Exploiting 3D Information for Robust Real-Time Tracking of Multiple Objects in Complex Scenarios

MSc Defense

Horst Possegger

Supervisor

Prof. Horst Bischof

Advisors

Sabine Sternig, Thomas Mauthner, and Peter M. Roth

Monocular Multiple Object Tracking

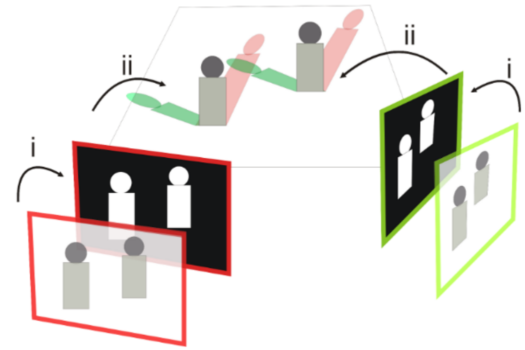
- Location estimates
  - Background modelling [1,2]
  - Appearance information [3,4]
  - Part-based detectors [5]

- Drawbacks
  - Occlusions
  - Lack of 3D information

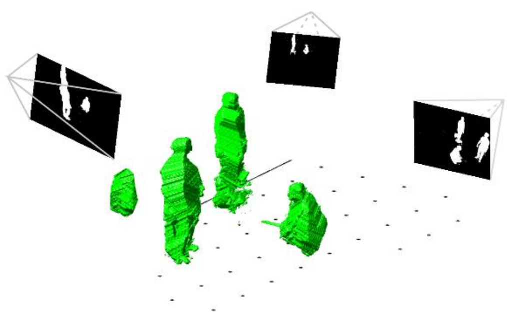
[1] Haritaoglu, Harwood, and Davis. *W4: Who? When? Where? What? A Real Time System for Detecting and Tracking People*. In Proc. AFGR, 1998.
[2] Intille and Bobick. *Visual Tracking Using Closed-Worlds*. In Proc. ICCV, 1995.
[3] Comaniciu, Ramesh, and Meer. *Kernel-Based Object Tracking*. PAMI, 2003.
[4] Seo, Choi, Kim, and Hong. *Where are the ball and players? Soccer Game Analysis with Color-based Tracking and Image Mosaick*. In Proc. ICIAP, 1997.
[5] Wu and Nevatia. *Detection and Tracking of Multiple, Partially Occluded Humans by Bayesian Combination of Edgelet based Part Detectors*. IJCV, 2007.

Multiple Views

- Planar projections [1,2]
  - Fuse 2D information
  - Common ground-plane
  - Projection artifacts



- Volumetric reconstructions [3,4]
  - No planarity assumption
  - Valuable tracking cue



[1] Khan and Shah. *Tracking Multiple Occluding People by Localizing on Multiple Scene Planes*. PAMI, 2009.
[2] Eshel and Moses. *Tracking in a Dense Crowd Using Multiple Cameras*. IJCV, 2010.
[3] Guan, Franco, and Pollefeys. *Multi-Object Shape Estimation and Tracking from Silhouette Cues*. In Proc. CVPR, 2008.
[4] Liem and Gavrila. *Multi-person tracking with overlapping cameras in complex, dynamic environments*. In Proc. BMVC, 2009.

Overview

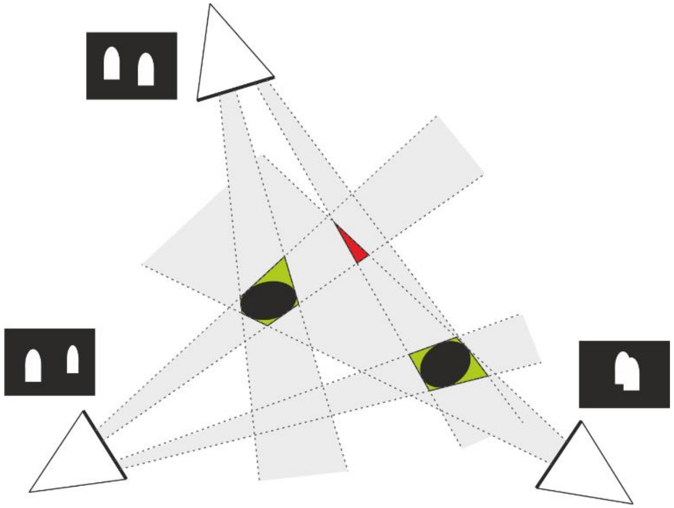
From visual hull reconstruction using foreground segmentations of calibrated cameras...

[Video]

Outline

- Volumetric mass density
- Resolving geometric ambiguities
- Experiments
- Conclusion

3D Reconstruction



- Constraints
  - Low number of views
  - Wide baselines
  - Different viewing angles

- Shape from Silhouette [1,2]
  - Volumetric reconstruction
  - Binary foreground segmentation (*e.g.*, [3])
  - Intersection of viewing cones - *visual hull* [4]
  - Noise sensitivity

[1] Martin and Aggarwal. *Volumetric Descriptions of Objects from Multiple Views*. PAMI, 1983.
[2] Baumgart. *Geometric Modeling for Computer Vision*. PhD thesis, Stanford University, CS Department, 1974.
[3] McFarlane and Schofield. *Segmentation and tracking of piglets in images*. MVA, 1995.
[4] Laurentini. *The Visual Hull Concept for Silhouette-Based Image Understanding*. PAMI, 1994.

Local Mass Densities

- Visual hull representation

$$v_i \in \mathcal{V} = \begin{cases} 1 & \text{if } v_i \text{ foreground} \\ 0 & \text{otherwise} \end{cases}$$

- Occupancy volume

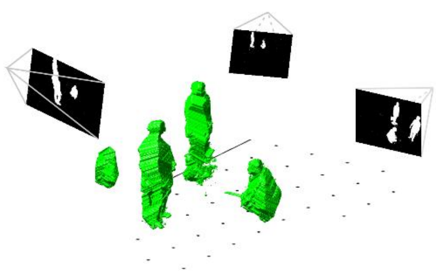
$$\mathcal{V}_{\mathrm{O}} = \{m_D(v_i) \,|\, \forall v_i \in \mathcal{V}\}\,, \qquad m_D(v_i) = \frac{\sum_{v_j \in N_{v_i}} v_j}{|N_{v_i}|}$$

- Local neighborhood
  - Depends on object class
  - People: upright aligned torso

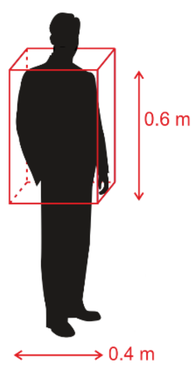
Neighborhood Formulation

- Upright aligned torso
- Approximation
  - Axis-aligned cuboid
  - Efficiency by integral structures

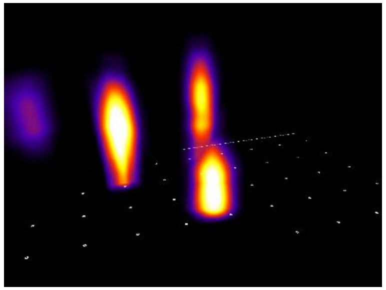
$$N_{v_i} = \left\{ v_j \;\middle|\; |v_{j,x} - v_{i,x}| \leq r \wedge |v_{j,y} - v_{i,y}| \leq r \wedge |v_{j,z} - v_{i,z}| \leq \frac{h}{2} \right\}$$



Visual hull.



Approximation.

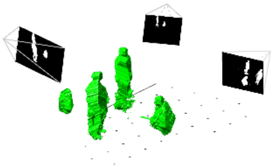


Occupancy volume.

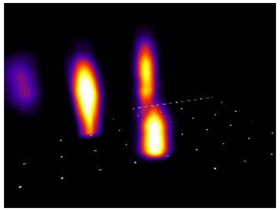
Tracking from Local Mass Densities

- Localization
  - (2+1)D tracking
  - Estimate xy-coordinates
  - Find vertical mass center
- Top-view occupancy map
  - Maximum along vertical axis
  - Indicates location of objects‘ mass center

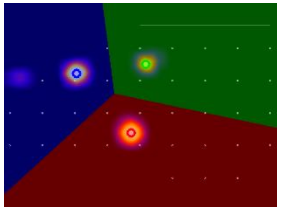
Input image.



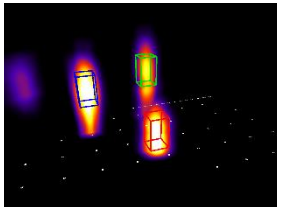
Visual hull.



Occupancy volume.



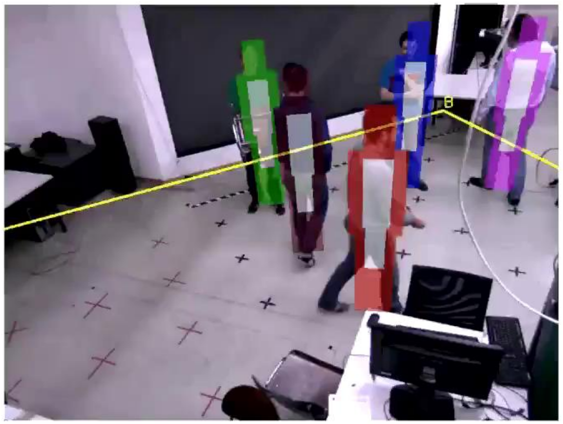
Occupancy map.

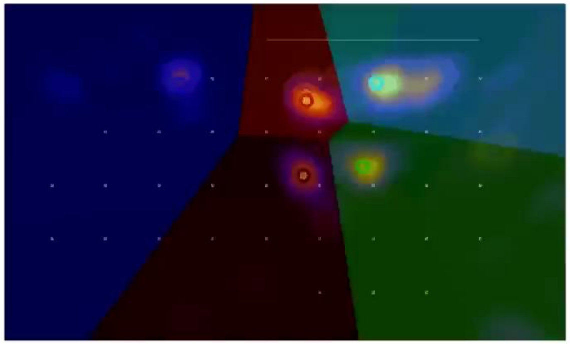


Vertical mass center.

Estimating *xy*-coordinates

- Tracking step
  - Bootstrap particle filter [1]
  - Second order auto-regressive transition model [2]
- Multiple objects
  - Individual particle filters
  - Voronoi tesselation
  - Restrict particle transition by Voronoi cell
  - Inspired by [3]





[Video]

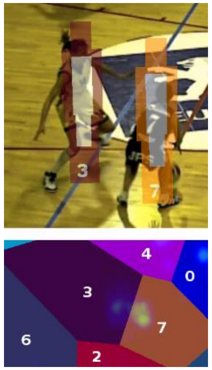
[1] Isard and Blake. *Condensation – Conditional Density Propagation for Visual Tracking*. IJCV, 1998.
[2] Pérez, Hue, Vermaak, and Gangnet. *Color-Based Probabilistic Tracking*. In Proc. ECCV, 2002.
[3] Kristan, Perš, Perše, and Kovačič. *Closed-world tracking of multiple interacting targets for indoor-sports applications*. CVIU, 2009.

Resolving Geometric Ambiguities

- Appearance information
  - Exploit 3D knowledge
  - Extract valuable features
  - Hue-saturation histograms

- Feature bag
  - For each camera c
  - FIFO update strategy

$$\mathcal{F}_i = \left\{ \mathcal{F}_i^{(c)} \right\}_{c=1}^{N_C}, \quad \mathcal{F}_i^{(c)} = \left\{ \mathbf{f}_l^{(c)} \right\}_{l=1}^{N_F}$$



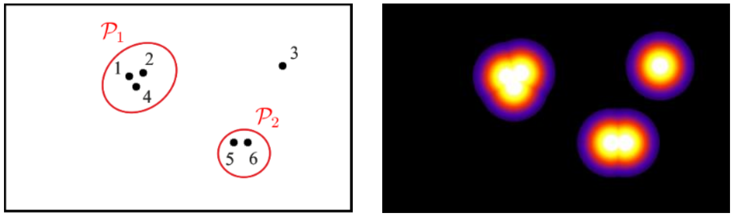
t = 1046.

t = 1051.

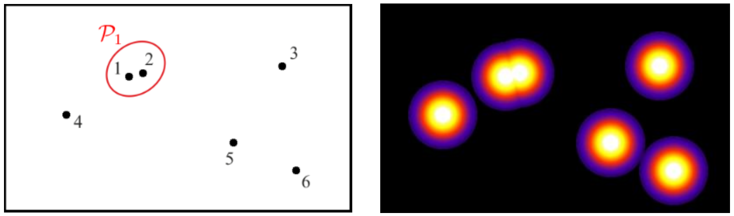
t = 1054.

t = 1060.

Merge-Split Approach



Merge step.



Split step.

- Build *conflict pools* $\mathcal{P}_m$
- Train one-vs-all logistic regression classifiers [1] on-demand

$$p_i\left(y_{i,l} \,|\, \mathbf{f}_l, \mathbf{w}_i\right) = \frac{1}{1 + e^{-y_{i,l}\mathbf{w}_i^\top \mathbf{f}_l}}, \quad y_{i,l} = \begin{cases} +1 & \forall \mathbf{f}_l \in \mathcal{F}_i \\ -1 & \forall \mathbf{f}_l \in \mathcal{F}_j : \forall j \in \mathcal{P}_m, j \neq i \end{cases}$$

- Detect seperate local maxima (NMS [2])

$$\hat{i} = \arg\max_i \; p_i\left(y_{i,\mathrm{NMS}} = +1 \,|\, \mathbf{f}_{\mathrm{NMS}}, \mathbf{w}_i\right)$$

[1] Fan, Chang, Hsieh, Wang, and Lin. *LIBLINEAR: A Library for Large Linear Classification*. JMLR, 2008.
[2] Neubeck and Van Gool. *Efficient Non-Maximum Suppression*. In Proc. ICPR, 2006.

Experiments

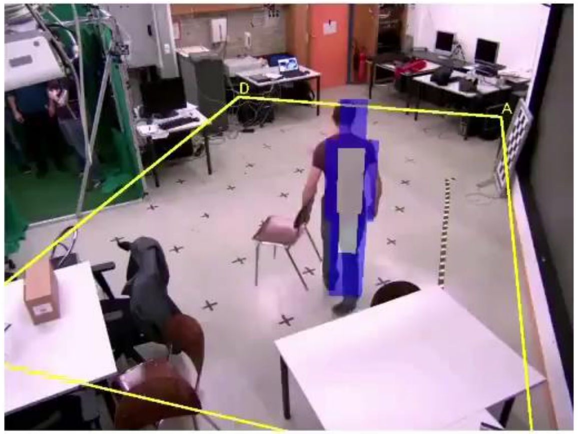
- Datasets

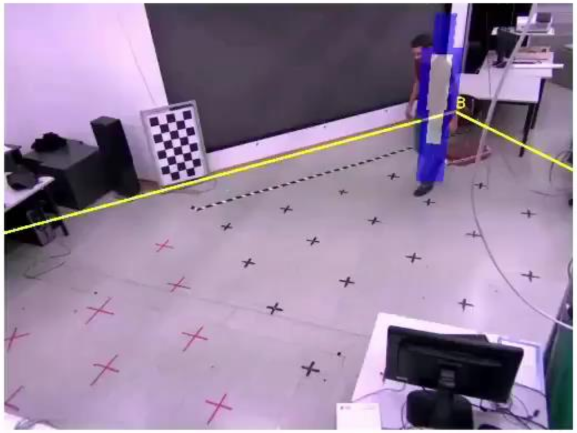
Dataset	Cameras	Objects	Frames	Resolution	Tracking Region
Changing Appearance (CHAP)	4	5	3760	1024×768	$7\,m \times 4\,m$
Leapfrogs (LEAF 1)	4	4	1800	1024×768	$7\,m \times 4\,m$
Leapfrogs (LEAF 2)	4	5	2400	1024×768	$7\,m \times 4\,m$
Musical Chairs (MUCH)	4	5	2400	1024×768	$7\,m \times 4\,m$
POSE	4	6	1820	1024×768	$7\,m \times 4\,m$
TABLE	4	5	1760	1024×768	$7\,m \times 4\,m$
APIDIS Basketball [1]	7	12	1500	1600×1200	$15\,m \times 15\,m$

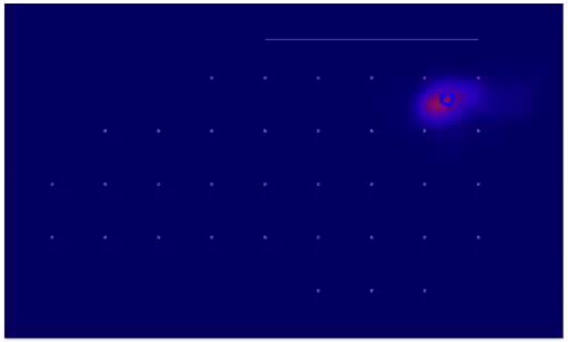
- Challenges
  - Spatial proximity / crowds / occlusions
  - Different poses
  - Out-of-plane motion
  - Similar / changing appearance

[1] Autonomous Production of Images based on Distributed and Intelligent Sensing (APIDIS), http://www.apidis.org/Dataset

Different Poses

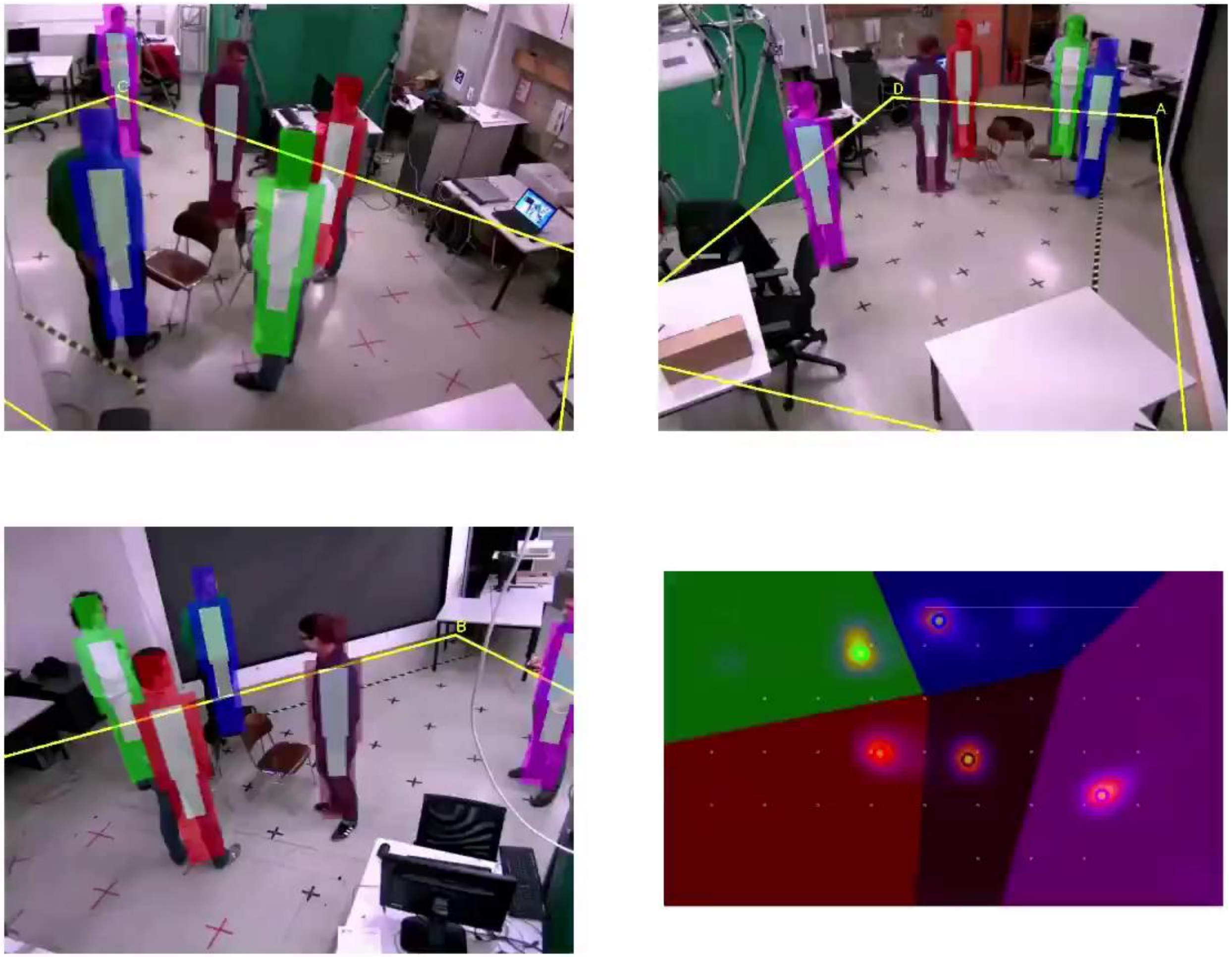






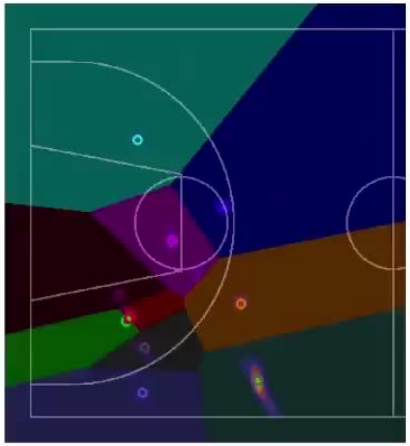
[Video]

Musical Chairs



[Video]

APIDIS Basketball



[Video]

Evaluation

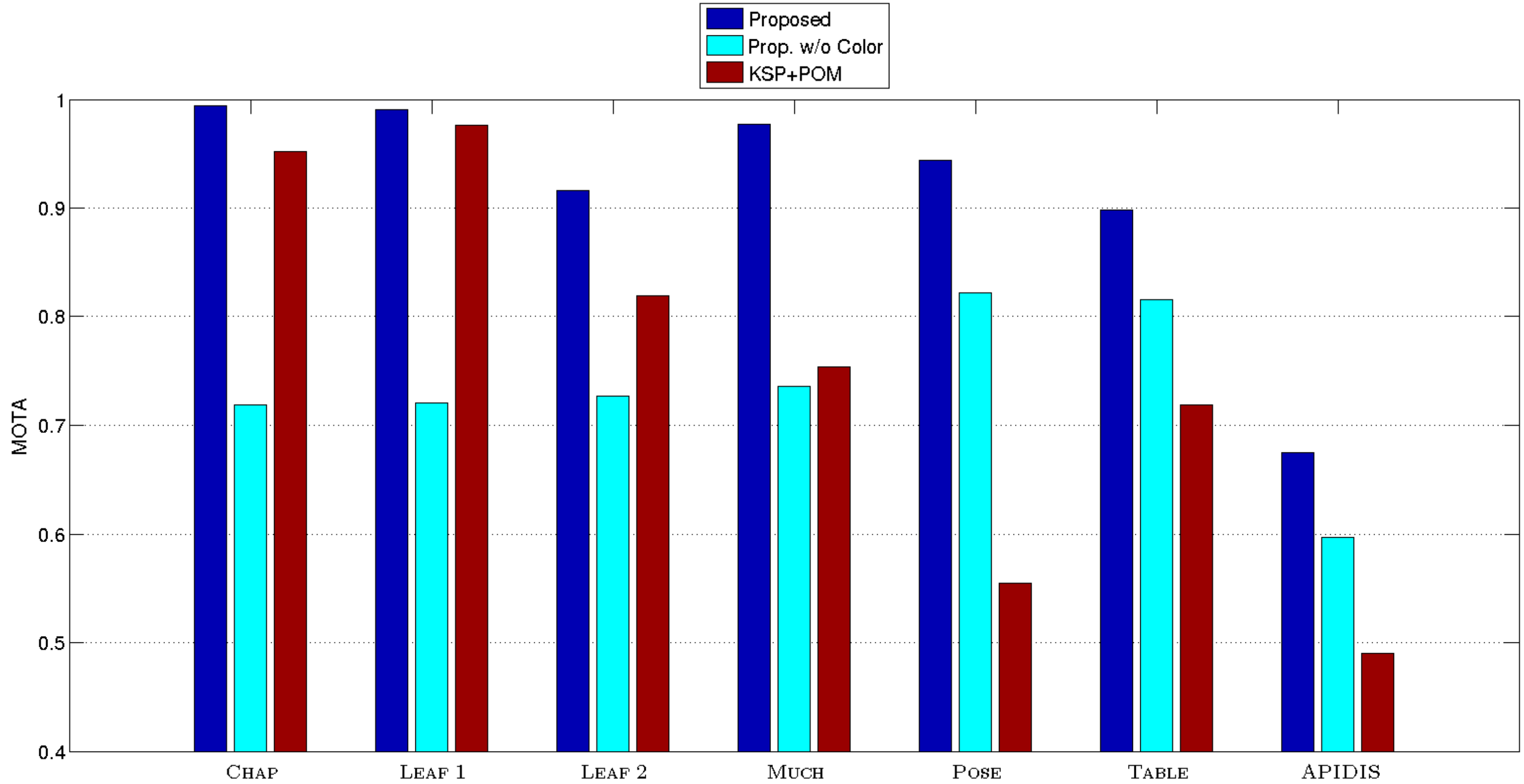
- CLEAR performance metrics [1]
  - Multiple object tracking accuracy (MOTA)
    - Object configuration errors
    - False positives, false negatives, identity switches
  - Multiple object tracking precision (MOTP)
    - Alignment of true positive trajectories *w.r.t.* ground truth
    - Ground-plane distance
- Comparison
  - KSP tracker [2]
  - POM detections [3]
  - Same input data

[1] Bernardin and Stiefelhagen. *Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics*. EURASIP JIVP, 2008.
[2] Berclaz, Fleuret, Türetken, and Fua. *Multiple Object Tracking using K-Shortest Paths Optimization*. PAMI, 2011.
[3] Fleuret, Berclaz, Lengagne, and Fua. *Multi-Camera People Tracking with a Probabilistic Occupancy Map*. PAMI, 2008

Tracking Performance

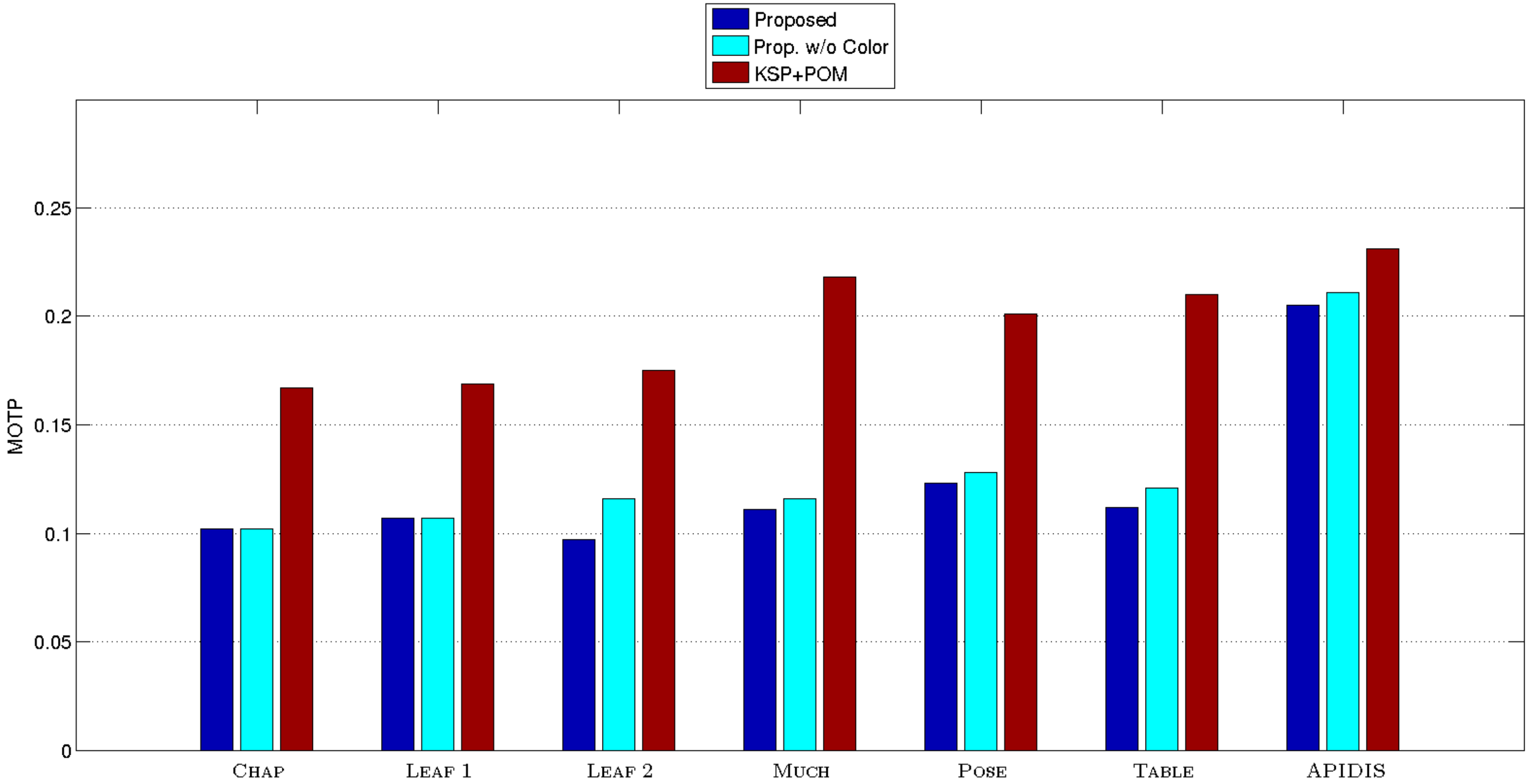
Dataset	Algorithm	MOTP [m]	MOTA	TP	FP	FN	IDS
CHAP	Proposed	**0.102**	**0.994**	1555	**2**	**6**	**1**
	Prop. w/o Color	**0.102**	0.719	1316	193	241	4
	KSP+POM	0.167	0.952	**1607**	50	21	7
LEAF 1	Proposed	**0.107**	**0.991**	464	**2**	2	**0**
	Prop. w/o Color	**0.107**	0.721	436	83	44	7
	KSP+POM	0.169	0.976	**495**	6	**1**	5
LEAF 2	Proposed	**0.097**	**0.916**	**930**	**41**	**41**	**0**
	Prop. w/o Color	0.116	0.727	856	115	117	34
	KSP+POM	0.175	0.819	913	87	66	24
MUCH	Proposed	**0.111**	**0.977**	**783**	**9**	**9**	**0**
	Prop. w/o Color	0.116	0.736	694	99	99	11
	KSP+POM	0.218	0.754	770	139	32	26
POSE	Proposed	**0.123**	**0.944**	**485**	**14**	**14**	**0**
	Prop. w/o Color	0.128	0.822	456	42	44	3
	KSP+POM	0.201	0.555	427	156	31	17
TABLE	Proposed	**0.112**	**0.898**	**621**	**32**	**28**	**6**
	Prop. w/o Color	0.121	0.816	596	57	55	8
	KSP+POM	0.210	0.719	573	105	58	14
APIDIS	Proposed	**0.205**	**0.675**	**656**	**88**	**172**	**9**
	Prop. w/o Color	0.211	0.597	625	121	202	10
	KSP+POM	0.231	0.490	607	156	220	46

Results - Accuracy



Higher is better.

Results - Precision



Measured in meters, lower is better.

Runtime Performance

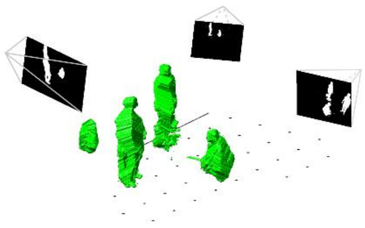
	Proposed	Prop. w/o Color	KSP	POM	KSP +POM
CHAP	9.89	12.67	43.49	0.02	0.02
LEAF 1	9.88	10.34	63.84	0.04	0.04
LEAF 2	7.65	9.04	229.77	0.05	0.05
MUCH	12.08	13.21	185.28	0.06	0.06
POSE	10.27	12.99	132.49	0.05	0.05
TABLE	8.03	9.60	208.51	0.07	0.07
APIDIS	4.42	6.16	80.70	0.03	0.03

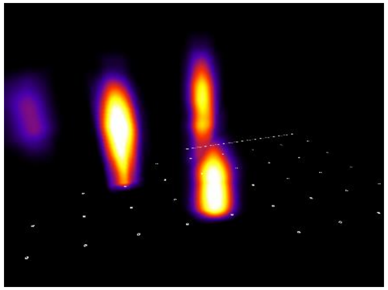
Average frame rate.

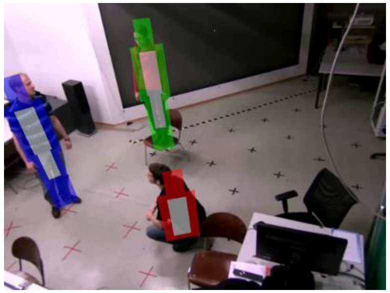
Conclusion

- Multiple object tracking
  - Calibrated camera network
  - Volumetric reconstruction

- Volumetric mass densities
  - Valuable cue for multi-object tracking
  - Not restricted to ground-plane motion
  - Handles various poses
  - Consistent identification by appearance information
  - Fast, on-line tracking







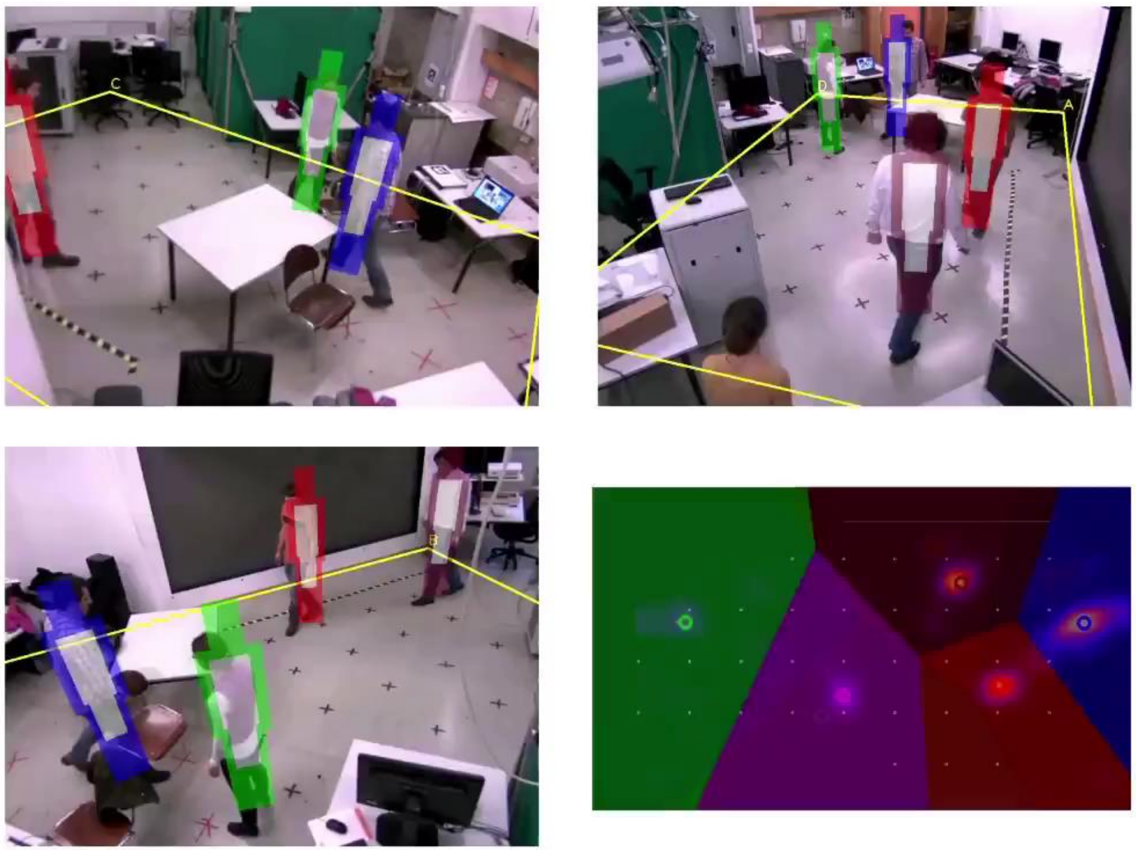
Outlook

- Incorporate a priori information
  e.g., static occluder inference [1]
- Specific challenges
  e.g., jersey number extraction [2]
- View planning
  e.g., Pan-Tilt-Zoom (PTZ) cameras
- Publications
  - CVWW
    OCG Best Student Paper Award
  - CVPR

[1] Guan, Franco, and Pollefeys. *3D Occlusion Inference from Silhouette Cues*. In Proc. CVPR, 2007.
[2] Shitrit, Berclaz, Fleuret, and Fua. *Tracking Multiple People under Global Appearance Constraints*. In Proc. ICCV, 2011.

Thank you!

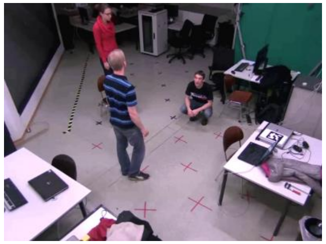
TABLE Dataset

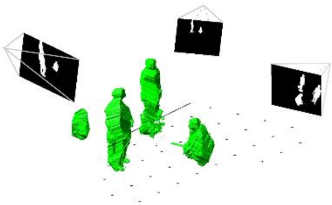
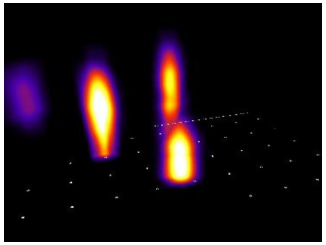
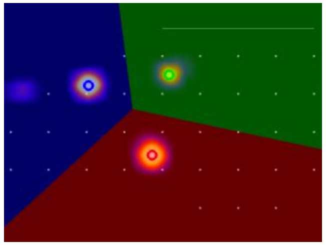
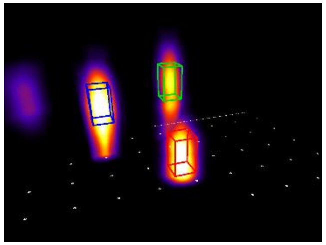
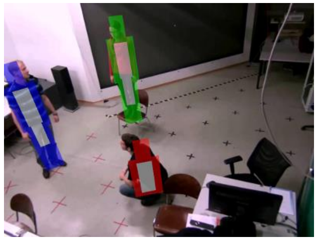
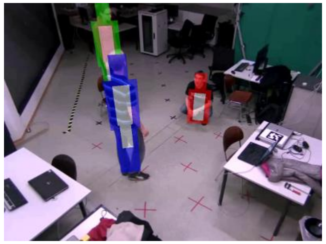


[Video]

Overview

- Visual hull reconstruction
- Local mass densities
  - Less sensitive to noise
  - Valuable tracking cue
- Estimate xy-coordinates
  - Particle filtering
  - Voronoi tesselation
- Find vertical mass center



Visual Hull Reconstruction

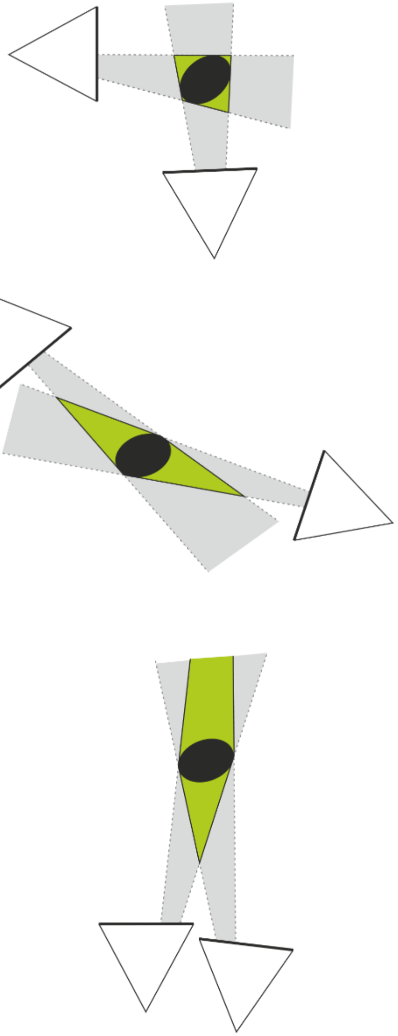
- Subset of cameras
  - Ability to view a voxel
  - Incorporate *chirality* [1]
  - Handle invalid image regions
    e.g., superimposed logos, rectification process

$$\text{visible}(v_i) = \left\{ c \,\middle|\, \text{visible}_c(v_i) = 1 \wedge sign\left(d_{v_i}^{(c)}\right) > 0 \right\}_{c=1}^{N_C}$$

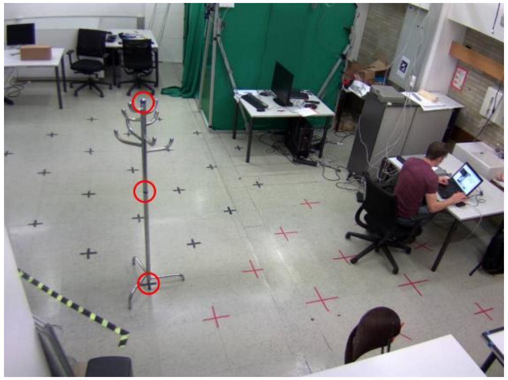
$$\text{visible}_c(v_i) = \begin{cases} 1 & \text{if } \text{project}_c(v_i) \in I_c \\ 0 & \text{otherwise} \end{cases}$$

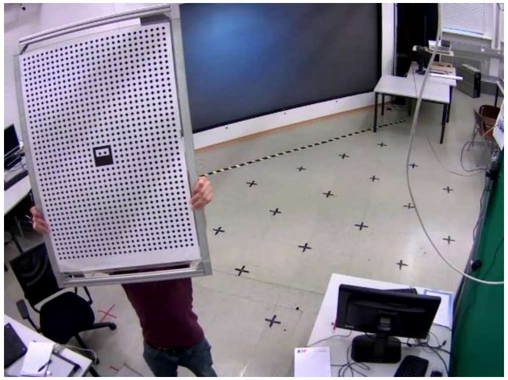
[1] Hartley. *Chirality*. IJCV, 1998.

Visual Hull Quality



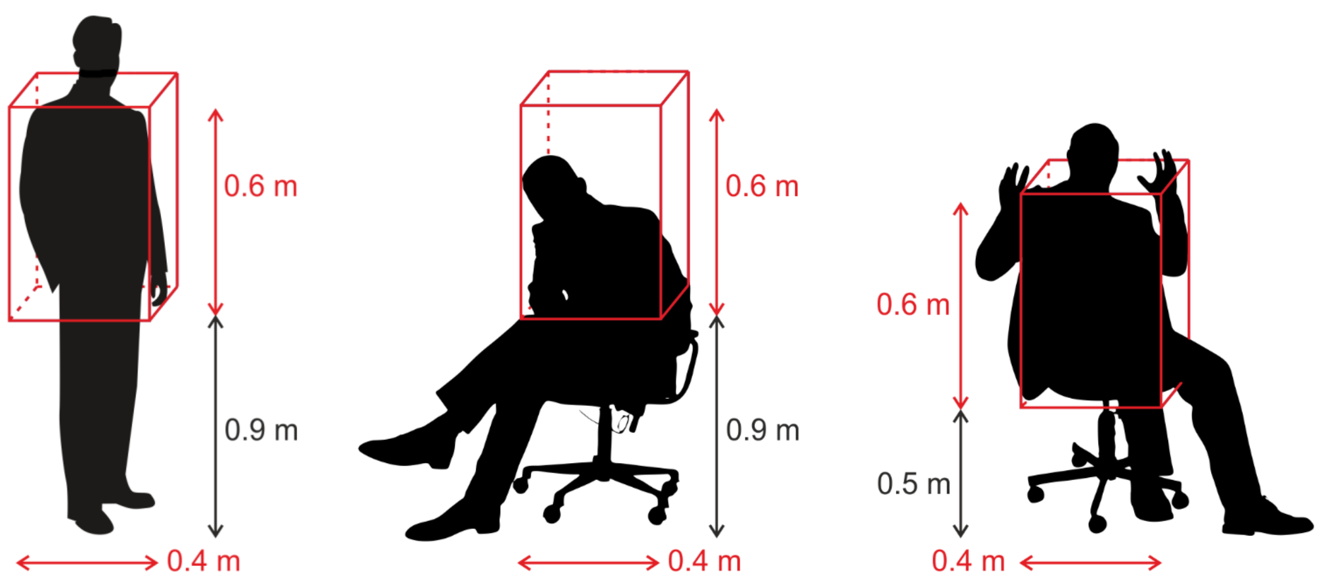
- Camera placement
  - Influence on visual hull quality
  - Large volumes for similar view-points
- Camera calibration
  - Intrinsic & extrinsic parameters
  - 3D targets, automatic methods, etc





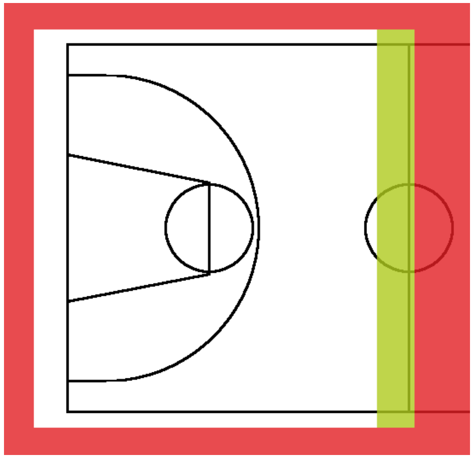
Efficiency – Occupancy Volume

- Mass density computation at specific height levels
- Efficient approximation for xy-coordinates

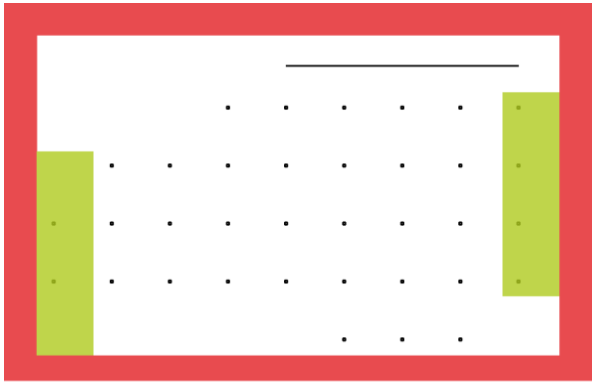


Autonomous Tracking

- Pre-defined entry regions
- Incoming objects
  - Detect via MSER extraction [1]
  - Re-identify using feature bags



APIDIS basketball court.



ICG laboratory.

[1] Matas, Chum, Urban, and Pajdla. *Robust Wide Baseline Stereo from Maximally Stable Extremal Regions*. In Proc. BMVC, 2002.