truth tracks in 7 minutes

Large-Scale Real-time Long-term Tracking



Real-time Multi-Target Tracking at 210 Megapixels/second in Wide Area Motion Imagery,
Basharat, Turek, Xu, Atkins, Stoup, Fieldhouse, Tunison, Hoogs, in WACV 2014

Large-Scale Real-time Long-term Tracking



- Tracks from an Area of Interest (AOI) processed as a single tile

- 6 min long track
- Includes inter-tile linking

Real-time Multi-Target Tracking at 210 Megapixels/second in Wide Area Motion Imagery,
Basharat, Turek, Xu, Atkins, Stoup, Fieldhouse, Tunison, Hoogs, in WACV 2014

Events & Actions

	Single-entity		Two-entity				Group
	Person	Vehicle	Person-Person	Person-Vehicle or Person-Object	Vehicle-Vehicle	Person-Facility or Vehicle-Facility	Person or Vehicle
Articulated Motion (Sub-entity)	Exploding Burning Digging Picking up Throwing Carrying Shooting Launching Limping Kicking Smoking Gesturing	Exploding Burning Shooting	Shaking hands Kissing Exchanging objects Kicking Carrying together	Exploding Burning Driving Opening/closing trunk Bicyling Loading/unloading Crawling under car Breaking window Shooting/launching Riding leading animal			Speaking to crowds Parade
Relative Motion (Track-level)	Walking Running Loitering	Starting Turning U-turn Stopping Aimless Driving Accelerating Decelerating	Following Meeting Gathering Moving as a group Dispersing	Getting in/out Dropping off Picking up	Overtaking or passing Moving together Maintaining distance Forming convoys Meeting	Entering Exiting Standing Dropping off Waiting at checkpoint Evading checkpoint Climbing atop Passing thru gate	Convoy Receiving line Queuing Troop formation

Data Requirements:

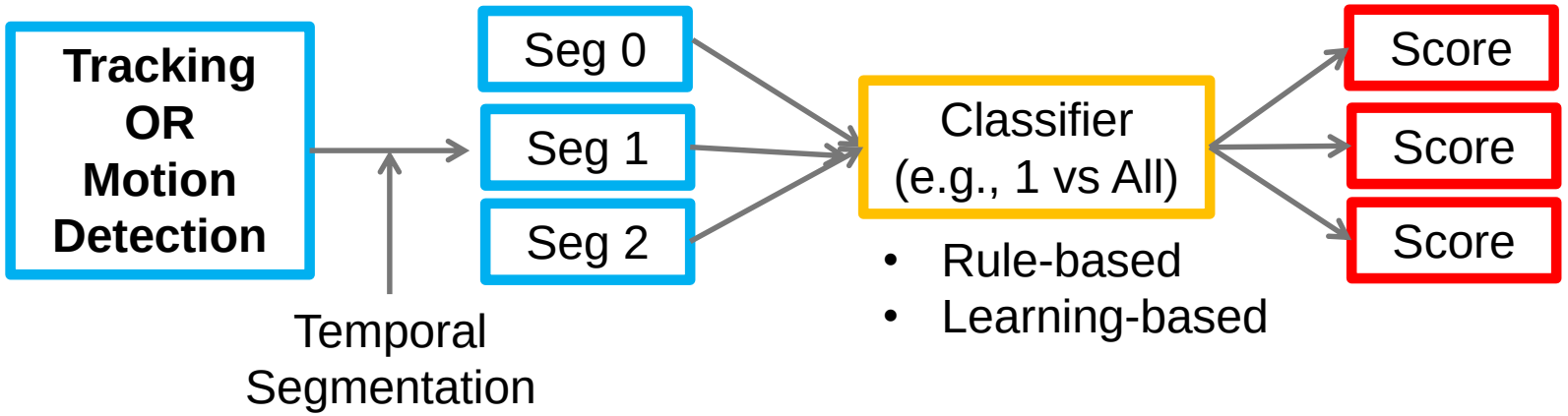
- **Low Resolution: possible by analyzing track-level information**
- **High Resolution: requires detailed pixel information**

Continuous Visual Event Recognition (CVER)

Common Architecture

- Foreground motion detection, e.g., tracking etc.
- Temporal segmentation, e.g., regular/variable units
- Classification, e.g., 1-vs-All, multi-way etc.
- Upper bound determined by weakest among above

Lots of blank intervals/space are challenging to optimize precision and recall

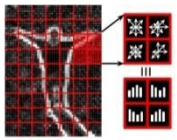


Human Actions in Aerial Video

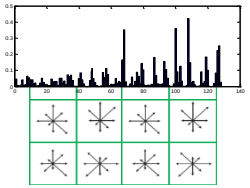
At low resolution, many actions look very similar

Event Models & Features

Pixel-based Features



spatio-temporal histogram of gradients [6][12]



Feature point descriptor: SIFT [10][11][15]

optical flow [2][6][9]

histogram of optical flow [6]

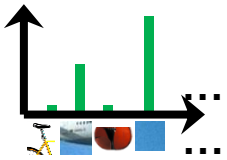
Macro Features

sensor metadata: gsd, pointing angles, etc.

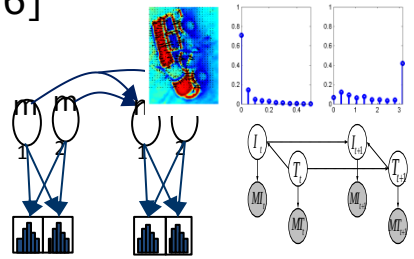
bounding box: area, aspect ratio, etc.

track level: speed, delta heading, curvature, etc. [1][12][13][8]

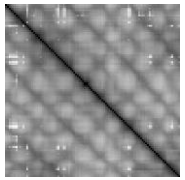
Models & Classification



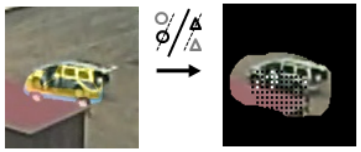
BoW + SVM [4][11][16]



Dynamic Bayesian networks [12][13]



periodicity of self-similarity matrices [14]



objects interaction modeling [1][7][8]

References

1. U. Gaur, B. Song, A. Roy-Chowdhury, Query-based Retrieval of Complex Activities using “Strings of Motion-Words”, IEEE Workshop on Motion and Video Computing, 2009.
2. Shandong Wu, Omar Oreifej, and Mubarak Shah, “Action Recognition in Videos Acquired by a Moving Camera Using Motion Decomposition of Lagrangian Particle Trajectories”, ICCV 2011.
3. Subhabrata Bhattacharya, Rahul Sukthankar, Rong Jin, and Mubarak Shah, “A Probabilistic Representation for Efficient Large Scale Visual Recognition Tasks”, IEEE CVPR, 2011.
4. Jingen Liu, Yang Yang, Imran Saleemi and Mubarak Shah, “Learning Semantic Features for Action Recognition via Diffusion Maps”, To appear in Computer Vision and Image Understanding.
5. Aniruddha Kembhavi, David Harwood, Larry S. Davis: Vehicle Detection Using Partial Least Squares. IEEE Trans. Pattern Anal. Mach. Intell. 33(6): 1250-1265 (2011)
6. C.-C. Chen and J. K. Aggarwal, "Recognizing Human Action from a Far Field of View", IEEE Workshop on Motion and Video Computing (WMVC), Utah, USA, December 2009.
7. J. T. Lee, M. S. Ryoo, and J. K Aggarwal, "View Independent Recognition of Human-vehicle Interactions using 3-D Models", IEEE Workshop on Motion and Video Computing (WMVC), Utah, USA, December 2009.
8. J. T. Lee*, C.-C. Chen*, and J. K. Aggarwal,, "Recognizing Human-Vehicle Interactions from Aerial Video without Training", Workshop of Aerial Video Processing in conjunction with CVPR (WAVP), Colorado Springs, CO, June 2011
9. N. M Nayak, B. Song, A. K. Roy-Chowdhury, " Dynamic Modeling of Streaklines for Motion Pattern Analysis in Video", CVPR Workshop on Machine Learning for Vision-based Motion Analysis, 2011.
10. Y.-G. Jiang, J. Yang, C.-W. Ngo, A. Hauptmann, “Representations of Keypoint-Based Semantic Concept Detection: A Comprehensive Study”, IEEE Trans. on Multimedia, 2010.

References (cont'd)

11. SF Chang, J He, YG Jiang, CW Ngo, A Yanagawa, Columbia University/VIREO-CityU/IRIT TRECVID2008 high-level feature extraction and interactive video search

12. Swears E., Hoogs A., Learning and Recognizing Complex Multi-Agent Activities with Applications to American Football Plays, Workshop on the Applications of Computer Vision , 2012

13. Zhi Zeng and Qiang Ji, Knowledge Based Activity Recognition with Dynamic Bayesian Network, ECCV 2010

14. Efros, A.A.; Berg, A.C.; Mori, G.; Malik, J.; , "Recognizing action at a distance," *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on* , vol., no., pp.726-733 vol.2, 13-16 Oct. 2003

15. David G. Lowe, Distinctive Image Features from Scale-Invariant Keypoints, International Journal of Computer Vision, 2004, Volume 60, Number 2, Pages 91-110

16. S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C-C. Chen, J.T. Lee, S. Mukherjee, J.K. Aggarwal, H. Lee, L. Davis, E. Swears, X Wang, Qiang Ji, K. Reddy, M, Shah, C.Vondrick, H. Pirsiavash, D. Ramanan, J. Yuen, A. Torralba, Bi Song, A. Fong, A. Roy-Chowdhury, and M. Desai, A Large-scale Benchmark Dataset for Event Recognition in Surveillance Video, CVPR 2011

17. Paul Over, George Awad, Jonathan Fiscus, Brian Antonishek, "TRECVID 2011--Goals, Tasks, Data, Evaluation Mechanisms and Metrics", in TRECVID '11 notebooks

18. S. Sadanand and J. J. Corso. "Action bank: A high-level representation of activity in video". In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

19. Q.V. Le, W.Y. Zou, S.Y. Yeung, A.Y. Ng. "Learning hierarchical spatio-temporal features for action recognition with independent subspace analysis" in *CVPR*, 2011.

20. Yang Wang and Greg Mori. "Hidden Part Models for Human Action Recognition: Probabilistic vs. Max-Margin".*IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, 33(7) pp.1310-1323 2011

Event Detections on WPAFB

Turn

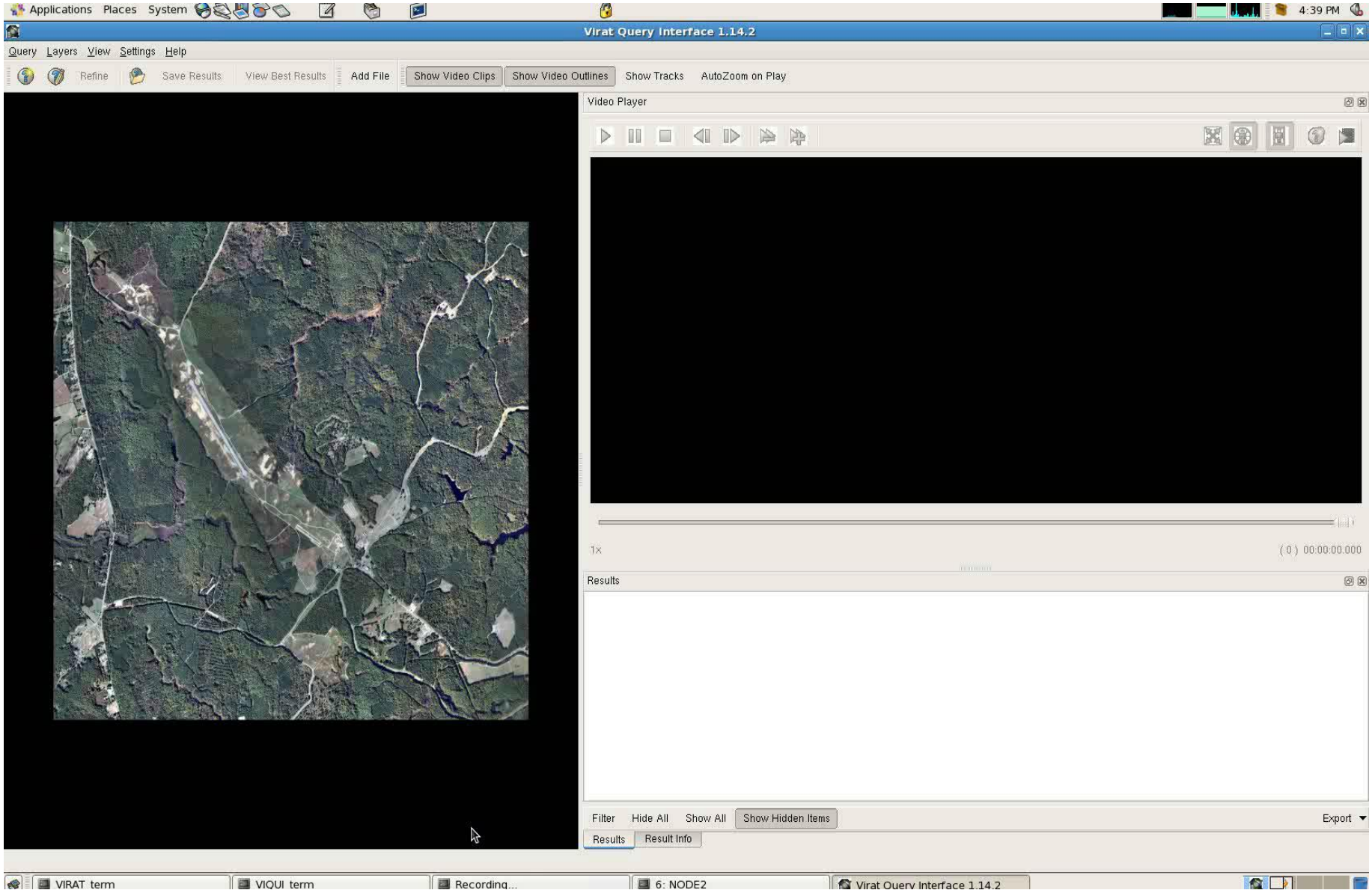
Stop

Meeting

Follow

Aimless Driving

Event Detections on FMV dataset



Activity-based Scene Understanding

Objective: Recognize stationary scene elements based on surrounding pedestrian and vehicle behaviors, as opposed to appearance features

WAMI Area of Interest(AOI)

(0.12 GSD, ~21 minutes, 3.3Hz)

Main Street Web Cam

(~0.05-0.18 GSD, ~8 hours, 2Hz)

- Buildings
- Intersection
- Cross-walk
- Roadway
- Parking-
- Sidewalk
- Doorway

Where are more scene elements similar to these?

Key Challenges (Multi-Modal Behaviors)

Different scene element instances can have significantly different behaviors

Roadways

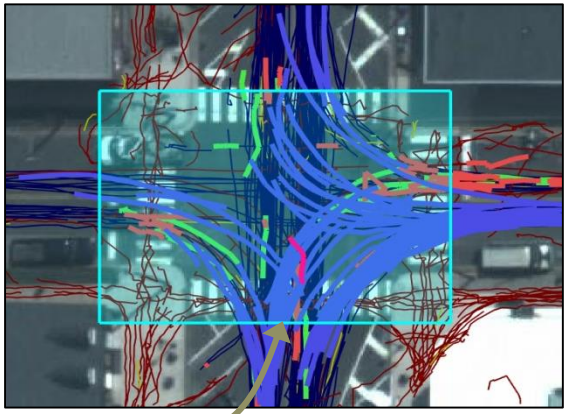
Many Modes

Few Modes

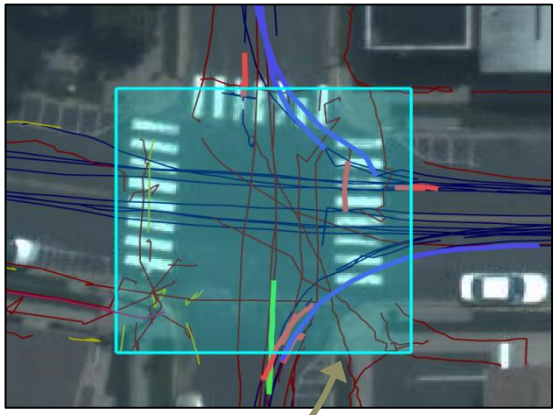
Event Legend

- Vehicle Driving
- Vehicle Turning
- Vehicle Starting
- Vehicle Stopping
- Person Walking

Intersection



Many Modes

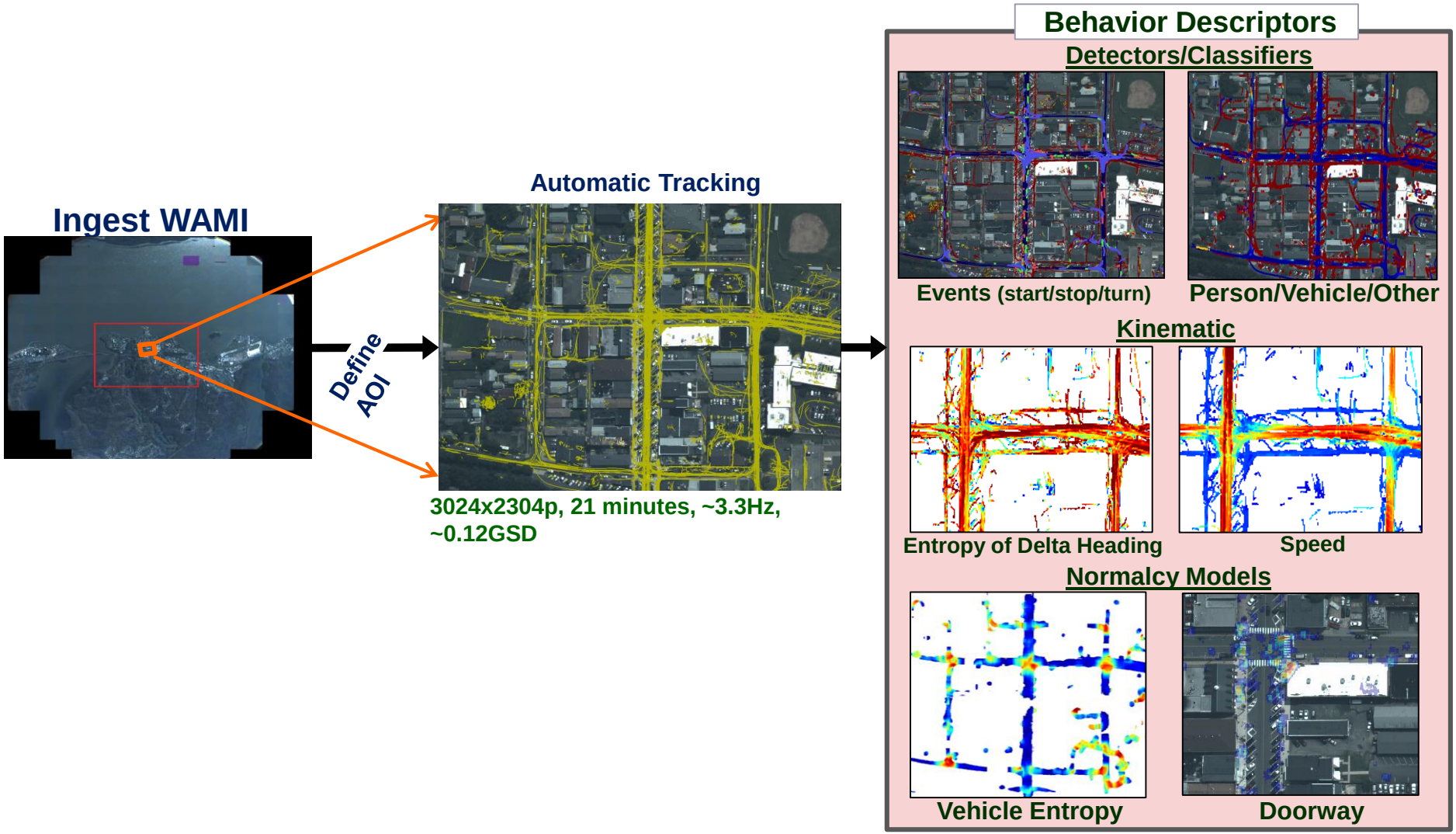


Few Modes

Pyramid Coding for Functional Scene Element Recognition in Video Scenes, Swears, Boyer, and Hoogs, in ICCV 2013

Features: Object, Motion, Statistics

Objective: Extract activity behavior descriptors using automatically computed tracks

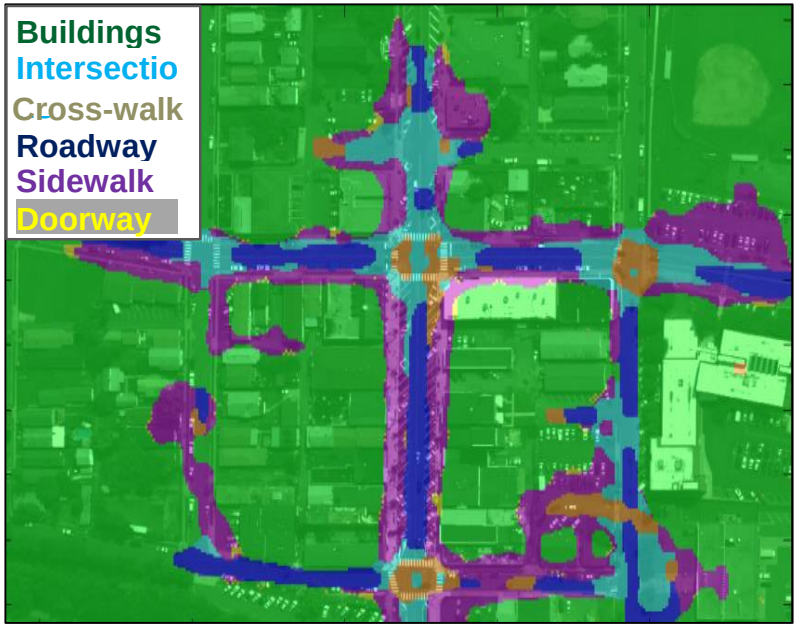


Pyramid Coding for Functional Scene Element Recognition in Video Scenes, Swears, Boyer, and Hoogs, in ICCV 2013

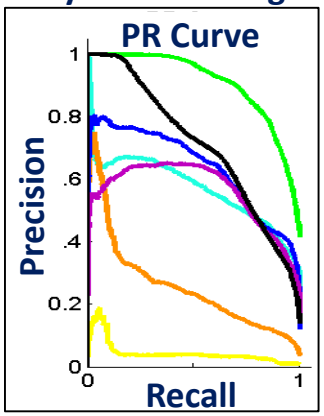
Activity-based Scene Understanding

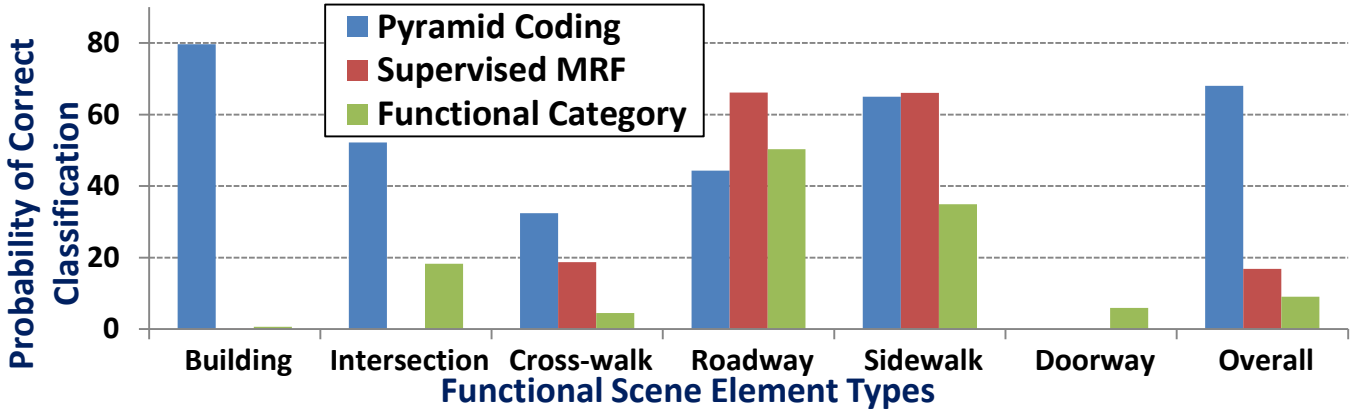
True Evaluated Scene Elements

Pyramid Coding, MRF Labels



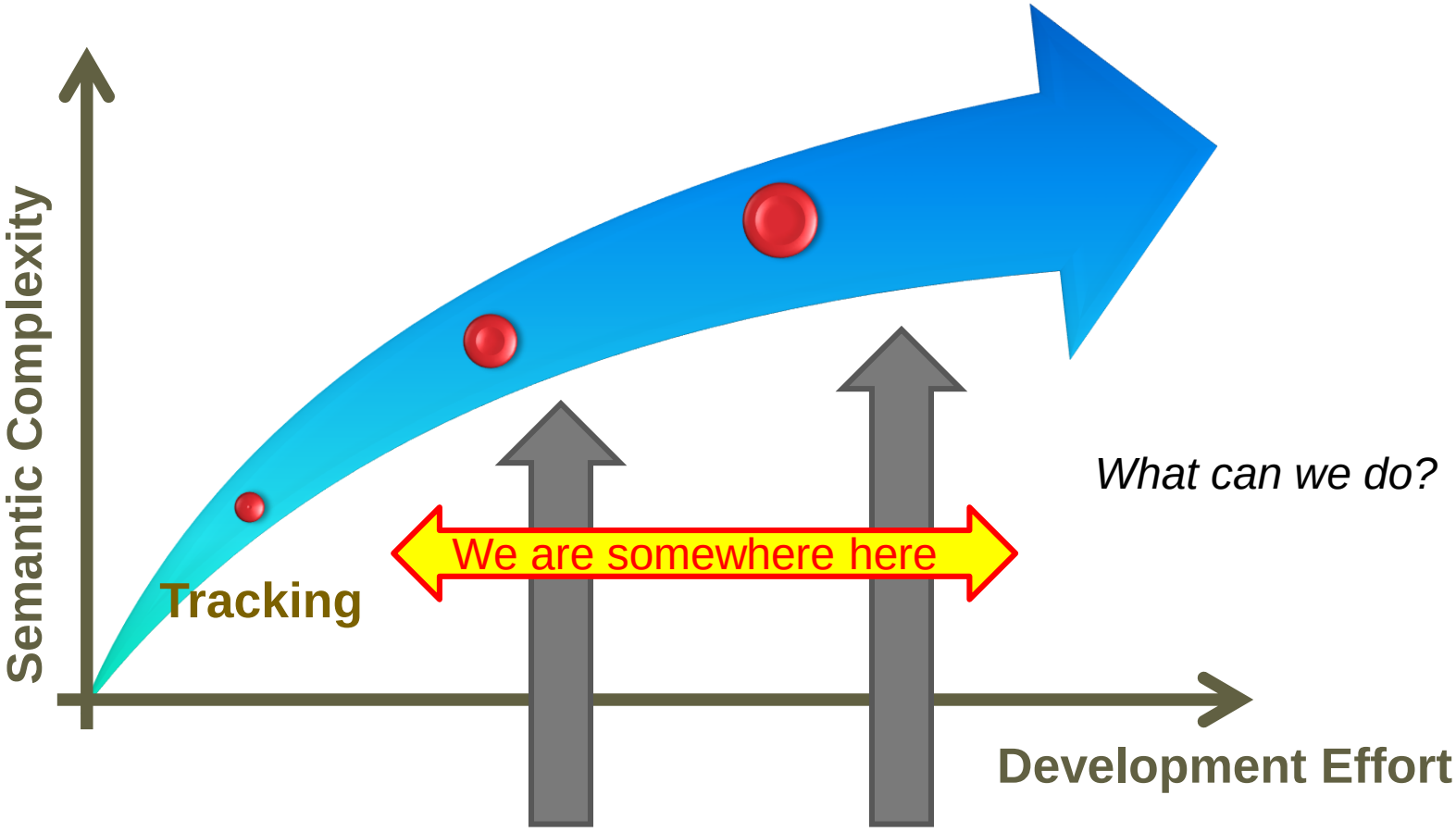
Pyramid Coding





Pyramid Coding for Functional Scene Element Recognition in Video Scenes, Swears, Boyer, and Hoogs, in ICCV 2013

Where are we and heading to?



FMV

- Event detection
- Event-based video indexing

WAMI

- Event detection
- Normalcy models & anomaly detection
- Complex activity recognition

Sports Video Analytics

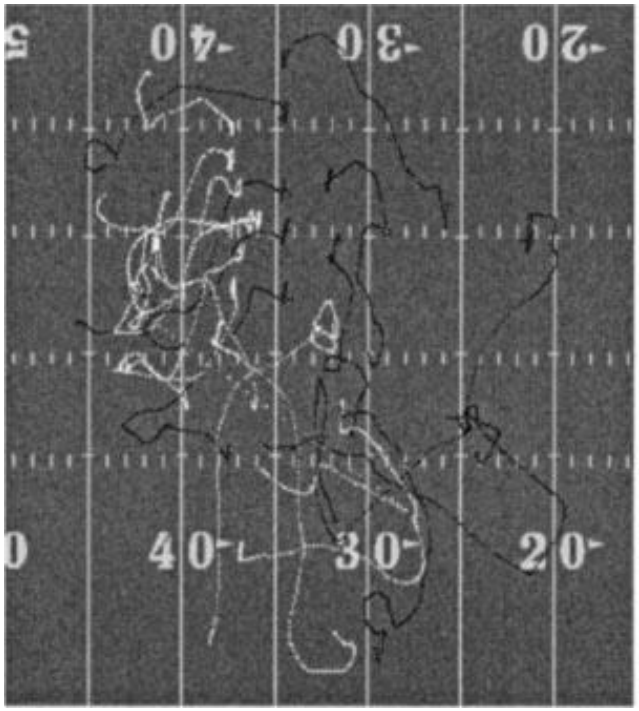
Sports Analytics

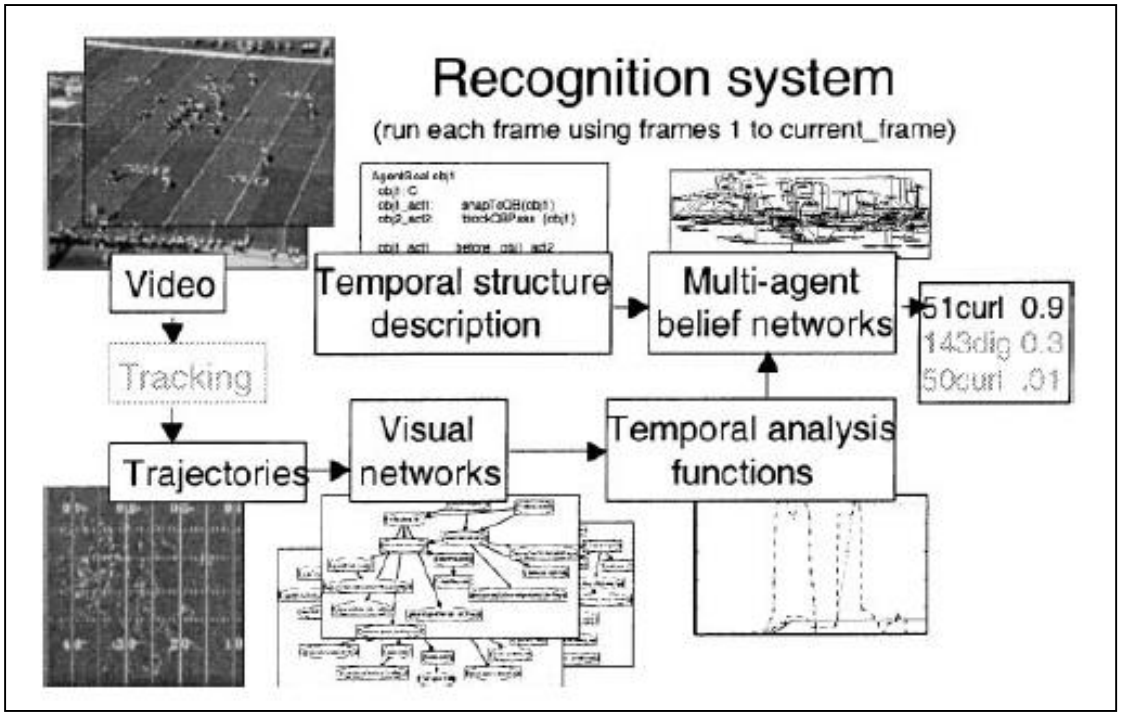
Application domains

- Player Tracking
- Event / Strategy recognition
- Event-based indexing
- (Semi) Automatic camera capture control
- Summarization

Early Work: Intille and Bobick

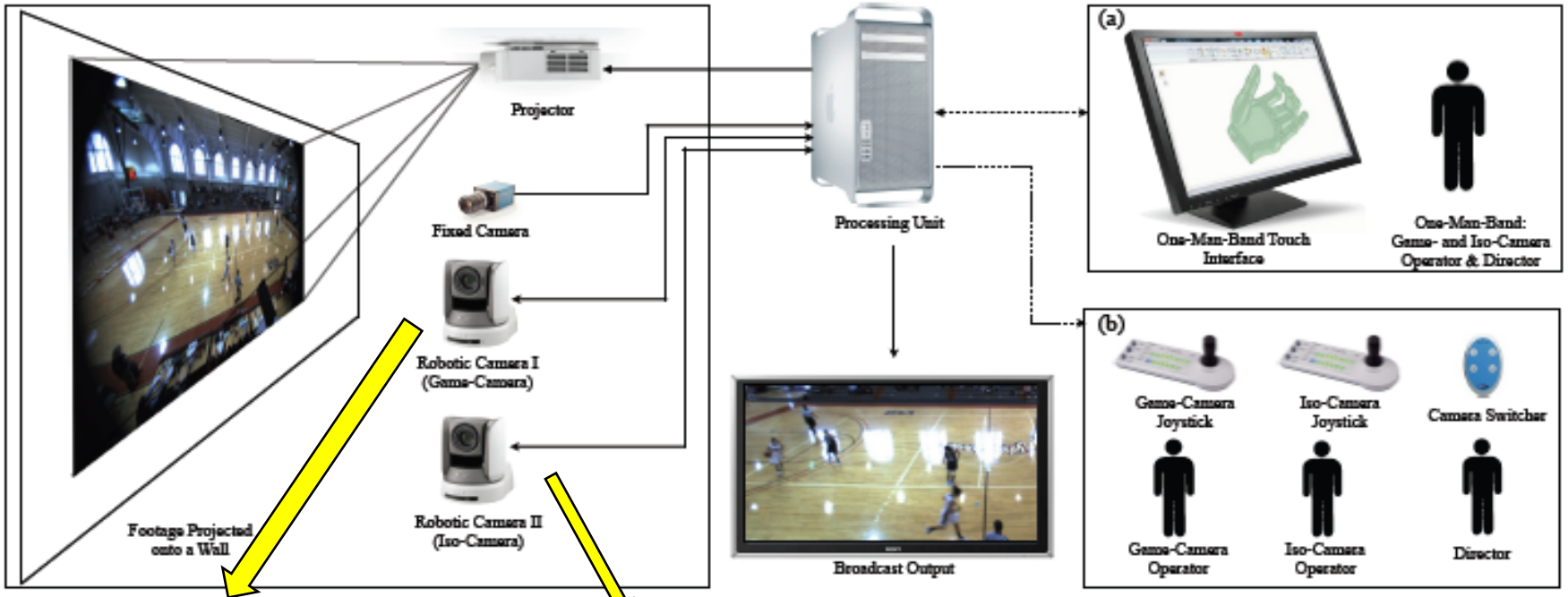
- Assumed Perfect Tracking information and role recognition
- Bayesian networks to agglomerate evidence and infer team playing strategies
- Limitation: Sports videos are chaotic, and we never get perfect features!

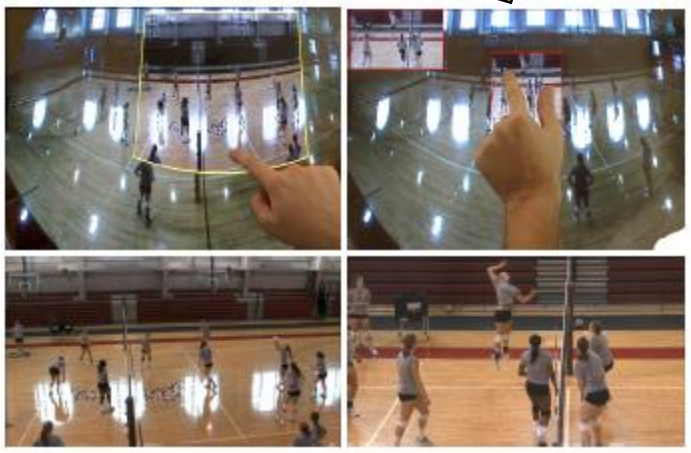




Recognizing Planned Multiperson Action, Intille and Bobick, Computer Vision and Image Understanding 81, 2001

Today: Towards Automated Sports Broadcasting





How do we help a single or a small number of crew(s) to manage the entire broadcasting feed?

One-Man-Band: A Touch Screen Interface for Producing Live Multi-Camera Sports Broadcasts, Foote, Carr, Lucey, Sheikh, Matthews, in ACM Multimedia 2013

Group Motion Prediction for Camera Control

Backdoor play (through pass) (soccer)

Compute motion fields based on player motions. Then, find convergence points to predict where ball (and players) are moving to.

Motion Fields to Predict Play Evolution in Dynamic Sports Scenes, Fkim, Grundmann, Shamir, Matthews, Hodgins, Essa, in CVPR 2010.

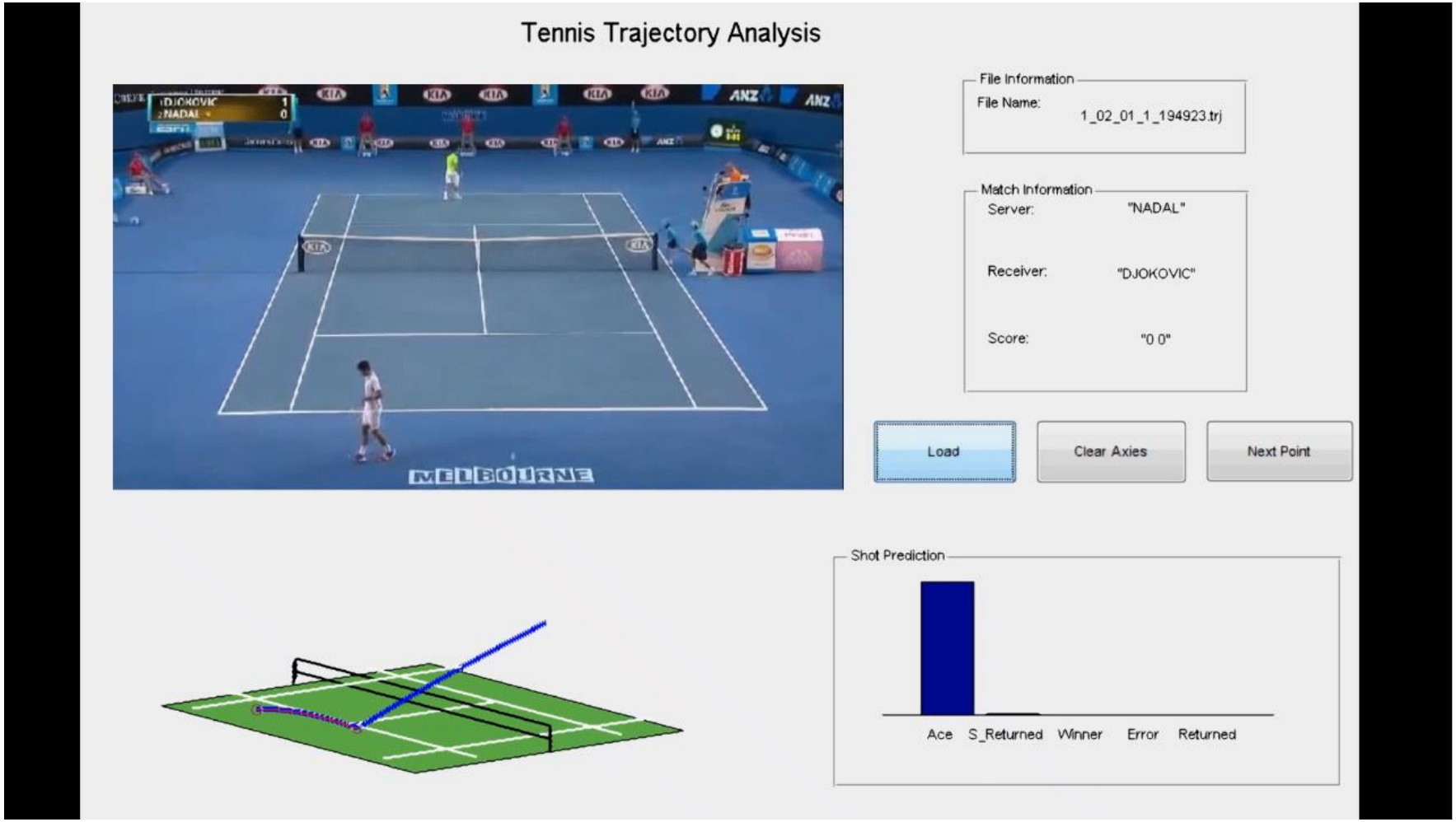
Prediction-based Video Re-targeting

Original Video Feed
among many camera feeds

Automatically computed
Re-targetted video

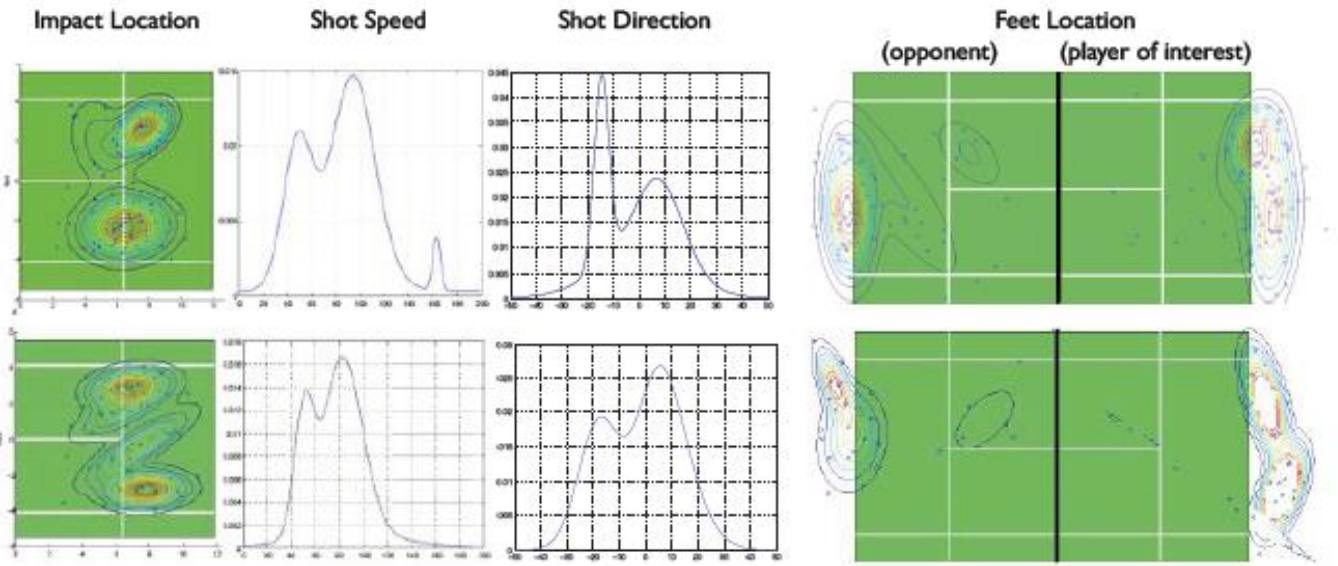
Motion Fields to Predict Play Evolution in Dynamic Sports Scenes, Fkim, Grundmann, Shamir, Matthews, Hodgins, Essa, in CVPR 2010.

Predicting Wins based on detailed trajectory analysis



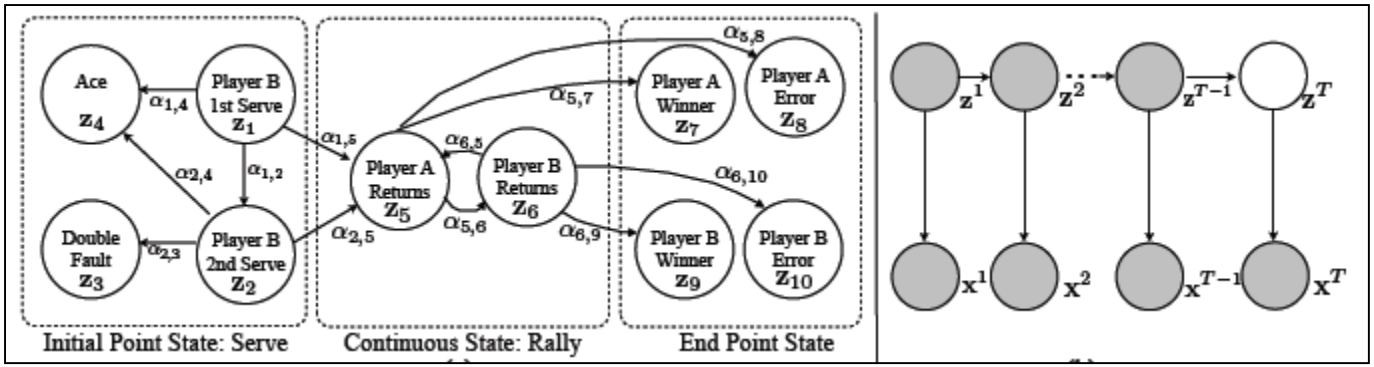
Sweet-Spot: Using Spatiotemporal Data to Discover and Predict Shots in Tennis, Wei, Lucey, Morgan, Sridharan, in MIT Sloan Sports Analytics Conference

Predicting Wins based on detailed trajectory analysis



Features

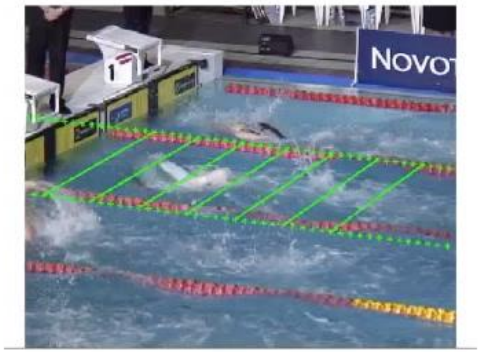
Bayesian networks for Prediction



Sweet-Spot: Using Spatiotemporal Data to Discover and Predict Shots in Tennis, Wei, Lucey, Morgan, Sridharan, in MIT Sloan Sports Analytics Conference

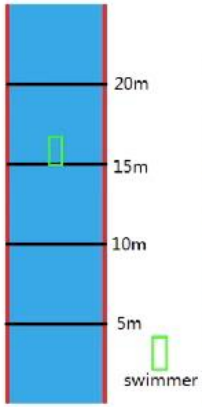
Swimming Video Analysis

Analyzing a Large Collection of Archived Swimming Videos



Calibration

Swimmer Location



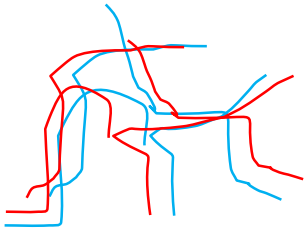
Part Tracking

L. Sha, P. Lucey, S. Morgan, S. Sridharan and D. Pease, "Understanding and Analyzing a Large Collection of Archived Swimming Videos", in *WACV* 2014

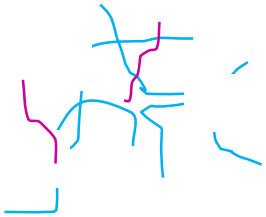
Understanding and Analyzing a large Collection of Archived Swimming Videos,
Sha, Lucey, Morgan, Sridharan, Pease, in in WACV 2014

Complex Play Recognition with Imperfect Tracks

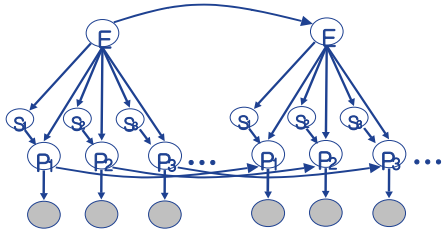
Which play is being run? How soon can we tell?



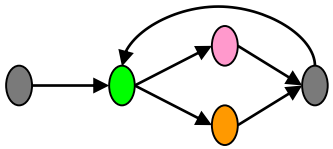
Spatial variability



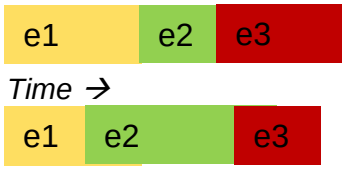
Fragmented and partial tracks



Complex object



Partial Temporal Ordering

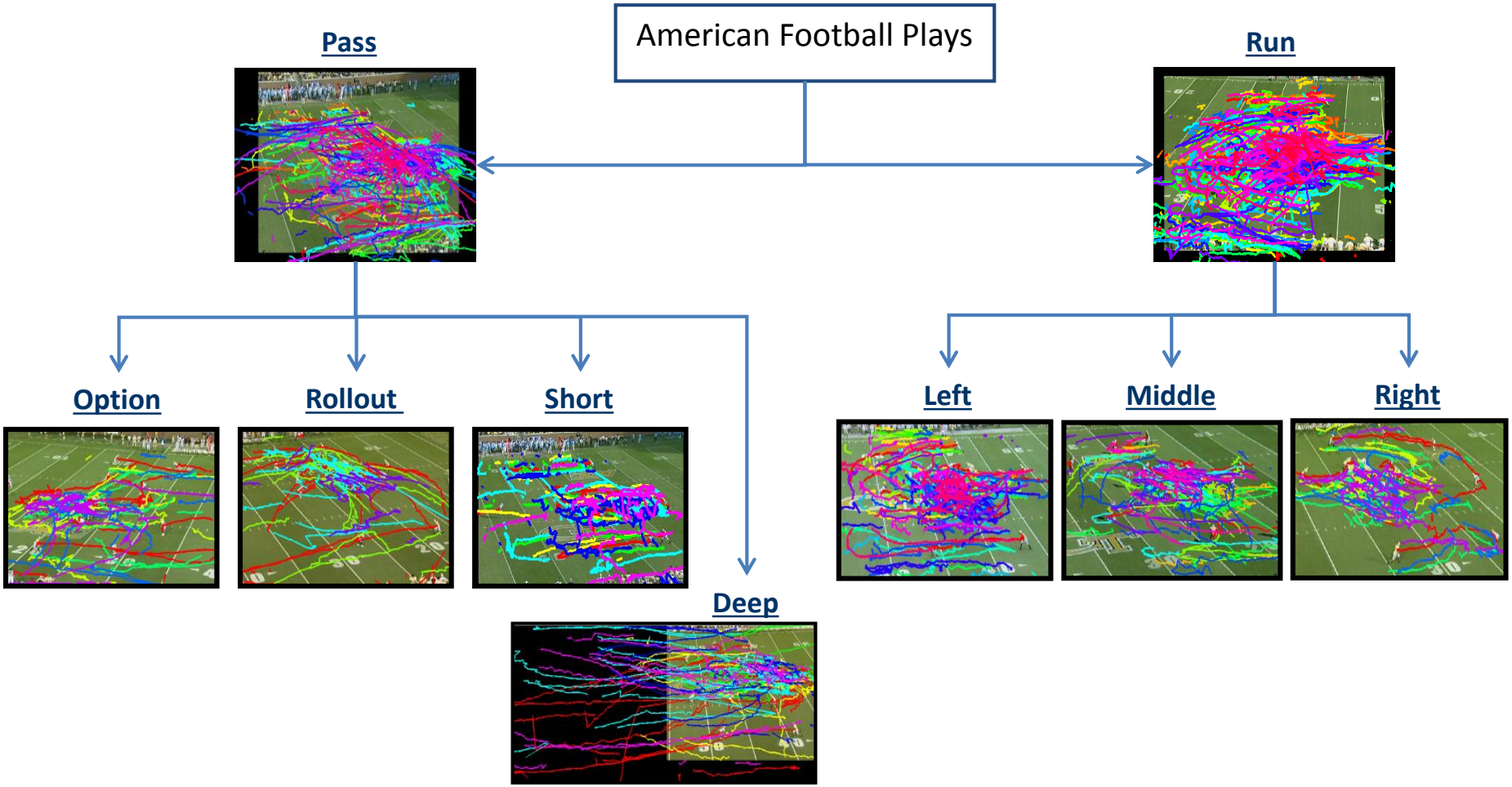


Temporal variability

Camera Motion

Active Deception

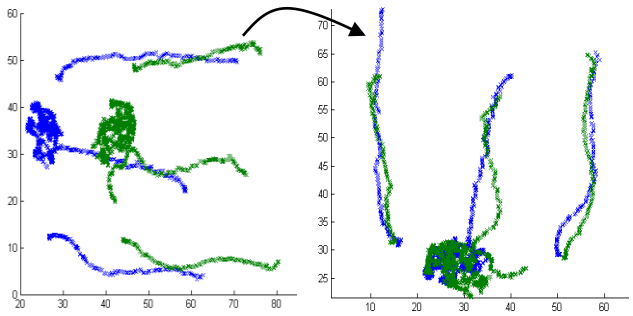
Play Taxonomy



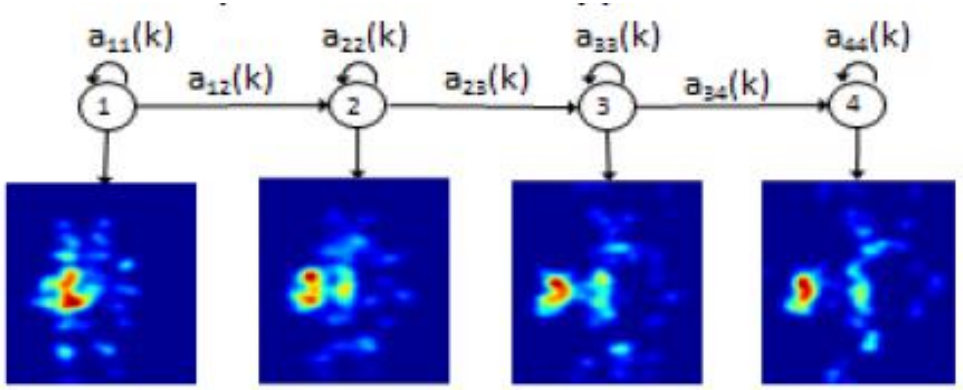
Learning and Recognizing Complex Multi-Agent Activities with Applications to American Football Plays, Swears and Hoogs in WACV 2012

Robust Play Recognition against Track Fragmentation

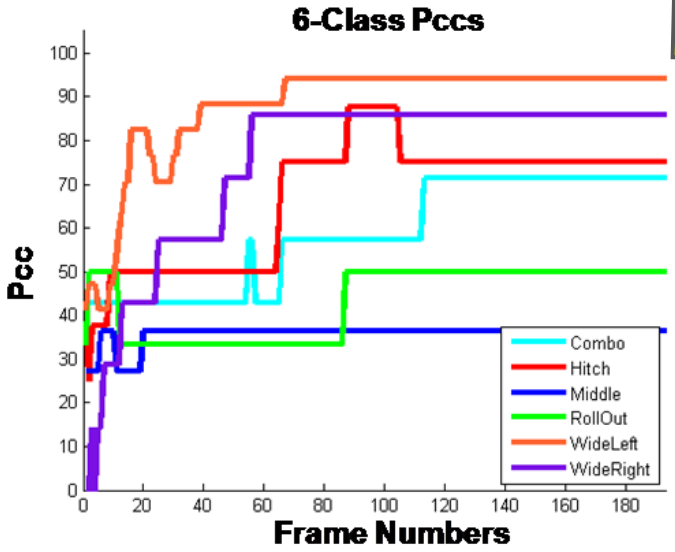
Track Normalization



Learn Non-Stationary HMM
using positions and speeds from tracks



Tracker ID is not important anymore!



Accuracy
Improves with
More observations

Learning and Recognizing Complex Multi-Agent Activities with Applications to American Football Plays, Swears and Hoogs in WACV 2012

Summary

Unconstrained Video Search

Sports Video Analysis

Aerial Video Analysis