Benchmarking Human Activity Recognition

Haowei Liu, Perceptual Computing, Intel
Rogerio Feris, IBM T.J. Watson Research Center
Ming-Ting Sun, University of Washington
6/21/2012

Introduction & Motivation

- Need datasets to benchmark different aspects of algorithms
- Need a common ground for researchers to evaluate/compare the performances of their approaches
- Need datasets that are good representatives of the problem being solved (solving the dataset vs. solving the problem)
- Focus and introduce the state of the art benchmarking video datasets for activity recognition

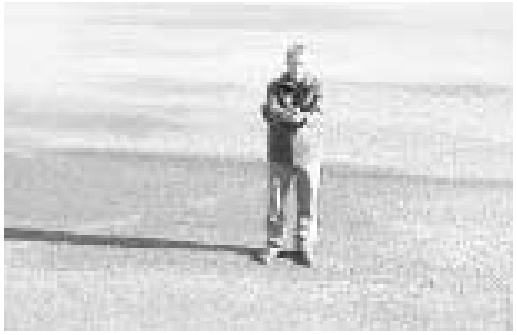
Outline

- Benchmark for Kinematic Activities
- Movie/Web Videos Benchmarks
- Benchmarks for Assisted Daily Life (ADL) Activities
- Video Surveillance Benchmarks
- Benchmarks for Group Activities
- Multi-Camera Benchmarks
- RGB-D Benchmarks
- Egocentric Benchmarks

Benchmarks for Kinematics Activities

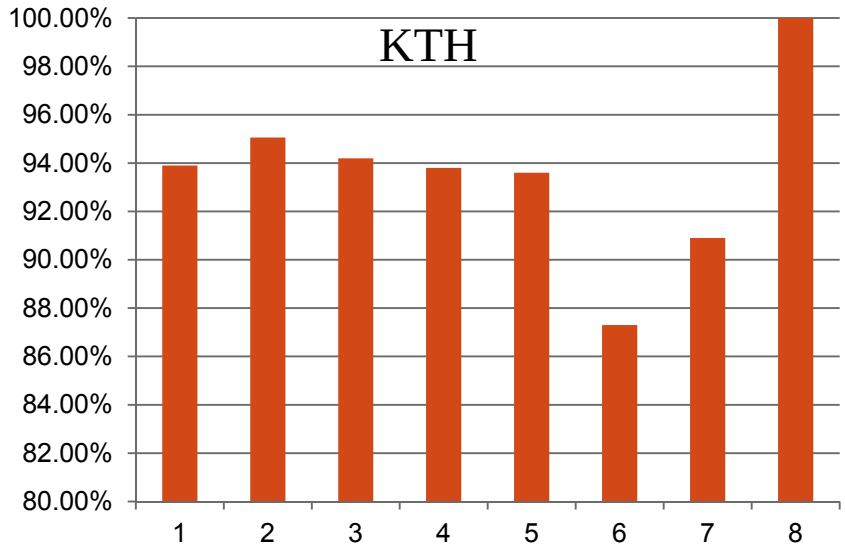
KTH* and Weizmann+

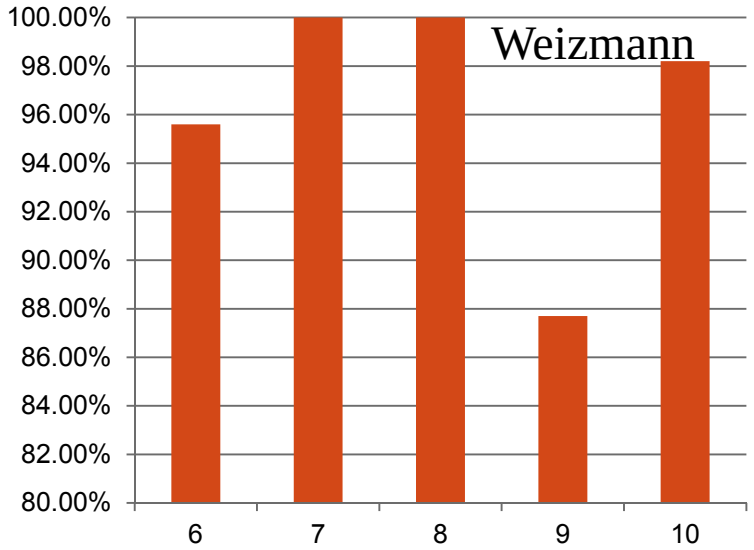
- Low resolution (<200x200)
- Few background clutters
- Mostly frontal and side-on camera viewing angles
- High accuracy reported by many papers already



KTH

Weizmann





*[SchuldtICPR2004Recognizing],+[GorelickICCV2005Action],[1][LeCVPR01Learning],[2][SubhabrataCVPR2011Probabilistic],[3][WangCVPR2011Action],[4][WangCVPR2011ActionST],[5][LiCVPR2011Activity],[6][XieCvpr11Unified],[7][ChenCVPR11Modeling],[8][SunICCV2011Action],[9][HoaiCVPR2011Joint],[10[BrendelICCV2011Learning]

Movie/Web Benchmarks

- Multiple camera viewing angles
- Camera motions
- Video qualities/ resolutions and clutters vary
- Multiple moving objects

HMDB51 Dataset*

- Large set (51) of activity categories
- High intra-category variations
- Drastic appearance, scale, position changes of actors. Variations in camera motion and viewpoints

*[KuehneICCV2011HMDB]

HMDB51 sample frames

Reported accuracy: 38.00% in [SadanandCVPR2012Action]

Movie/Web Benchmarks

Other datasets and reported results

	Classes	Clips	Resolution	Accuracy
UCF Youtube[1]	11	3185	240*320	84.2%[9]
UCF 50[3]	50	>5000	240*320	76.4%[12]
UCF Sports[2]	9	182	480*720	95.0%[12]
Coffee/Cigarette[4]	2	264	240*500	57%[8]
Hollywood1[5]	8	400	240*500	56.8%[11]
Hollywood2[6]	12	1707	240*500	58.3%[9]
Olympics[7]	16	800	360*450	77.3%[10]
TRECVID MED[14]	15	32061	vary	38%~54% [14]

[1][LiuCVPR09Recognizing],[2][RodriguezCVPR2008Action],[3][UCF50],
[4][LaptevICCV07Retrieve],[5][LaptevCVPR2008Learning],[6][MarszalekCVPR2009Actions]
,[7][NieblesECCV2010Modeling],[8][GaidonCVPR2011Actom],[9][WangCVPR2011Action],[10]
[BrendelICCV2011Learning], [11][GilbertPAMI2010Action],[12][SadanandCVPR2012Action
], [13] [TangCVPR2012Learning], [14][TRECVID2011]

Benchmarks for Assisted Daily Living (ADL) Activities

ADL65 Dataset*

- Large set (65 categories) of high-resolution kitchen activities
- Fine-grained activities (low inter-class variability)
- Detailed annotations including time intervals and poses
- Provide classification & detection tasks
- Similar dataset: URADL+
  - high-res 10 kitchen activities.
  - reported accuracy: 96% in [WangCVPR2011ActionST]
- Scene and object info highly correlates with the activities

sample frames

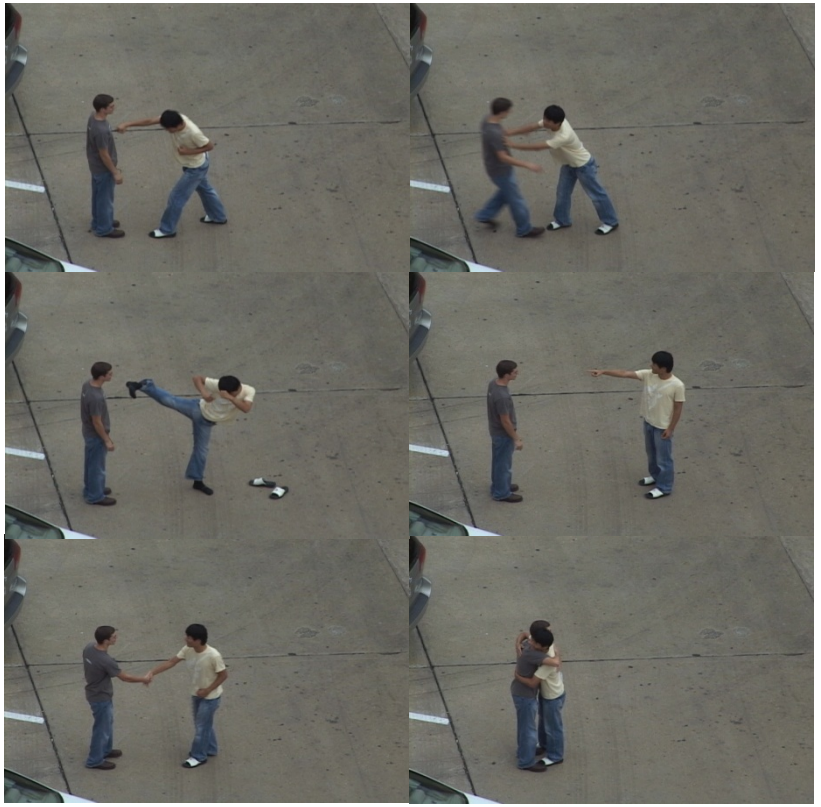
Reported average precision for
Classification: 59.2%
Detection: 45.0%
in [RohrbachCVPR2012Database]

*[RohrbachCVPR2012Database]
+[MessingICCV2009Activity]

Benchmarks for Group Activities

UT-Interaction*

- 6 classes of 2-person interaction activities
- Detailed annotation with time intervals/bounding boxes
- Camera jittering
- Pedestrians in the background
- Concurrent activities
- Similar datasets: Collective Dataset[1], BEHAVE[2]



UT-Interaction sample frames

	Accuracy
AmerICCV2011Chain	75.75%
BrendelICCV2011Learning	78%
RyooICCV2011Early	85%
GaurICCV2011SFG	72%

*[RyooICPR2010Overview]
[1][ChoiICCV2009Collective]
[2][BlunsdenBMVA2010Behave]

Benchmarks for Long Term Surveillance

Virat Ground Video Dataset*

- Realistic scenarios (non-actors)
- Multi spatial-temporal resolutions
- Diverse scenes (16 scenes) and event types (23)
- Multiple objects and concurrent activities
- Different camera perspectives
- Detailed annotations including time intervals, bounding boxes, and tracks
- People/facility, people/vehicle interaction

sample frames

sample annotations

* [OhCVPR2011Virat]

Benchmarks for Long Term Surveillance

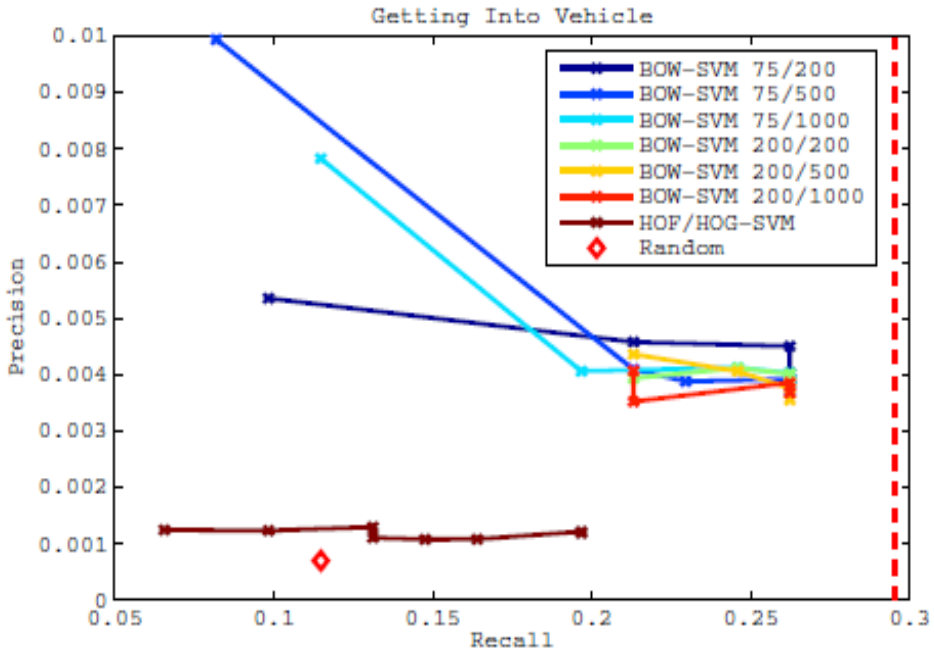
Virat Aerial Video Dataset

- Camera motion
- Low resolution of human figures
- Similarity across actions from high altitude
- Time-varying viewpoints and scales
- Shadows and interrupted tracking



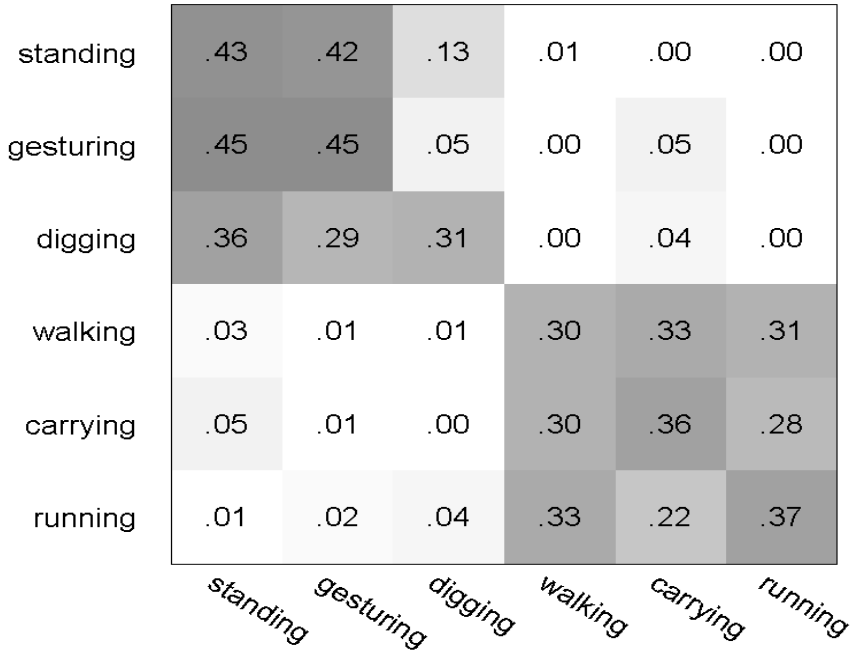
sample frames for different scenes and viewpoints

Benchmarks for Long Term Surveillance Evaluations

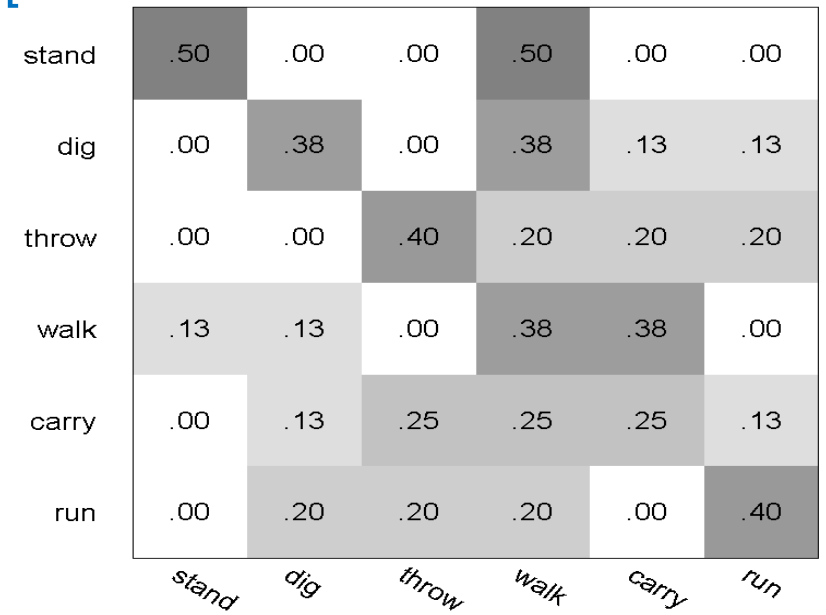


[OhCVPR2011Virat]

Average hit rate on ground dataset:
33% in [OhCVPR2011Virat]
Average accuracy on aerial dataset:
38% in [ChenCVPR11Modeling]



[SubhabrataCVPR2011Probabilistic]



[ChenCVPR11Modeling]

Benchmarks for Long Term Surveillance

TRECVID 2011 Surveillance Event Detection (SED) Dataset

- ~100 hours indoor airport surveillance videos
- 7 events including 2 single person events, 2 person-object interaction events and 3 multi-person events
- Different background clutters and traffic due to different camera placements
- Best Normalized Detection Cost Rate (NDCR): 0.8~2

$$NDCR = P_{miss} + \frac{Rate_{FA}}{Rate_{event}}$$

Perfect NDCR: 0




camera placements: controlled access door, waiting area, debarkation area, elevator door, transit area

Multi-Camera Benchmarks

IXMAS Dataset*

- Provide 5-view videos of 13 kinematic activities
- Provide silhouette, reconstructed volumes, and calibration information
- 3D information is available
- Suitable to evaluate view dependent models

Wave

Pick up

sample frames

*[WeinlandICCV07Action]

Multi-Camera Benchmarks

IXMAS Dataset - Evaluations

- Accuracy by using multi-cameras*

1,2,3,4,5	1,2,3,4	1,2,3,5	1,2,3	1,3,5	1,3	2,4	3,5
88.20%	88.20%	89.40%	87.70%	88.40%	86.60%	82.40%	83.30%

Recognition using multi-view information (5~15% improvements over single view)

- View transfer evaluations+

Evaluating cross-view model effectiveness

%	1	2	3	4	5
1		81.8	88.1	87.5	81.4
2	87.5		82	92.3	74.2
3	85.3	82.6		82.6	76.5
4	82.1	81.5	80.2		70
5	78.8	73.8	77.7	78.7	

+[LiCVPR2012Discriminative]

*[WuCVPR2011Action]

RGB-D Benchmarks

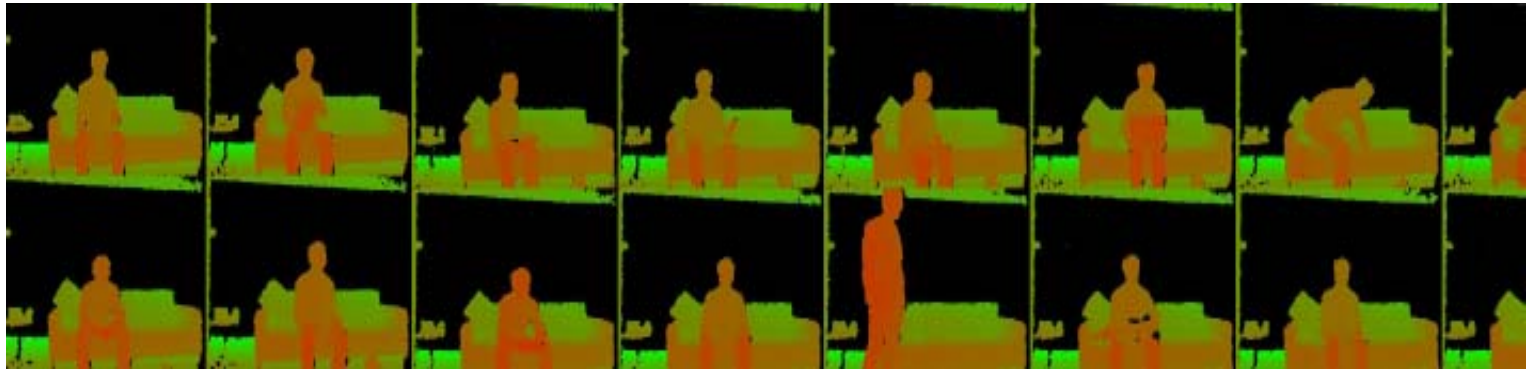
MSRAction3D and DailyActivity3D*

- Both recorded with commercially available depth sensor
- MSRAction3D consists of 20 kinematic activities
- DailyActiviy3D consists of 16 living room activities involving different objects e.g. writing on a paper, answering phone

Accuracy: 88.2% *

MSRAction3D sample frames

Accuracy: 85.7%*



DailyActivity3D sample frames

*[WangCVPR2012Mining]

Egocentric Benchmarks

Egocentric ADL Datasets[1]

- Complex object interactions (42 objects)
- Large set of actions (18 actions), and sites (20 homes)
- Longer activities
- Large variations of object appearance
- Scene/Clutter variations
- Similar datasets: GTEA[2] (GeorgiaTech Egocentric) and Intel Egocentric Dataset[3]
- Strong priors for hand locations

appearance variations

scene variations

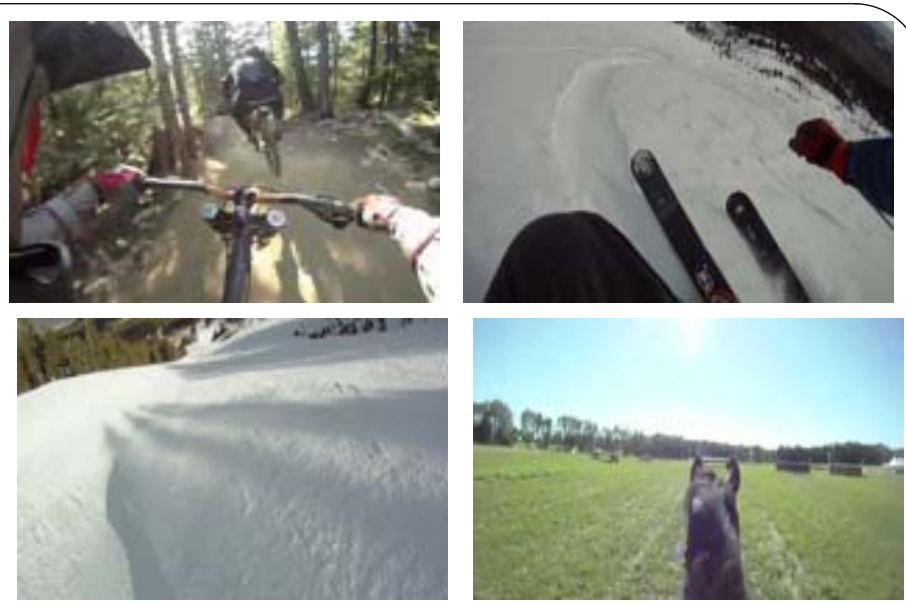
Reported classification accuracy: 40.6% in [PirsiavashCVPR2012 Detecting]

[1][PirsiavashCVPR2012Detecting]
[2][FathiICCV2011Understanding]
[3][RenEgo2009Ego]

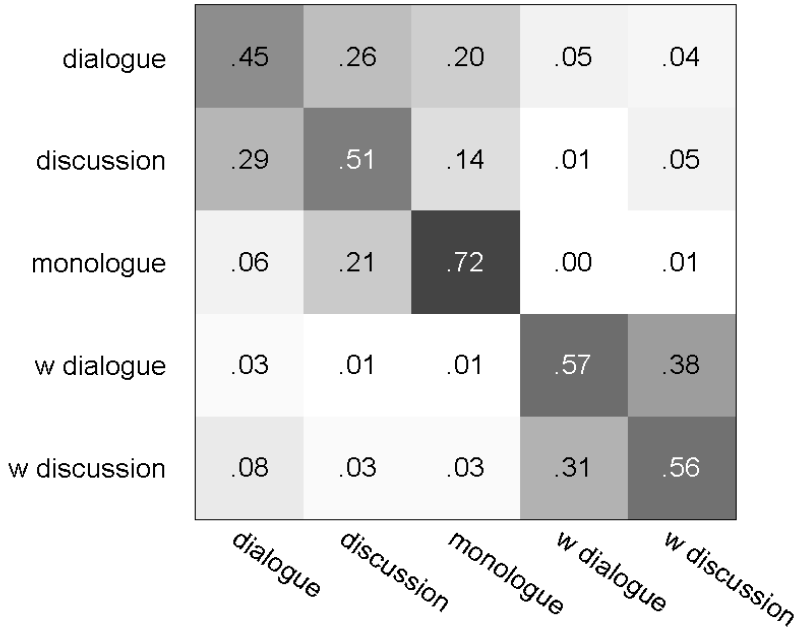
Egocentric Benchmarks

Other Egocentric Datasets

- UEC Sports[1]
  - Large set of outdoor activities
  - Large motion and blur
- GeorgiaTech[2] –
  - Egocentric social activity recognition
- Other datasets: Egocentric novelty detection[3], and egocentric video summary[4]



UEC Dataset



FPSI Result in [FathiCVPR2012Social]

[1][KitaniCVPR2011Fast],
[2][FathiCVPR2012Social],
[3][AghazadehCVPR2011Novelty],
[4][LeeCVPR2012Discovering]

Summary Table

	classes	res.	seqs	frames	url	annotation
KTH	6	160*120	600	~500k	http://www.nada.kth.se/cvap/actions/	N.A
Weizmann	10	180*144	90	~8k	http://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html	extracted foreground mask
HMDB51	51	~240*480	7000	~800K	http://serre-lab.clps.brown.edu/resources/HMDB/	meta info regarding video quality, angle, and camera motion
UCF Youtube	11	240*320	1168	~80K	http://www.cs.ucf.edu/vision/public_html/data.html#UCF%20YouTube%20Action%20Dataset	N.A
UCF Sports	9	480*720	182	~6K	http://www.cs.ucf.edu/vision/public_html/data.html#UCF%20Sports%20Action%20Dataset	N.A
UCF 50	50	240*320	>500	>150K	http://www.cs.ucf.edu/vision/public_html/data.html#UCF50	N.A
Coffee/Cigarette	2	240*500	1	36K	http://www.di.ens.fr/~laptev/download.html	Space-time cuboid, key frame and head position
Hollywood1	8	240*500	~600	~400K	http://www.di.ens.fr/~laptev/download.html	Time interval
Hollywood2	12	240*500	~600	~600K	http://www.di.ens.fr/~laptev/download.html	Time interval

Summary Table

	classes	res.	seqs	frames	url	annotation
MED11	10	N.A	10K	35M	http://www-nlpir.nist.gov/projects/tv2011	event intervals
ADL65	65	1624x1224	44	~881K	N.A	time-interval and body pose
URADL	10	1280x720	50	~75K	http://www.cs.rochester.edu/~rmessing/uradl/	N.A
Ut-interaction	6	480*720	20	~36K	http://cvrc.ece.utexas.edu/SDHA2010/Human Interaction.html	time-interval and bounding boxes of subjects
Collective	5	480*720	44	~13K	http://www.eecs.umich.edu/vision/activity-dataset.html	locations of the subjects, bounding box, and pose info
BEHAVE	10	480*640	4	~300K	http://groups.inf.ed.ac.uk/vision/BEHAVEDATA/INTERACTIONS/	Interval, group id, bounding box
IXMAS	13	390x291	36	~40K*	http://4drepository.inrialpes.fr/	silhouettes, reconstructed volume
VIRAT	23	1920 x 1080	N.A	~1620K	http://www.viratdata.org/	object tracks, subject bounding boxes, event interval
SED11	7	720x576	N.A	~9M	http://www-nlpir.nist.gov/projects/tv2011	event intervals, bounding boxes

Summary Table

	classes	res.	seqs	frames	url	annotation
MSRAction3D	20	320*240	320	9.6K	http://research.microsoft.com/en-us/um/people/zliu/actionrecorsrc/default.htm	joints
DailyAvtivity 3D	16	320*240	567	~15K	http://research.microsoft.com/en-us/um/people/zliu/actionrecorsrc/default.htm	joints
EgoADL	18	N.A	N.A	1M	N.A	object bounding box, tracks and labels, hand positions

Conclusions and Challenges

- We covered most of the major benchmarking datasets – good starting points for people new to the field
- Creating a benchmark that captures the level of complexity of real world problems is still hard
- Few reported cross dataset performance or model generality evaluation
- The trend is to create large scale benchmarking datasets with detailed annotations - need better tools (e.g. LabelMe Video) or technology (e.g. Machine-aided Mechanical Turk) for good quality and large annotation tasks

Bibliography

[AghazadehCVPR2011Novelty] Omid Aghazadeh, Josephine Sullivan, and S. Carlsson, "Novelty Detection from an Ego-Centric Perspective," in CVPR, 2011.

[AmerICCV2011Chain] Mohamed Amer and S. Todorovic, "A Chains Model for Localizing Participants of Group Activities in Videos," in ICCV, 2011.

[BrendelICCV2011Learning] William Brendel and S. Todorovic, "Learning Spatiotemporal Graphs of Human Activities," in ICCV, 2011.

[BlunsdenBMVA2010Behave] S Blunsden and R. B. Fisher, "The BEHAVE video dataset: ground truthed video for multi-person behavior classification," *Annals of the BMVA,* pp. 1-11, 2010.

[ChenCVPR11Modeling] Chia-Chih Chen and J. K. Aggarwal, "Modeling Human Activities as Speech," in CVPR, 2011.

[ChoiICCV2009Collective] W. Choi, K. Shahid, and S. Savarese., "What are they doing?Collective activity classification using spatio-temporal relationship among people," in ICCV, 2009.

[FathiICCV2011Understanding] Alireza Fathi, Ali Farhadi, and J. Rehg, "Understanding Egocentric Activities," in ICCV, 2011.

[FathiCVPR2012Social] Alireza Fathi, Jessica Hodgins, and J. Rehg, "Social Interactions: A First-Person Perspective," in CVPR, 2012.

[GaurICCV2011SFG] U. Gaur, Y. Zhu, B. Song, and A. Roy-Chowdhury, "A “String of Feature Graphs” Model for Recognition of Complex Activities in Natural Videos," in ICCV, 2011.

[GaidonCVPR2011Actom] Adrien Gaidon, Zaid Harchaoui, and C. Schmid, "Actom Sequence Models for Efficient Action Detection," in CVPR, 2011.

Bibliography (continued)

[GilbertPAMI2010Action] Andrew Gilbert, John Illingworth, and R. Bowden, "Action Recognition using Mined Hierarchical Compound Features," TPAMI, 2010.

[HoaiCVPR2011Joint] Minh Hoai, Zhen-Zhong Lan, and F. D. l. Torre, "Joint Segmentation and Classification of Human Actions in Video," in CVPR, 2011.

[KitaniCVPR2011Fast] Kris Kitani, Takahiro Okabe, Yoichi Sato, and A. Sugimoto, "Fast Unsupervised Ego-Action Learning for First-Person Sports Videos," in CVPR, 2011.

[KuehneICCV2011HMDB] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: A Large Video Database for Human Motion Recognition," in ICCV, 2011.

[LaptevICCV07Retrieve] Ivan Laptev and Patrick Perez, "Retrieving actions in movies," in ICCV, 2007.

[LaptevCVPR2008Learning] Ivan Laptev, Marcin Marszałek, Cordelia Schmid, and Benjamin Rozenfeld, "Learning realistic human actions from movies," in CVPR, 2008.

[LeeCVPR2012Discovering] Yong Jae Lee, Joydeep Ghosh, and K. Grauman, "Discovering Important People and Objects for Egocentric Video Summarization," in CVPR, 2012.

[LeCVPR2011Learning] Quoc Le, Will Zhou, and A. Ng, "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis," in CVPR, 2011.

Bibliography (continued)

[LiCVPR2012Discriminative] Ruonan Li and T. Zickler, "Discriminative Virtual Views for Cross-View Action Recognition," in CVPR, 2012.

[LiCVPR2011Activity] Binlong Li, Mustafa Ayazoglu, Teresa Mao, Octavia Camps, and M. Sznaier, "Activity Recognition using Dynamic Subspace Angles," in CVPR, 2011.

[LiuCVPR09Recognizing] Jingen Liu, Jiebo Luo, and Mubarak Shah, "Recognizing realistic actions from videos in the wild," in CVPR, 2009.

[MarszalekCVPR2009Actions] Marcin Marszałek, Ivan Laptev, and Cordelia Schmid, "Actions in Context," in CVPR, 2009.

[MessingICCV2009Activity] Ross Messing, Chris Pal, and Henry Kautz, "Activity recognition using the velocity histories of tracked keypoints," in ICCV, 2009.

[NieblesECCV2010Modeling] Juan Niebles, Chih-Wei Chen, and F.-F. Li, "Modeling Temporal Structure of Decomposable Motion Segments for Activity Classification," in ECCV 2010.

[OhCVPR2011Virat] Sangmin Oh and e. al., "A Large-scale Benchmark Dataset for Event Recognition in Surveillance Video," in CVPR, 2011.

[PirsiavashCVPR2012Detecting] Hamed Pirsiavash and D. Ramanan, "Detecting Activities of Daily Living in First-person Camera Views," in CVPR, 2012.

Bibliography (continued)

[RenEgo2009Ego] Xiaofeng Ren and M. Philipose, "Egocentric Recognition of Handled Objects: Benchmark and Analysis. ," in Egovision Workshop, 2009.

[RohrbachCVPR2012Database] Marcus Rohrbach, Sikandar Amin, Mykhaylo Andriluka, and B. Schiele, "A Database for Fine Grained Activity Detection of Cooking Activities," in CVPR, 2012.

[RodriguezCVPR2008Action] Mikel Rodriguez, Javed Ahmed, and Mubarak Shah, "Action MACH: a spatio-temporal maximum average correlation height filter for action recognition," in CVPR, 2008.

[RyooICPR2010Overview] M. S. Ryoo, Chia-Chih Chen, J. K. Aggarwal, and A. Roy-Chowdhury, "An Overview of Contest on Semantic Description of Human Activities," in ICPR, 2010.

[RyooICCV2011Early] M. S. Ryoo, "Human Activity Prediction:Early Recognition of Ongoing Activities from Streaming Videos," in ICCV, 2011

[SadanandCVPR2012Action] Sreemanananth Sadanand and J. Corso, "Action Bank: A High-Level Representation of Activity in Video," in CVPR, 2012.

[SubhabrataCVPR2011Probabilistic] Subhabrata Bhattacharya, R. Sukthankar, R. Jin, and M. Shah, "A Probabilistic Representation for Efficient Large Scale Visual Recognition Tasks," in CVPR, 2011.

Bibliography (continued)

[SunICCV2011Action] Chuan Sun, Imran Junejo, and H. Foroosh, "Action Recognition using Rank-1 Approximation of Joint Self-Similarity Volume," in ICCV, 2011.

[TangCVPR2012Learning] Kevin Tang, Li Fei-Fei, and D. Koller, "Learning Latent Temporal Structure for Complex Event Detection," in CVPR, 2012.

[TRECVID2011] P. Over, "An Overview of the Goals, Tasks, Data, Evaluation Mechanisms, and Metrics," in TRECVID, 2011.

[UCF50] http://www.cs.ucf.edu/vision/public_html/data.html

[WangCVPR2011Action] Heng Wang, Alexander Klaser, Cordelia Schmid, and C.-L. Liu, "Action Recognition by Dense Trajectories," in CVPR, 2011.

[WangCVPR2011ActionST] Jiang Wang, Zhuoyuan Chen, and Y. Wu, "Action Recognition with Multiscale Spatio-Temporal Contexts," in CVPR, 2011.

[WangCVPR2012Mining] Jiang Wang, Zicheng Liu, Ying Wu, and J. Yuan, "Mining Actionlet Ensemble for Action Recognition with Depth Cameras," in CVPR, 2012.

[WeinlandICCV07Action] Daniel Weinland, Edmond Boyer, and Remi Ronfard, "Action Recognition from Arbitrary Views using 3D Exemplars," in ICCV, 2007.

[WuCVPR2011Action] Xinxiao Wu, Dong Xu, Lixin Duan, and J. Luo, "Action Recognition using Context and Appearance Distribution Features," in CVPR, 2011.

Bibliography (continued)

[WuICCV2011Action] Shandong Wu, Omar Oreifej, and M. Shah, "Action Recognition in Videos Acquired by a Moving Camera Using Motion Decomposition of Lagrangian Particle Trajectories," in ICCV, 2011.

[XieCvpr2011Unified] Yuelei Xie, Hong Chang, Zhe Li, Luhong Liang, Xilin Chen, and D. Zhao, "A Unified Framework for Locating and Recognizing Human Actions," in CVPR, 2011.