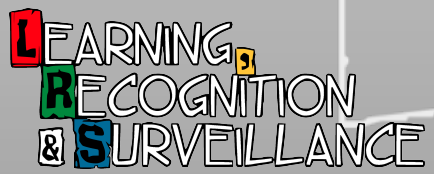
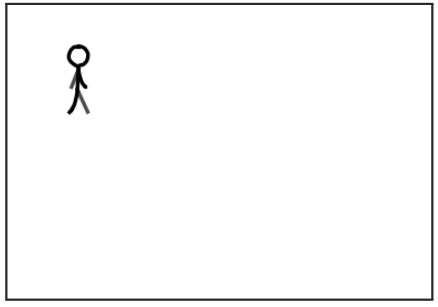
Contextual Cues for Causal Visual Tracking

Horst Possegger

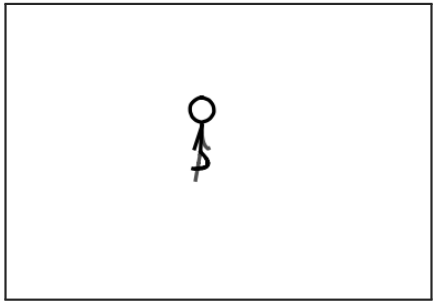
April 24, 2018



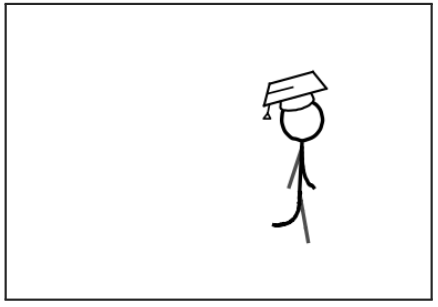
A Quest for the Holy Grail



2013/04/09 08:00



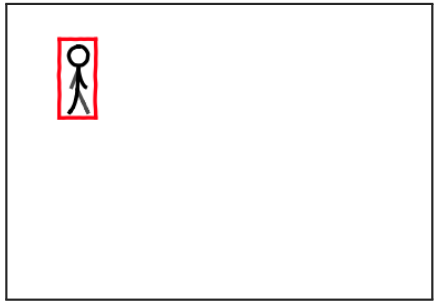
...



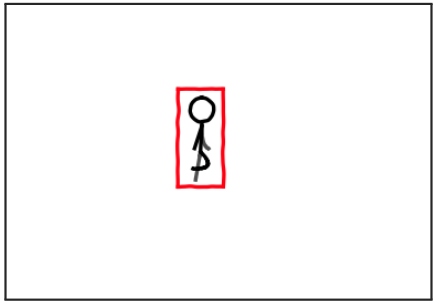
2018/04/24 15:01

- Teaching computers to understand scenes:

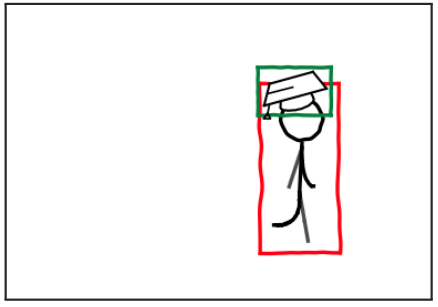
A Quest for the Holy Grail



2013/04/09 08:00



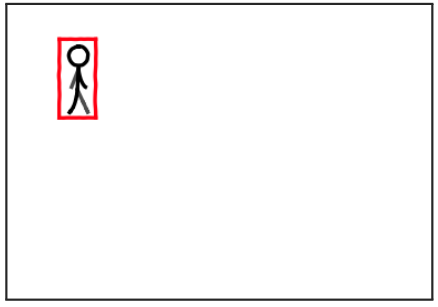
...



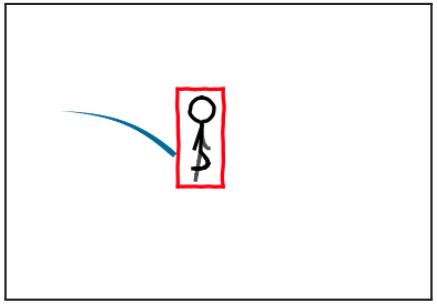
2018/04/24 15:01

- Teaching computers to understand scenes:
  - Object detection and recognition.

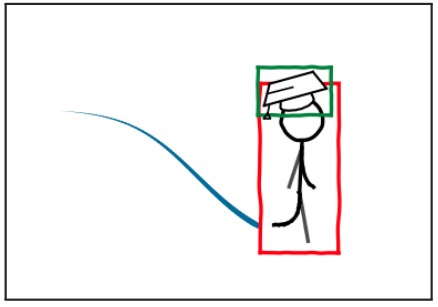
A Quest for the Holy Grail



2013/04/09 08:00



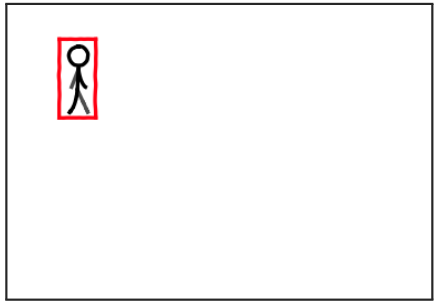
...



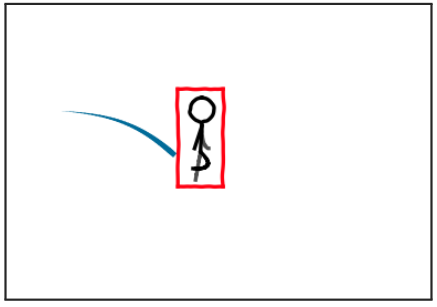
2018/04/24 15:01

- Teaching computers to understand scenes:
  - Object detection and recognition.
  - Localization and tracking.

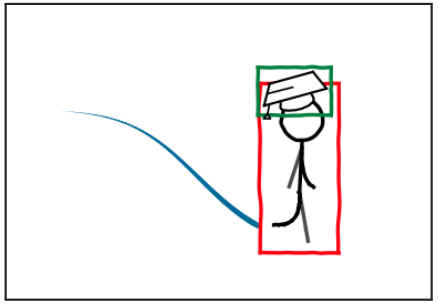
A Quest for the Holy Grail



2013/04/09 08:00



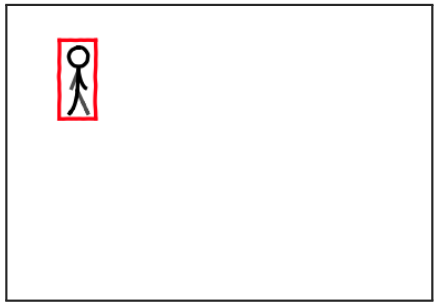
...



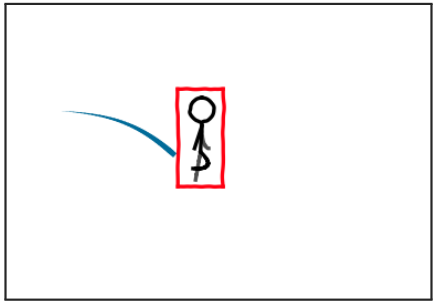
2018/04/24 15:01

- Teaching computers to understand scenes:
  - Object detection and recognition.
  - Localization and tracking.
  - Activity recognition and reasoning.

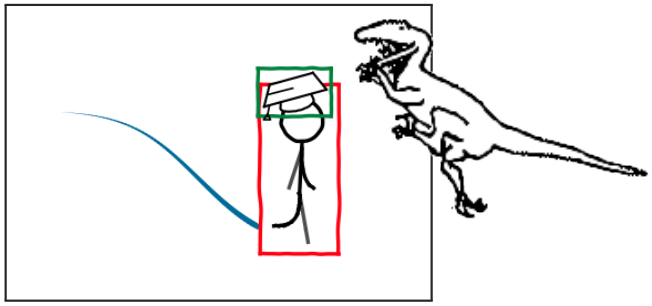
A Quest for the Holy Grail



2013/04/09 08:00



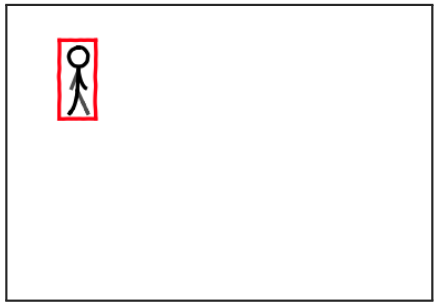
...



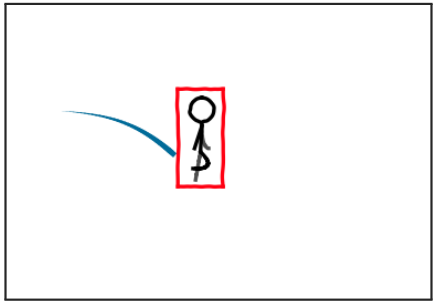
2018/04/24 15:01

- Teaching computers to understand scenes:
  - Object detection and recognition.
  - Localization and tracking.
  - Activity recognition and reasoning.

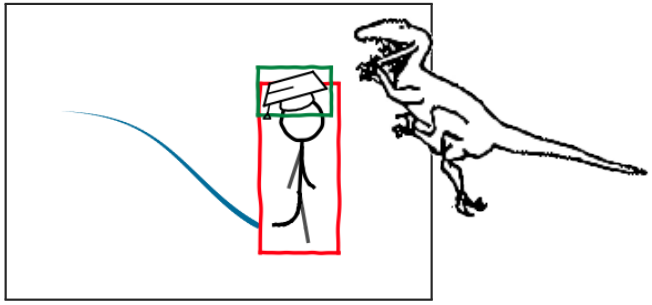
A Quest for the Holy Grail



2013/04/09 08:00



...



2018/04/24 15:01

- Teaching computers to understand scenes:
  - Object detection and recognition.
  - **Localization and tracking (by contextual cues).**
  - Activity recognition and reasoning.

3 The Importance of Context

- Auxiliary information about the scene.
  For example: spatial layout, objects, intents and (inter-)actions.

3 The Importance of Context

- Auxiliary information about the scene.
  For example: spatial layout, objects, intents and (inter-)actions.
- Human visual system:
  - Uses context to **focus attention on challenging** scenarios [1].
  - Leverage prior knowledge & visual stimuli.

[1] Cavanagh and Alvarez. *Tracking Multiple Targets with Multifocal Attention.* TICS 9(7), 2005.

The Importance of Context

- Auxiliary information about the scene.
  For example: spatial layout, objects, intents and (inter-)actions.
- Human visual system:
  - Uses context to **focus attention on challenging** scenarios [1].
  - Leverage prior knowledge & visual stimuli.
- Visual tracking algorithms:
  - Appearance – object versus close surroundings.
  - Motion – target dynamics.
  - Other cues often overlooked!

[1] Cavanagh and Alvarez. *Tracking Multiple Targets with Multifocal Attention.* TICS 9(7), 2005.

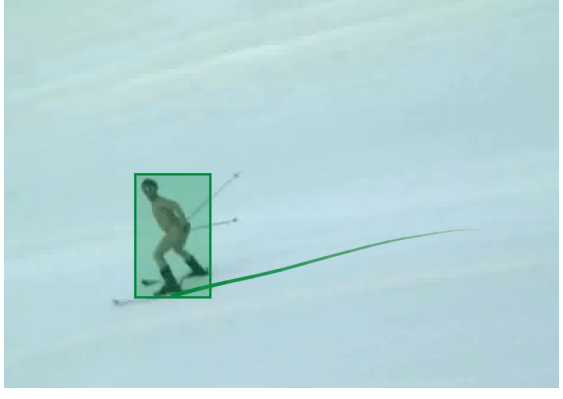
Visual Tracking Requirements

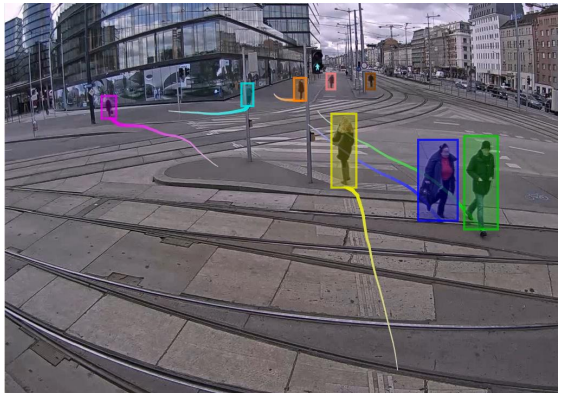
- Robustness.
  Handle appearance variations **but** prevent drifting.

Visual Tracking Requirements

- Robustness.

  Handle appearance variations **but** prevent drifting.
- Efficiency.
  - Time-critical applications.
  - Sports: outdoor (mountain biking, skiing) and indoor (handball).
  - Surveillance: automated pedestrian traffic lights.

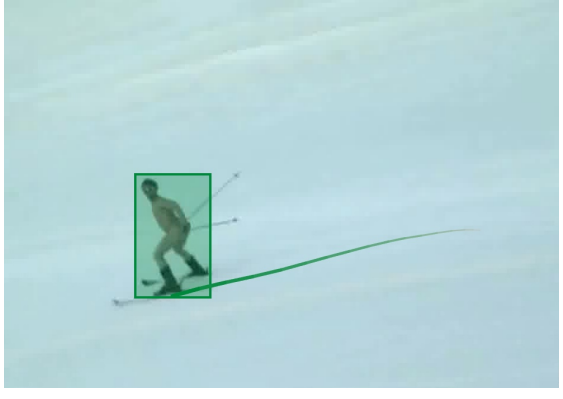
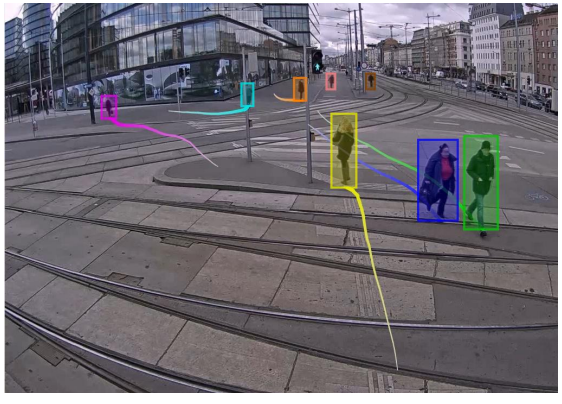




Visual Tracking Requirements

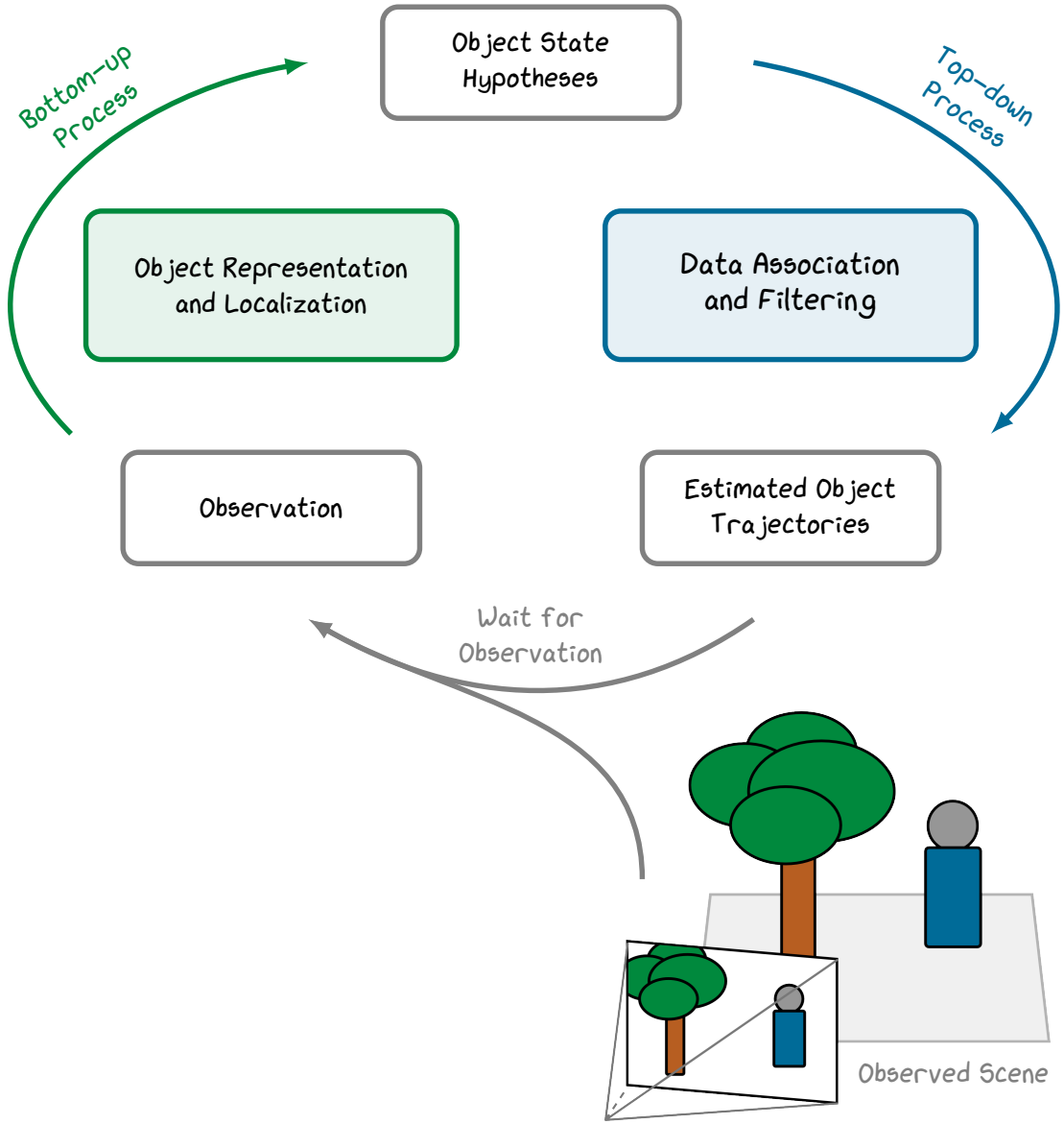
- Robustness.

  Handle appearance variations **but** prevent drifting.
- Efficiency.
  - Time-critical applications.
  - Sports: outdoor (mountain biking, skiing) and indoor (handball).
  - Surveillance: automated pedestrian traffic lights.
- Causal tracking:
  - State inference **solely** from previous/current observation.
  - Cannot alter reported trajectories.

Contributions to the Tracking Loop & Outline

- Visual Tracking Loop:
  Two major components [1].

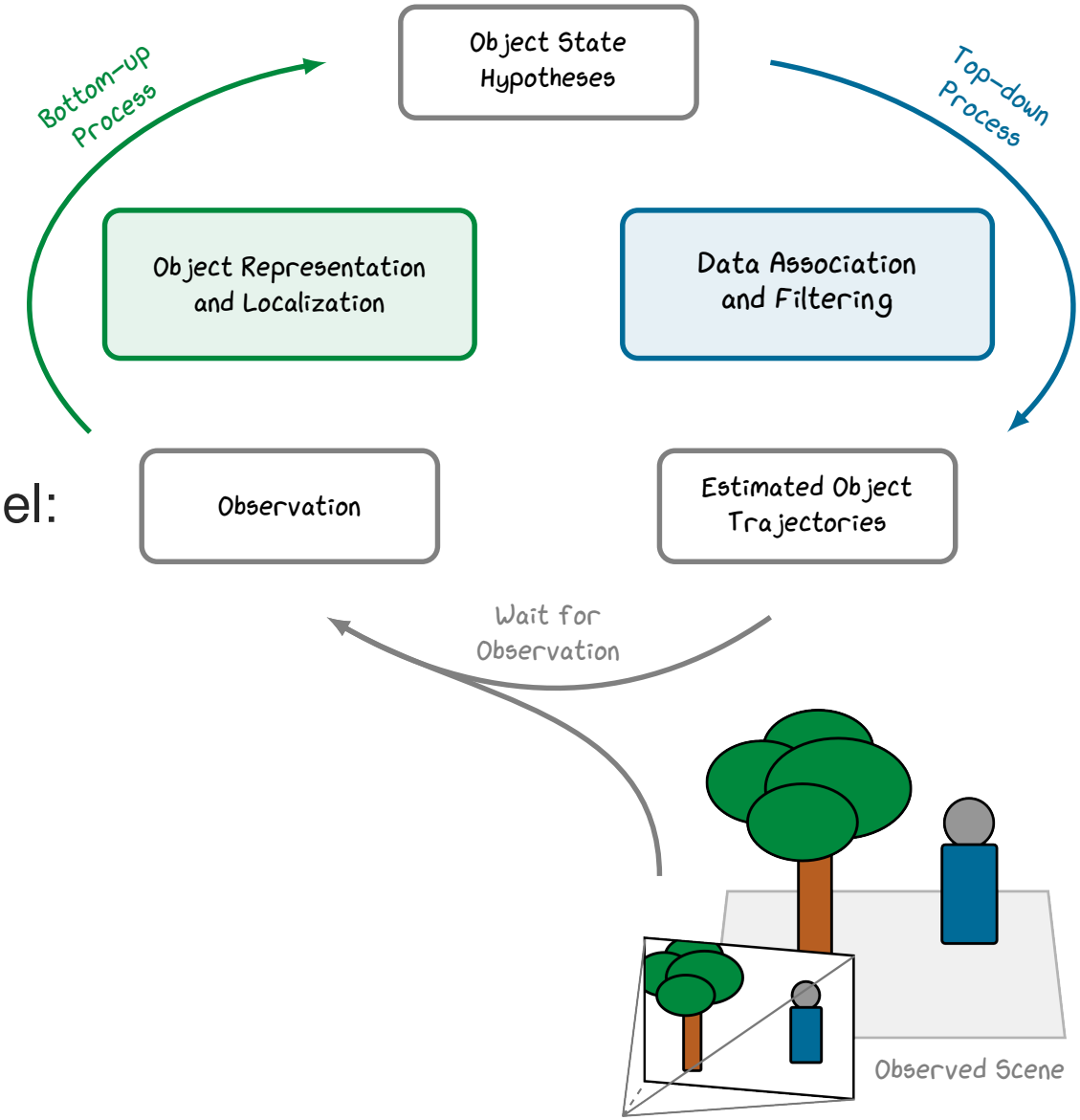


[1] Comaniciu *et al. Real-Time Tracking of Non-Rigid Objects using Mean Shift*. CVPR'00.

Contributions to the Tracking Loop & Outline

- Visual Tracking Loop:
  Two major components [1].

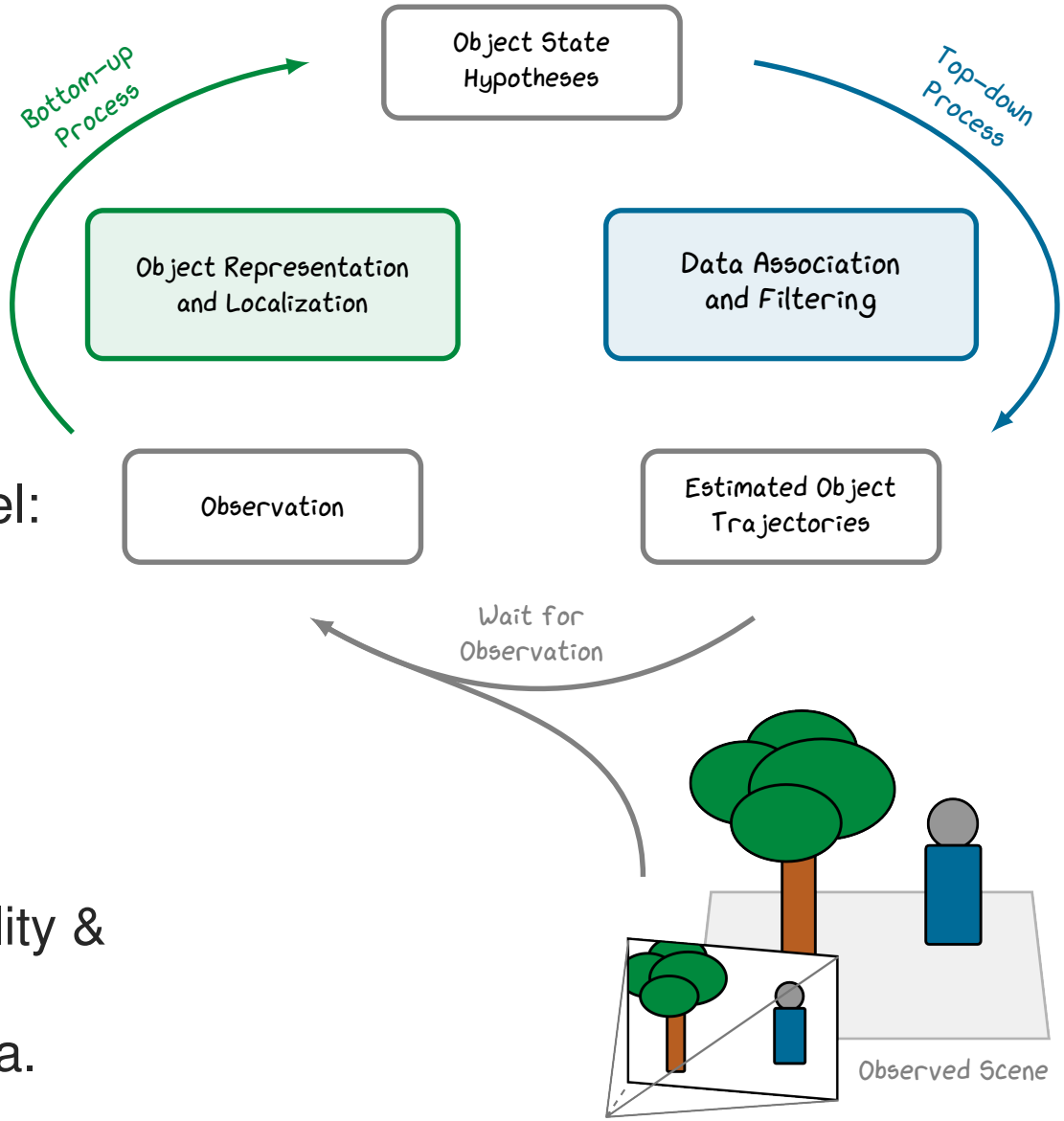
- Distractor-aware object model:
  - Leverage appearance.
  - Reduce drifting.



[1] Comaniciu *et al. Real-Time Tracking of Non-Rigid Objects using Mean Shift*. CVPR'00.

Contributions to the Tracking Loop & Outline

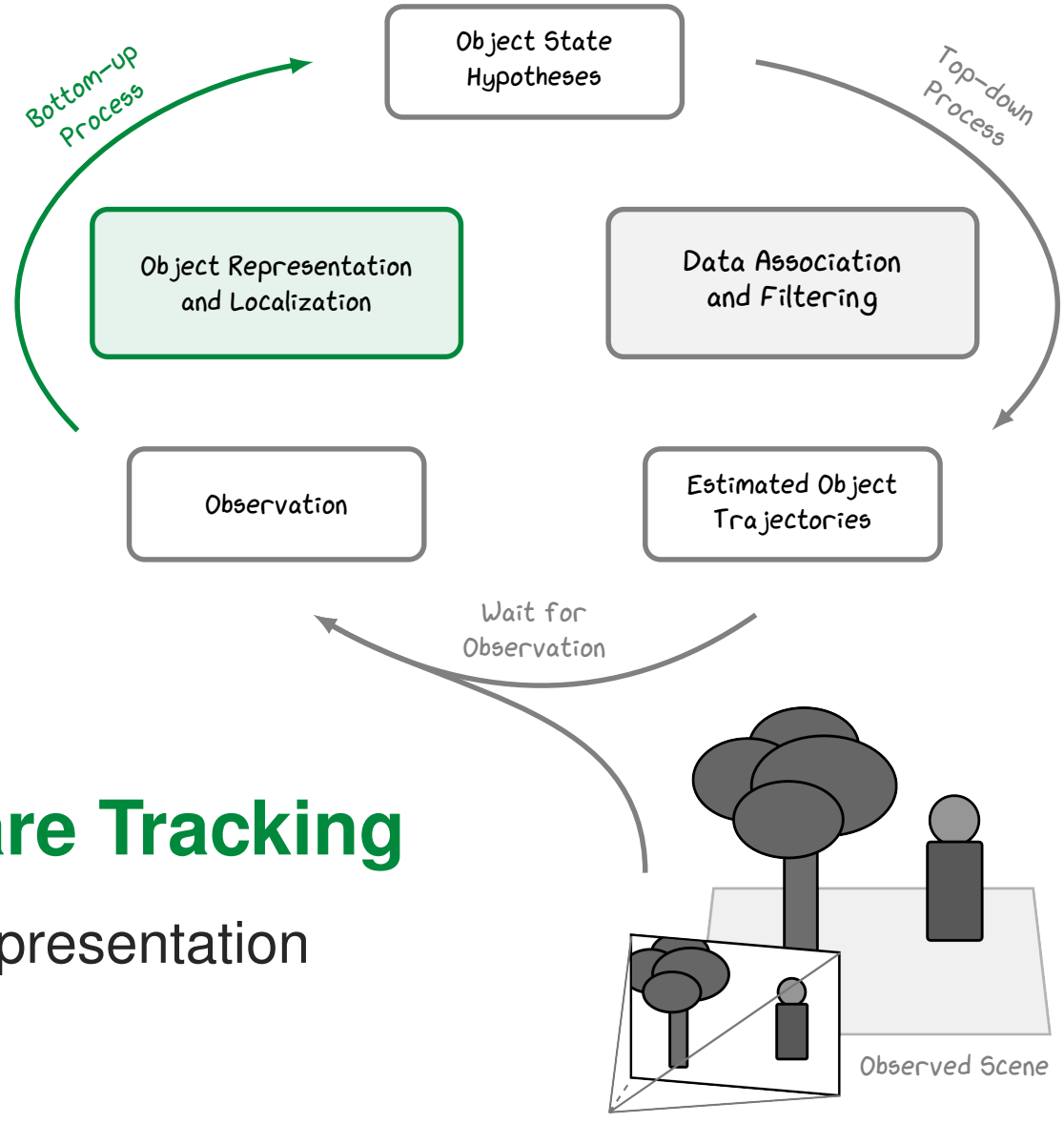
- Visual Tracking Loop:
  Two major components [1].



- Distractor-aware object model:
  - Leverage appearance.
  - Reduce drifting.

- Occlusion geodesics:
  - Leverage physical plausibility & occlusion knowledge.
  - Robust association schema.

[1] Comaniciu *et al. Real-Time Tracking of Non-Rigid Objects using Mean Shift*. CVPR'00.

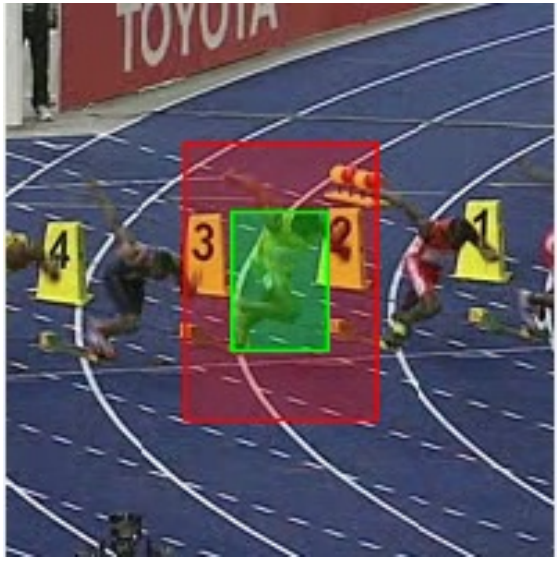


Distractor-aware Tracking

Appearance and Representation

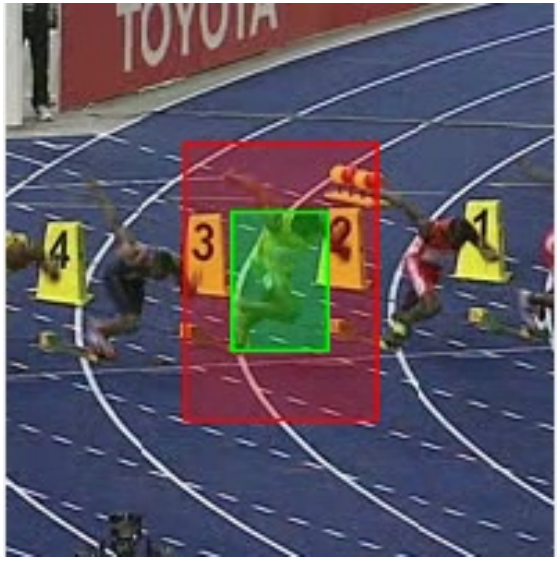
Exploiting Color Distributions

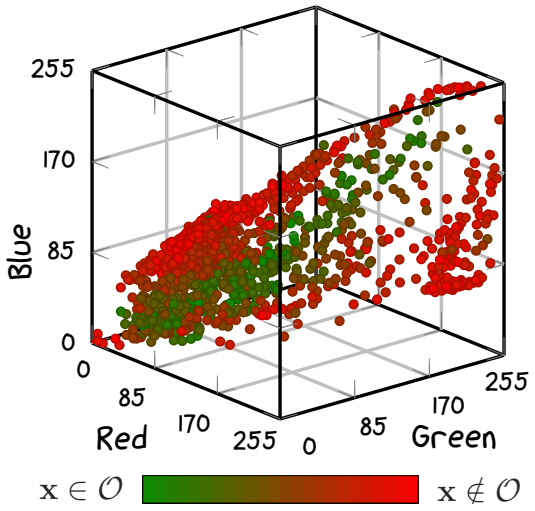
Sequence `bolt` - Object vs. Surroundings

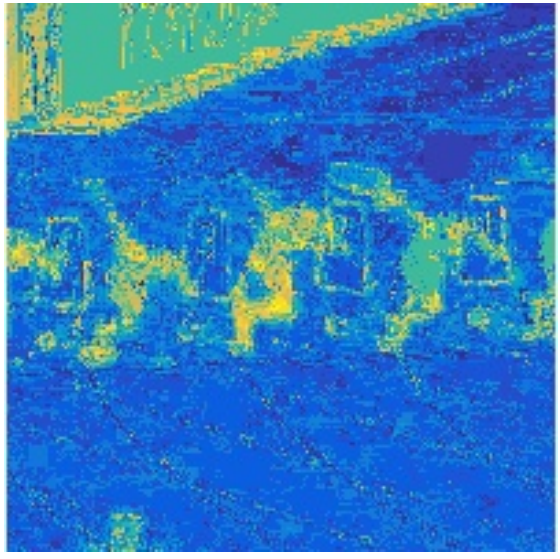


Exploiting Color Distributions

Sequence `bolt` - Object vs. Surroundings

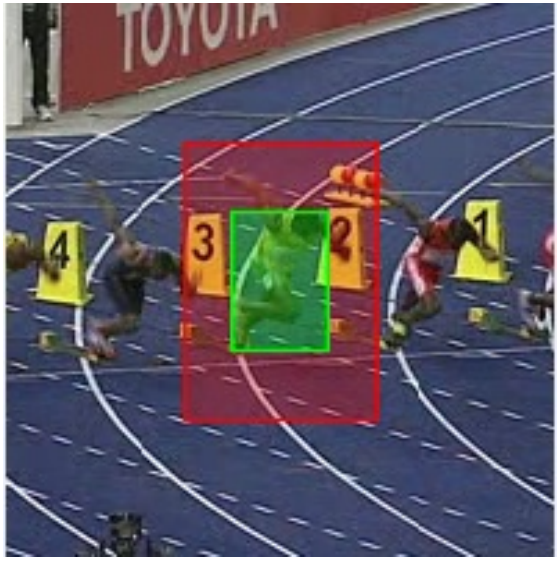


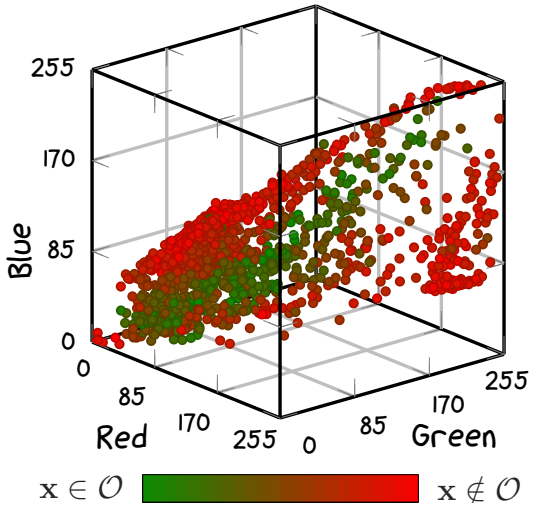


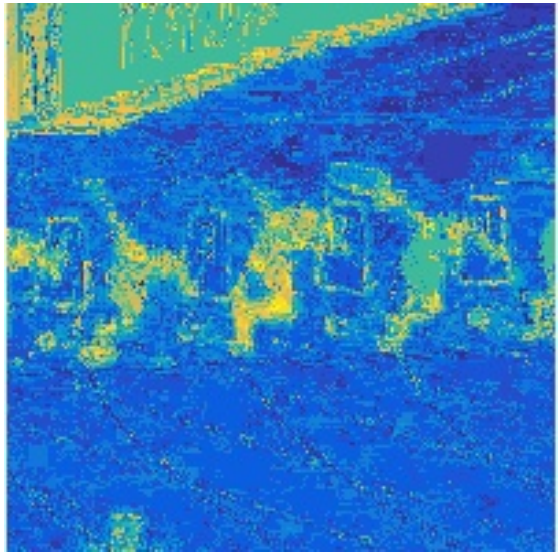


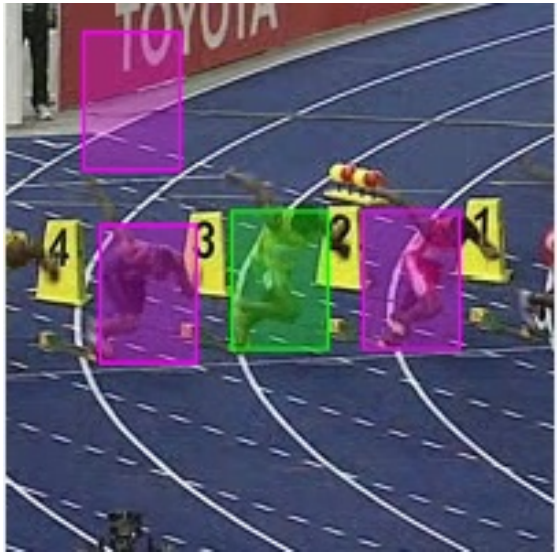
Exploiting Color Distributions

Sequence `bolt` - Object vs. Distractors



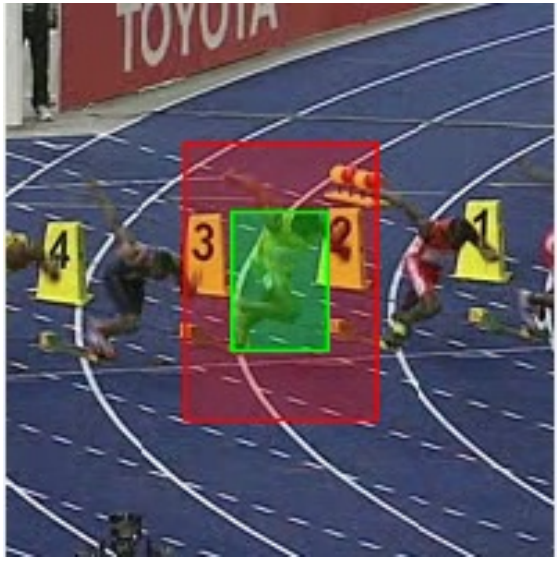


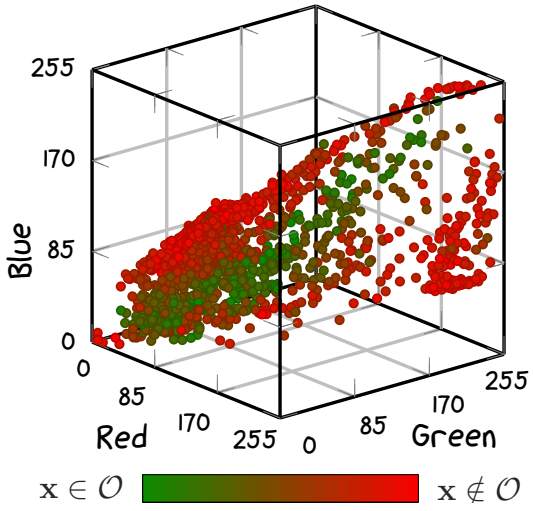


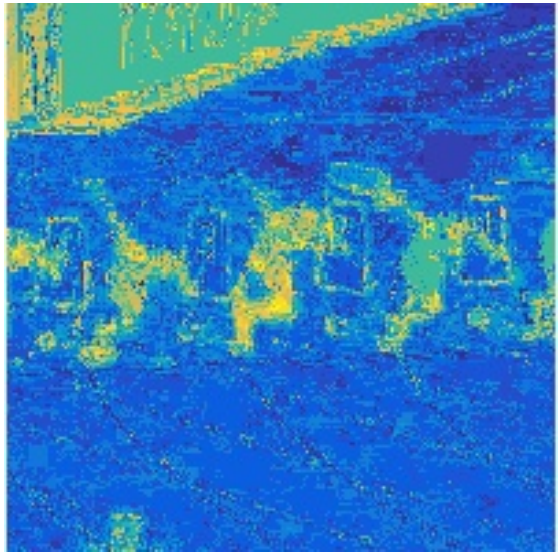


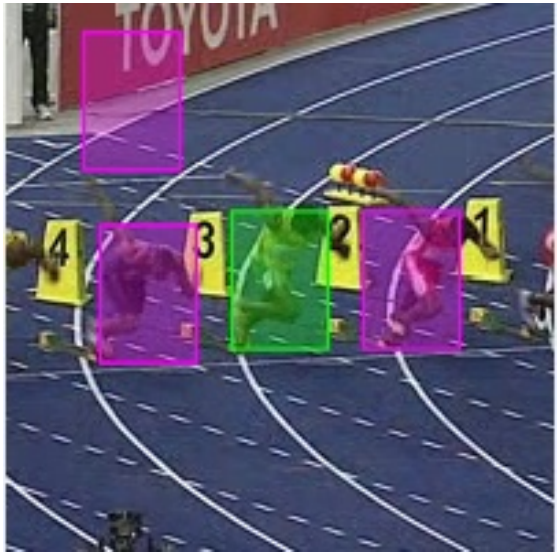
Exploiting Color Distributions

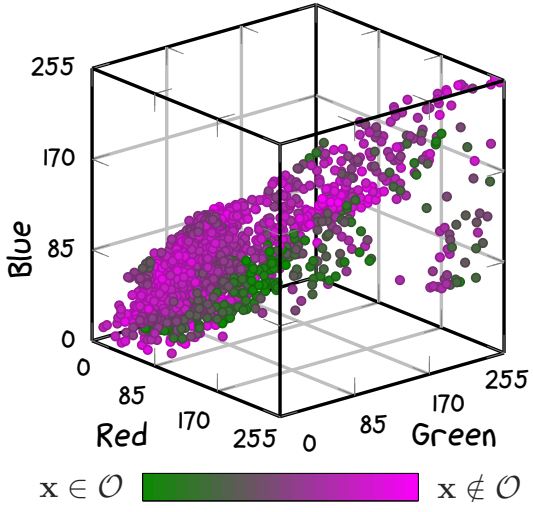
Sequence `bolt` - Object vs. Distractors

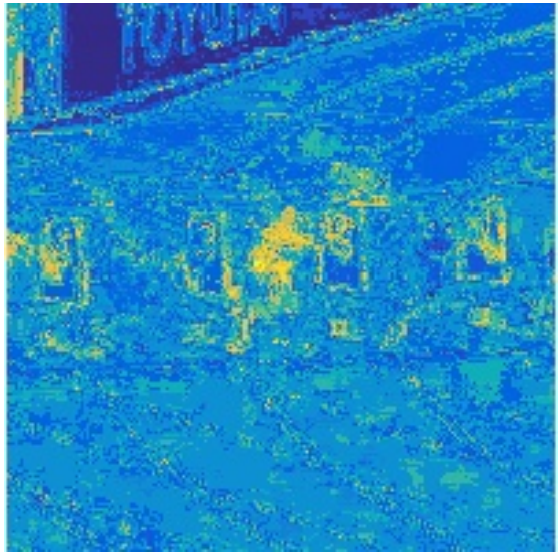












Object-versus-Surroundings

- Bayes' theorem:

$$p(\underbrace{\mathbf{x} \in \mathcal{O}}_{\text{Object pixel}} \mid \overbrace{b_{\mathbf{x}}}^{\text{Histogram bin}}) = \frac{p(b_{\mathbf{x}} \mid \mathbf{x} \in \mathcal{O})\, p(\mathbf{x} \in \mathcal{O})}{p(b_{\mathbf{x}})}.$$

Object-versus-Surroundings

- Bayes' theorem:

$$p(\underbrace{\mathbf{x} \in \mathcal{O}}_{\text{Object pixel}} \mid \overbrace{b_{\mathbf{x}}}^{\text{Histogram bin}}) = \frac{p(b_{\mathbf{x}} \mid \mathbf{x} \in \mathcal{O})\, p(\mathbf{x} \in \mathcal{O})}{p(b_{\mathbf{x}})}.$$

- $\mathbf{x} \in \mathcal{O}$ is hard to get, axis-aligned boxes are easy to get
  $\rightarrow$ relax posterior:

$$p(\mathbf{x} \in \mathcal{O} \mid b_{\mathbf{x}}) \approx \frac{p(b_{\mathbf{x}} \mid \mathbf{x} \in O)\, p(\mathbf{x} \in O)}{\sum p(b_{\mathbf{x}} \mid \mathbf{x} \in \Omega)\, p(\mathbf{x} \in \Omega)}.$$

$$\Omega \in \underbrace{\{O, S\}}_{\text{Object region or surroundings}}$$

Object-versus-Surroundings

- Bayes' theorem:

$$p(\underbrace{\mathbf{x} \in \mathcal{O}}_{\text{Object pixel}} \mid \overbrace{b_{\mathbf{x}}}^{\text{Histogram bin}}) = \frac{p(b_{\mathbf{x}} \mid \mathbf{x} \in \mathcal{O})\, p(\mathbf{x} \in \mathcal{O})}{p(b_{\mathbf{x}})}.$$

- $\mathbf{x} \in \mathcal{O}$ is hard to get, axis-aligned boxes are easy to get
  $\rightarrow$ relax posterior:

$$p(\mathbf{x} \in \mathcal{O} \mid b_{\mathbf{x}}) \approx \frac{p(b_{\mathbf{x}} \mid \mathbf{x} \in O)\, p(\mathbf{x} \in O)}{\sum p(b_{\mathbf{x}} \mid \mathbf{x} \in \Omega)\, p(\mathbf{x} \in \Omega)}.$$

$$\Omega \in \underbrace{\{O, S\}}_{\text{Object region or surroundings}}$$

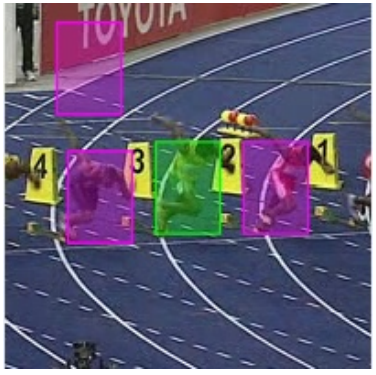
- Leverage color histograms to compute likelihood and prior terms:

$$\underbrace{p_{O,S}}_{\text{Object-versus-Surroundings}}(\mathbf{x} \in \mathcal{O} \mid b_{\mathbf{x}}) = \frac{\overbrace{H_O^I(b_{\mathbf{x}})}^{\text{Histogram over region } O \text{ of image } I} \quad +1}{H_O^I(b_{\mathbf{x}}) + H_S^I(b_{\mathbf{x}}) \quad \underbrace{+2}_{\text{Laplace smoothing}}}.$$

Object-versus-Distractors

- Identify distracting regions D at time t.
  - D is a by-product of localization.
  - Leverage color histograms:

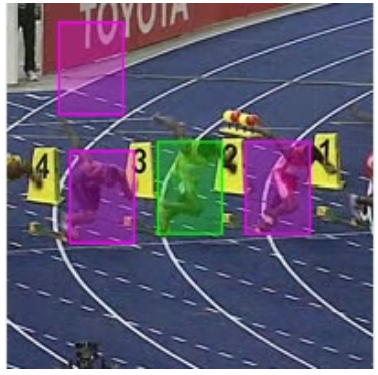
$$\underbrace{p^t_{O,D}}_{\text{Object-versus-Distractors at time } t}(\mathbf{x} \in \mathcal{O} \,|\, b_{\mathbf{x}}) = \frac{H^I_O(b_{\mathbf{x}}) + 1}{H^I_O(b_{\mathbf{x}}) + H^I_D(b_{\mathbf{x}}) + 2}.$$



Object-versus-Distractors

- Identify distracting regions D at time t.
  - D is a by-product of localization.
  - Leverage color histograms:

$$\underbrace{p^t_{O,D}}_{\text{Object-versus-Distractors at time } t}(\mathbf{x} \in \mathcal{O} \mid b_{\mathbf{x}}) = \frac{H^I_O(b_{\mathbf{x}}) + 1}{H^I_O(b_{\mathbf{x}}) + H^I_D(b_{\mathbf{x}}) + 2}.$$



- Update model:

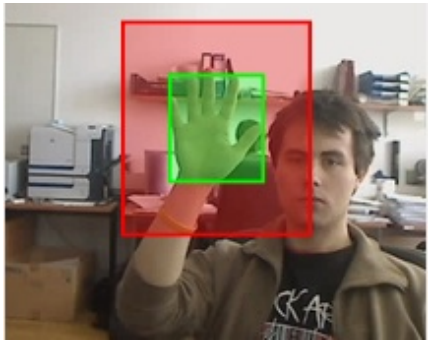
$$p^{1:t}_{O,D}(\mathbf{x} \in \mathcal{O} \mid b_{\mathbf{x}}) = \eta_D \underbrace{p^t_{O,D}(\mathbf{x} \in \mathcal{O} \mid b_{\mathbf{x}})}_{\text{Current frame}} + (1 - \eta_D) \underbrace{p^{1:t-1}_{O,D}(\mathbf{x} \in \mathcal{O} \mid b_{\mathbf{x}})}_{\text{Accumulated model}}$$

- Similarly, update object-versus-surroundings model $p^{1:t}_{O,S}(\cdot)$.

Localization & Scale Adaptation

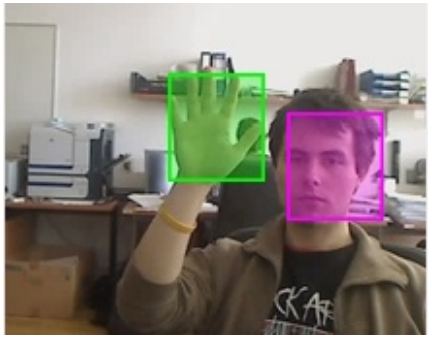
Challenges

Inputs I, O and S



$p^{1:t}_{O,S}(\mathbf{x} \in \mathcal{O} \mid b_{\mathbf{x}})$

Inputs I, O and D



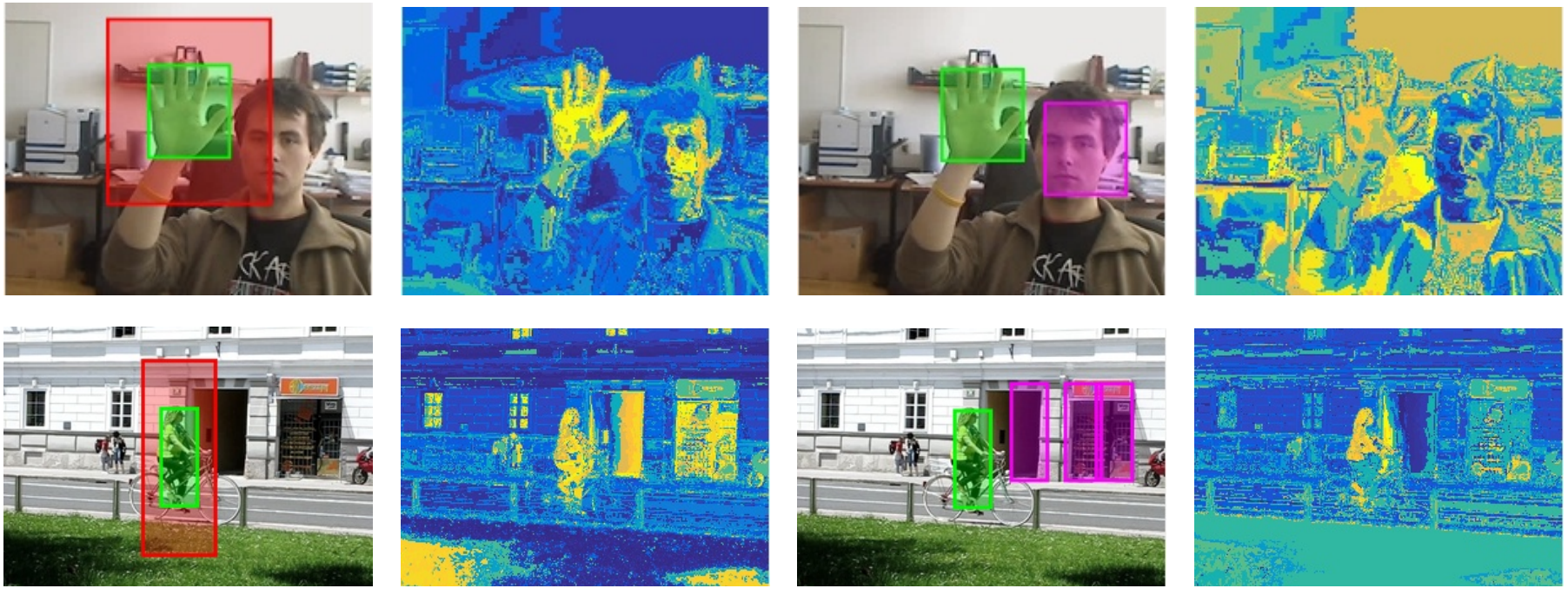
$p^{1:t}_{O,D}(\mathbf{x} \in \mathcal{O} \mid b_{\mathbf{x}})$

- Top: suppressed distracting region on face.

Localization & Scale Adaptation

Challenges

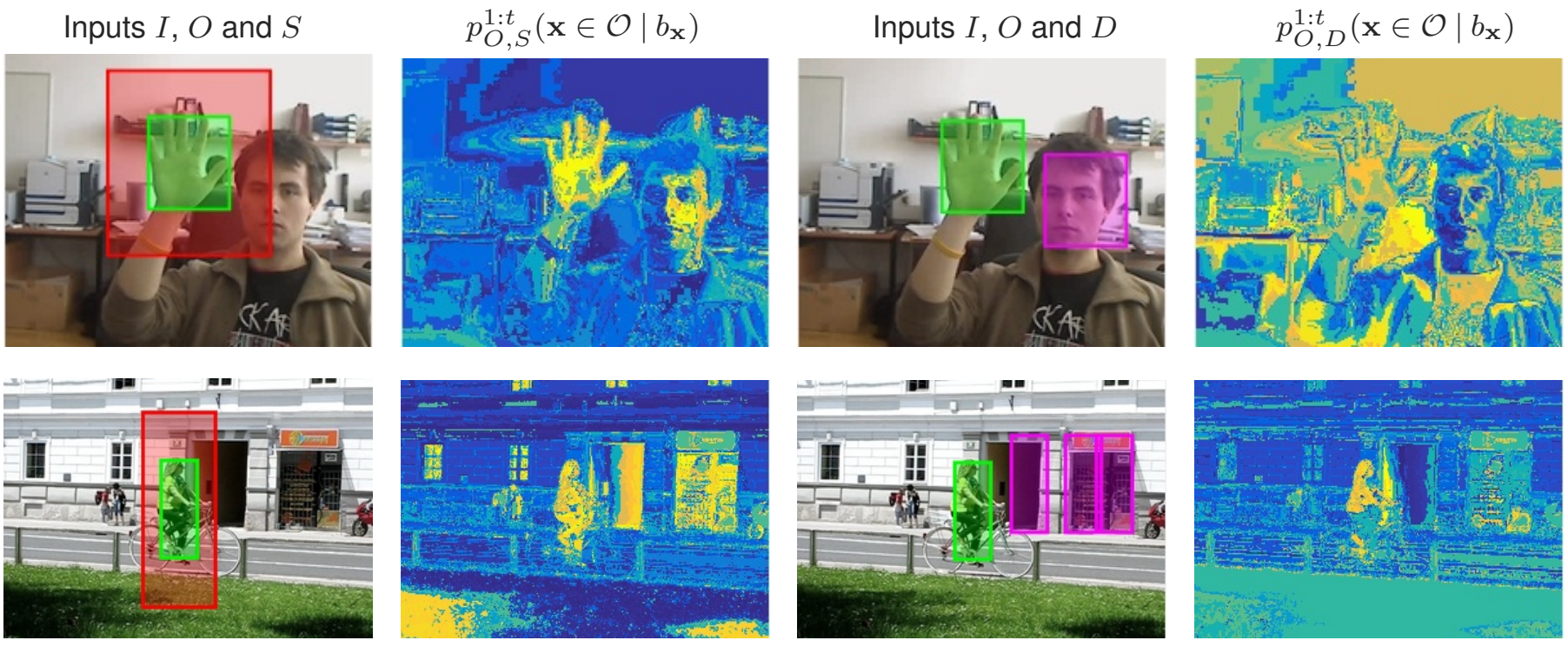
Inputs I, O and S | $p^{1:t}_{O,S}(\mathbf{x} \in \mathcal{O} \mid b_{\mathbf{x}})$ | Inputs I, O and D | $p^{1:t}_{O,D}(\mathbf{x} \in \mathcal{O} \mid b_{\mathbf{x}})$



- Top: suppressed distracting region on face.
- Bottom: suppressed dark regions at the doorways.

Localization & Scale Adaptation

Challenges



- Top: suppressed distracting region on face.
- Bottom: suppressed dark regions at the doorways.
- Cannot simply combine models **for both** localization and scaling.

Localization

Non-maximum suppression (NMS)

- Densely sample object hypotheses $O_{i,j}^t$ within search window.
- Perform NMS:

$$O_\star^t = \underset{O_{i,j}^t}{\arg\max} \underbrace{\left(\rho_S(O_{i,j}^t) + \rho_D(O_{i,j}^t) \right)}_{\text{Appearance term}} \underbrace{\exp\left(-\frac{\left\| \mathbf{c}^{t-1} - \mathbf{c}_{i,j}^t \right\|_2^2}{2\sigma^2} \right)}_{\text{Motion term}}$$

Localization

Non-maximum suppression (NMS)

- Densely sample object hypotheses $O_{i,j}^t$ within search window.
- Perform NMS:

$$O_\star^t = \arg\max_{O_{i,j}^t} \underbrace{\left(\rho_S(O_{i,j}^t) + \rho_D(O_{i,j}^t) \right)}_{\text{Appearance term}} \underbrace{\exp\left(-\frac{\left\| \mathbf{c}^{t-1} - \mathbf{c}_{i,j}^t \right\|_2^2}{2\sigma^2} \right)}_{\text{Motion term}}$$

$$\rho_S(O_{i,j}^t) = \frac{1}{2} \left(\underbrace{\frac{1}{\left| O_{i,j}^t \right|} \sum_{\mathbf{x} \in O_{i,j}^t} p_{O,S}^{1:t-1}(\mathbf{x} \in \mathcal{O} \mid b_\mathbf{x})}_{\text{Mean likelihood over } O_{i,j}^t} + \underbrace{\frac{1}{\left| \overline{O_{i,j}^t} \right|} \sum_{\mathbf{x} \in \overline{O_{i,j}^t}} p_{O,S}^{1:t-1}(\mathbf{x} \in \mathcal{O} \mid b_\mathbf{x})}_{\text{Center region of } O_{i,j}^t} \right)$$

Localization

Non-maximum suppression (NMS)

- Densely sample object hypotheses $O_{i,j}^t$ within search window.
- Perform NMS:

$$O_\star^t = \arg\max_{O_{i,j}^t} \underbrace{\left(\rho_S(O_{i,j}^t) + \rho_D(O_{i,j}^t) \right)}_{\text{Appearance term}} \underbrace{\exp\left(-\frac{\left\| \mathbf{c}^{t-1} - \mathbf{c}_{i,j}^t \right\|_2^2}{2\sigma^2} \right)}_{\text{Motion term}}$$

$$\rho_S(O_{i,j}^t) = \frac{1}{2}\left(\underbrace{\frac{1}{\left|O_{i,j}^t\right|} \sum_{\mathbf{x} \in O_{i,j}^t} p_{O,S}^{1:t-1}(\mathbf{x} \in \mathcal{O} \mid b_\mathbf{x})}_{\text{Mean likelihood over } O_{i,j}^t} + \underbrace{\frac{1}{\left|\overline{O_{i,j}^t}\right|} \sum_{\mathbf{x} \in \overline{O_{i,j}^t}} p_{O,S}^{1:t-1}(\mathbf{x} \in \mathcal{O} \mid b_\mathbf{x})}_{\text{Center region of } O_{i,j}^t} \right)$$

$$\rho_D(O_{i,j}^t) = \frac{1}{|O_{i,j}^t|} \sum_{\mathbf{x} \in O_{i,j}^t} p_{O,D}^{1:t-1}(\mathbf{x} \in \mathcal{O} \mid b_\mathbf{x})$$

Localization

Non-maximum suppression (NMS)

- Densely sample object hypotheses $O_{i,j}^t$ within search window.
- Perform NMS:

$$O_\star^t = \arg\max_{O_{i,j}^t} \underbrace{\left(\rho_S(O_{i,j}^t) + \rho_D(O_{i,j}^t) \right)}_{\text{Appearance term}} \underbrace{\exp\left(-\frac{\left\| \mathbf{c}^{t-1} - \mathbf{c}_{i,j}^t \right\|_2^2}{2\sigma^2} \right)}_{\text{Motion term}}$$

$$\rho_S(O_{i,j}^t) = \frac{1}{2} \Bigg(\underbrace{\frac{1}{\left|O_{i,j}^t\right|} \sum_{\mathbf{x} \in O_{i,j}^t} p_{O,S}^{1:t-1}(\mathbf{x} \in \mathcal{O} \mid b_\mathbf{x})}_{\text{Mean likelihood over } O_{i,j}^t} + \underbrace{\frac{1}{\left|\overline{O_{i,j}^t}\right|} \sum_{\mathbf{x} \in \overline{O_{i,j}^t}} p_{O,S}^{1:t-1}(\mathbf{x} \in \mathcal{O} \mid b_\mathbf{x})}_{\text{Center region of } O_{i,j}^t} \Bigg)$$

$$\rho_D(O_{i,j}^t) = \frac{1}{|O_{i,j}^t|} \sum_{\mathbf{x} \in O_{i,j}^t} p_{O,D}^{1:t-1}(\mathbf{x} \in \mathcal{O} \mid b_\mathbf{x})$$

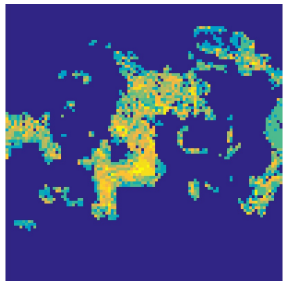
- Retrieve distractors:

$$D = \left\{ O_{i,j}^t \mid \rho_S(O_{i,j}^t) \geq \tau_\nu \, \rho_S(O_\star^t) \right\} \setminus \left\{ O_\star^t \right\}$$

Scale Adaptation

Scale from X

- Pre-segment via $p_{O,S}(\cdot)$ likelihood maps.



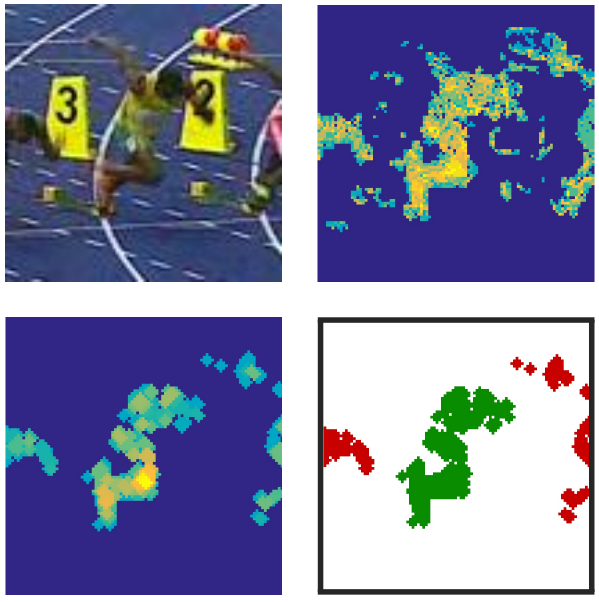
Scale Adaptation

Scale from X

- Pre-segment via $p_{O,S}(\cdot)$ likelihood maps.
- Segmentation via connected components.

[+] Easily understandable heuristics.

[−] Speed tradeoff.



Scale Adaptation

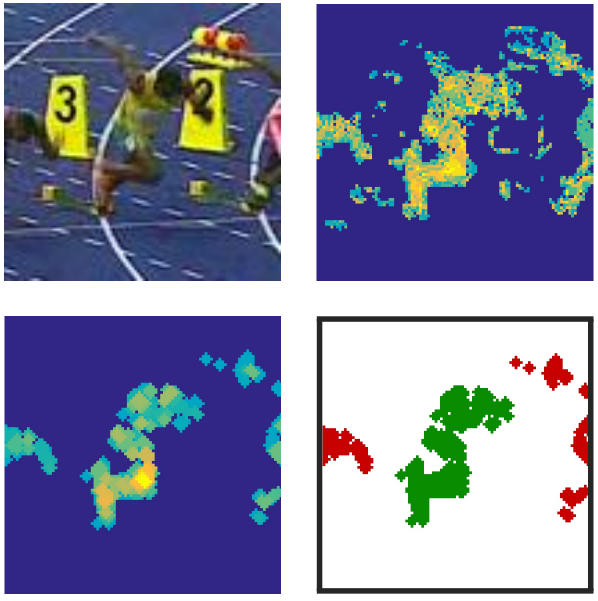
Scale from X

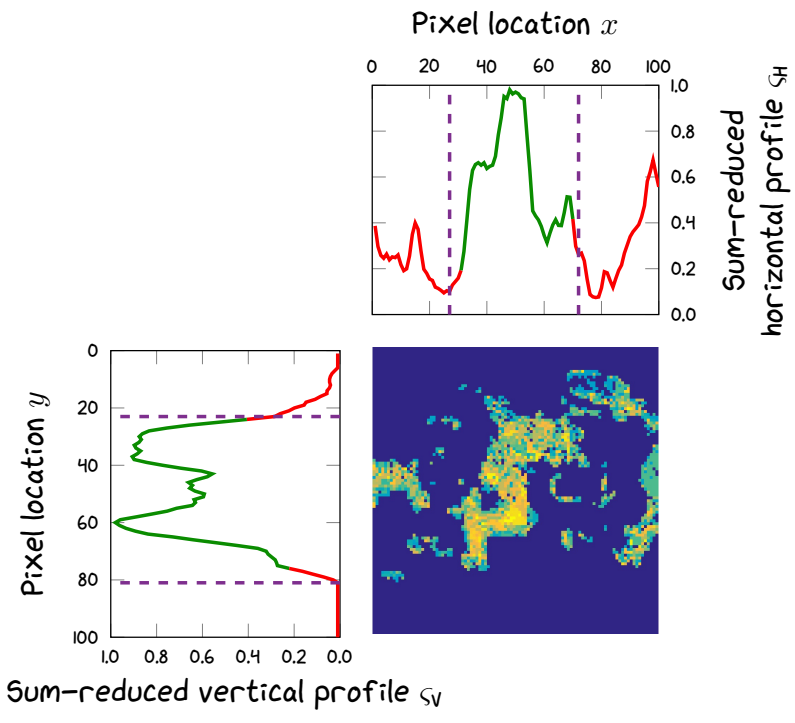
- Pre-segment via $p_{O,S}(\cdot)$ likelihood maps.
- Segmentation via connected components.

[+] Easily understandable heuristics.
[−] Speed tradeoff.

- Sum reduction of likelihood maps.

[+] Fast, useful in practice.
[−] Accuracy for benchmarks.





Scale Adaptation

Scale from X

- Pre-segment via $p_{O,S}(\cdot)$ likelihood maps.
- Segmentation via connected components.

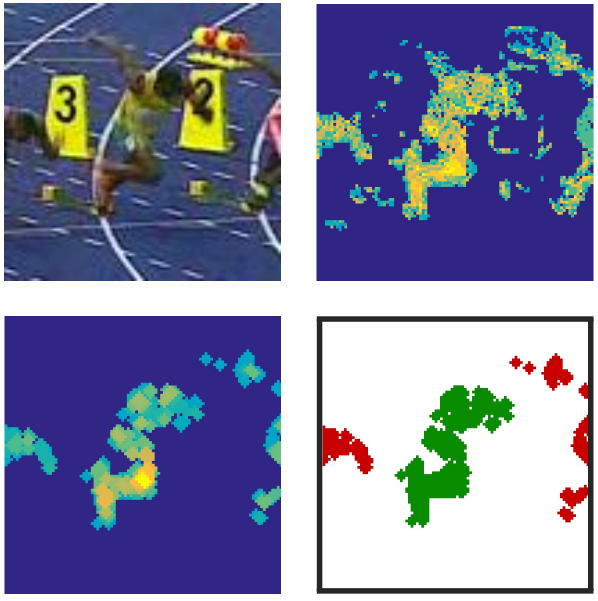
[+] Easily understandable heuristics.
[−] Speed tradeoff.

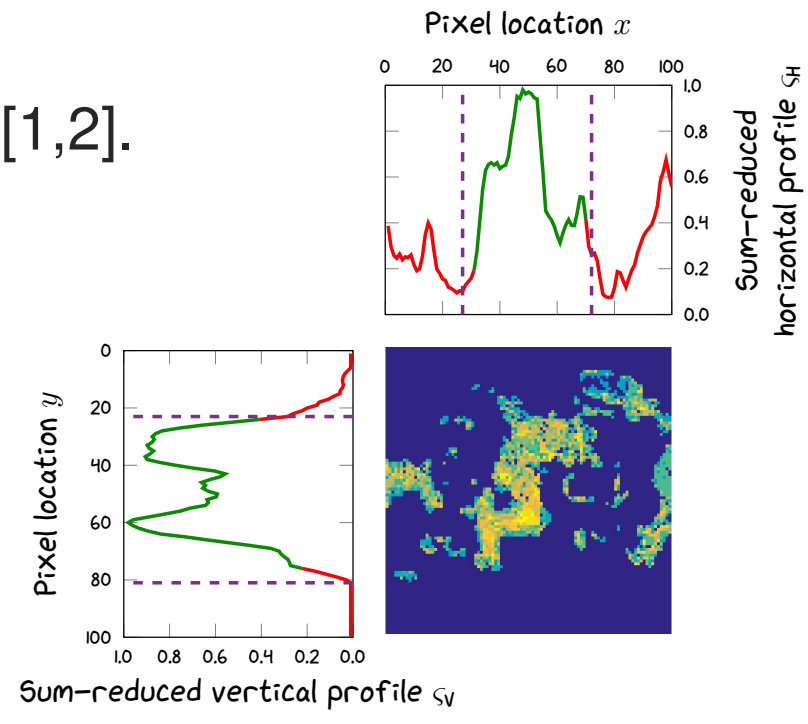
- Sum reduction of likelihood maps.

[+] Fast, useful in practice.
[−] Accuracy for benchmarks.

- Bounding box refinement, similar to [1,2].

[+] Instance-specific regression task.
[−] Clutter & strong deformations.



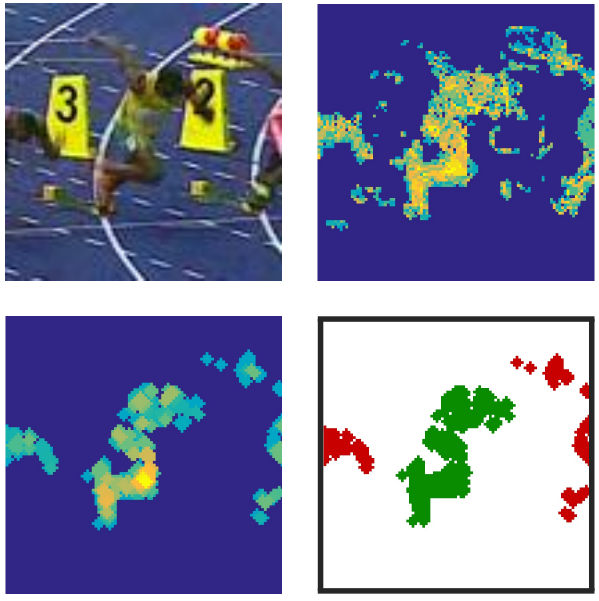


[1] Felzenszwalb *et al. Object Detection with Discriminatively Trained Part Based Models.* TPAMI 32(9), 2010.

[2] Girshick *et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation.* CVPR'14.

Scale Adaptation

Scale from X



- Pre-segment via $p_{O,S}(\cdot)$ likelihood maps.
- Segmentation via connected components.

[+] Easily understandable heuristics.
[−] Speed tradeoff.

- Sum reduction of likelihood maps.

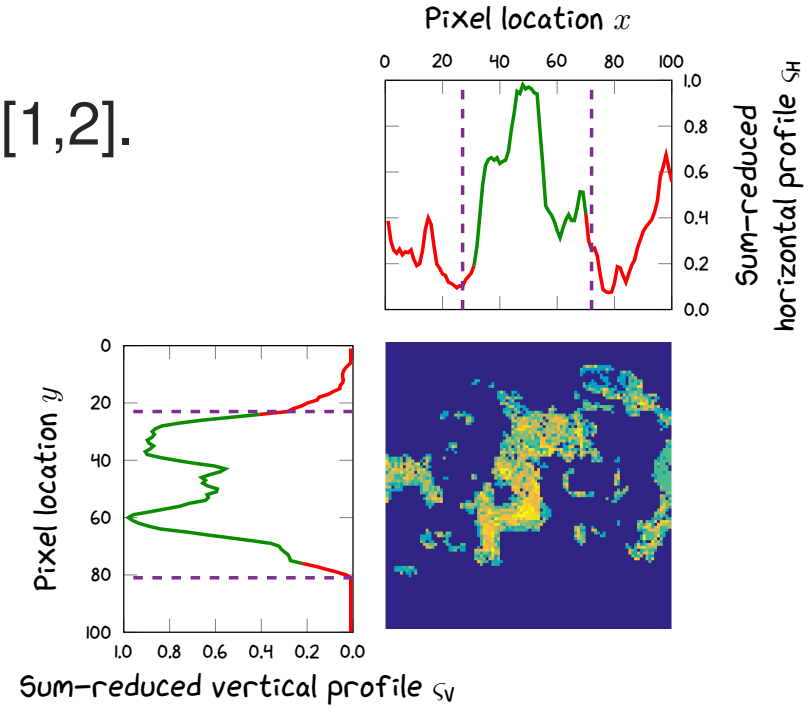
[+] Fast, useful in practice.
[−] Accuracy for benchmarks.

- Bounding box refinement, similar to [1,2].

[+] Instance-specific regression task.
[−] Clutter & strong deformations.

- Segmentation (Graph cuts, TVseg).

[+] Accurate (on high-res images).
[−] Slow; typically low-res videos.

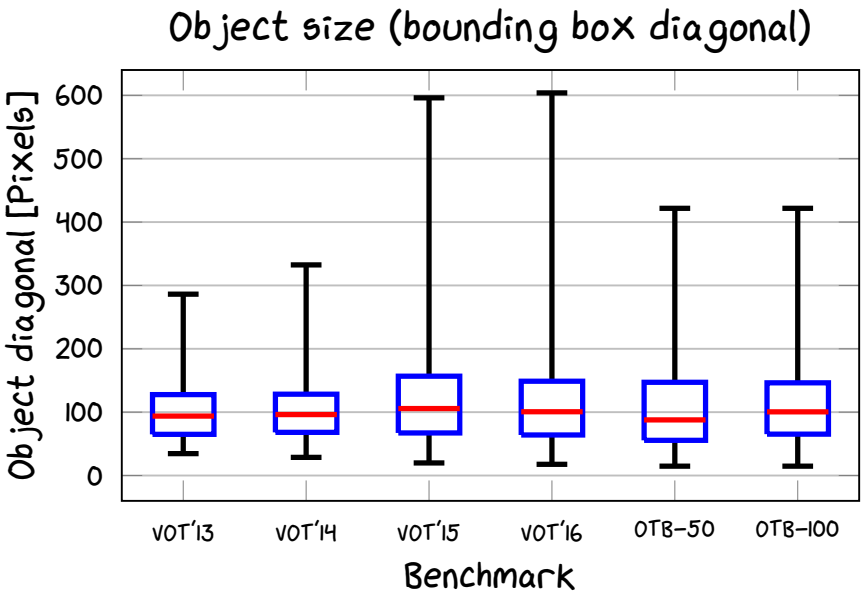


[1] Felzenszwalb *et al. Object Detection with Discriminatively Trained Part Based Models.* TPAMI 32(9), 2010.

[2] Girshick *et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation.* CVPR'14.

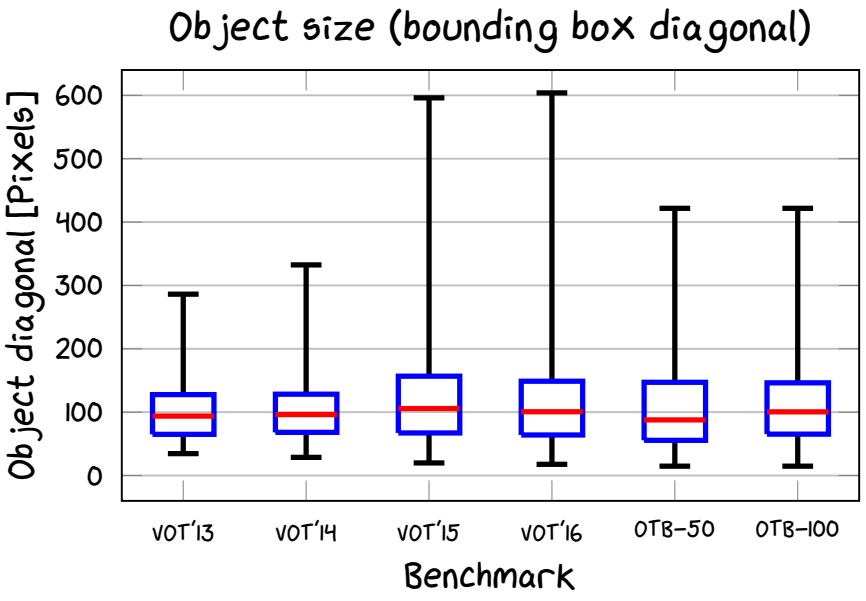
Evaluation

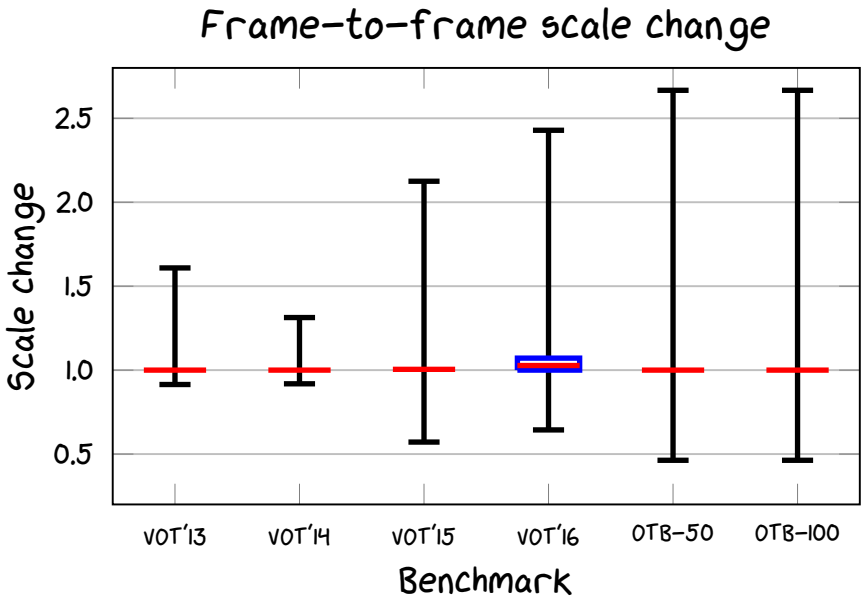
Benchmark Characteristics



Evaluation

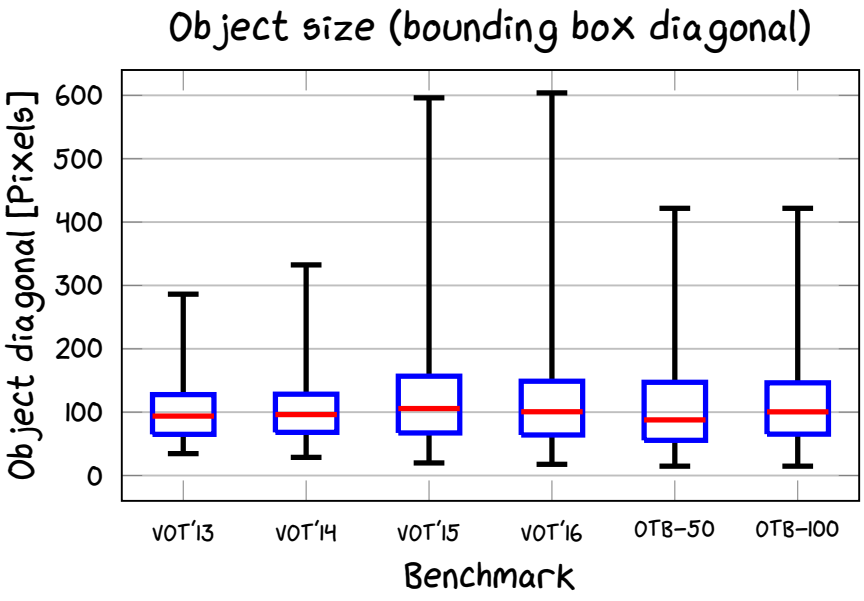
Benchmark Characteristics

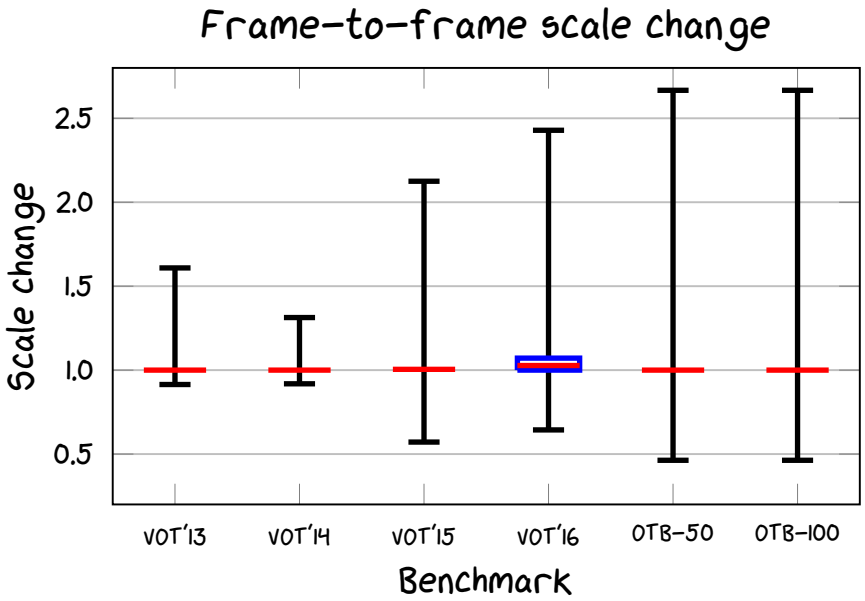


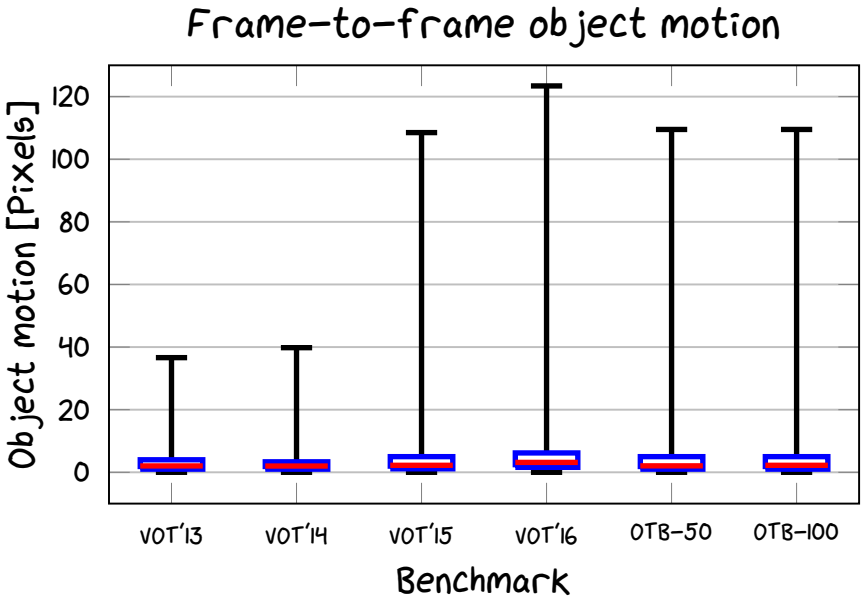


Evaluation

Benchmark Characteristics

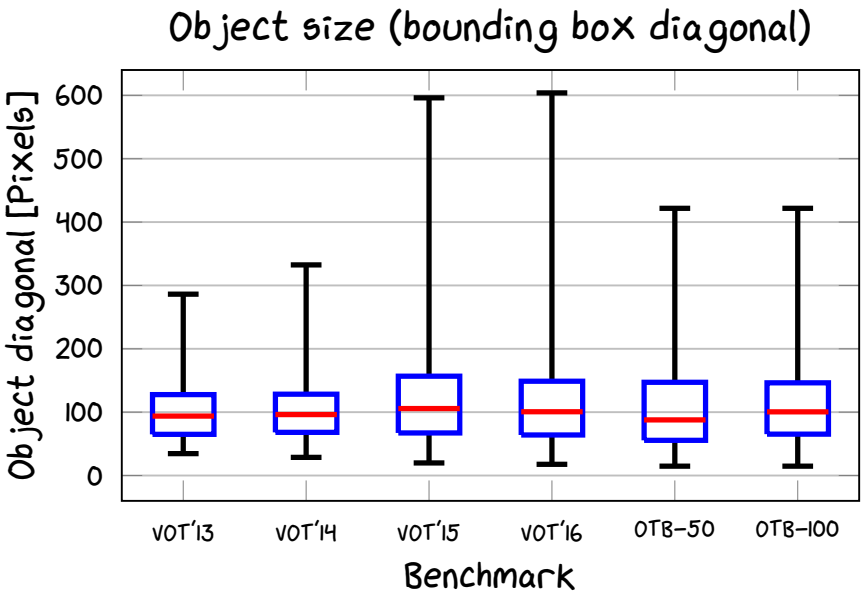


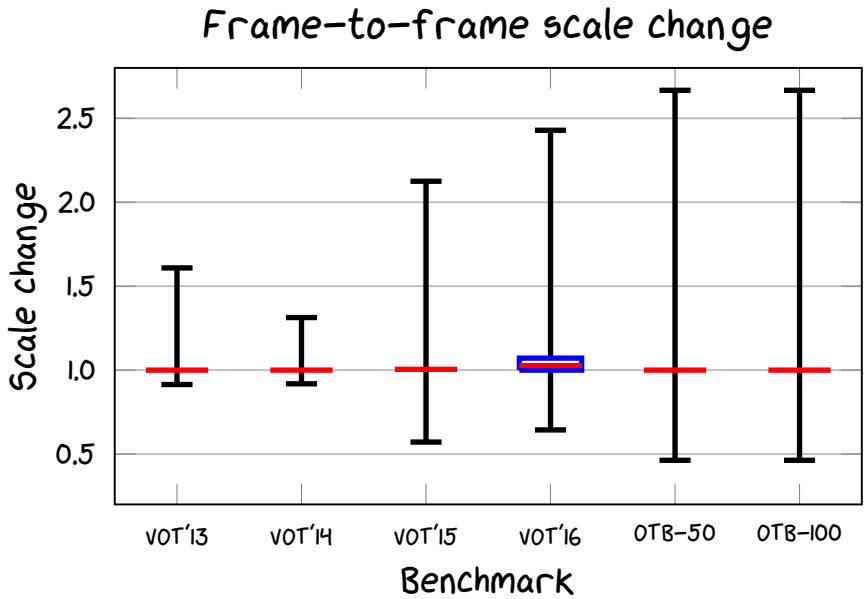


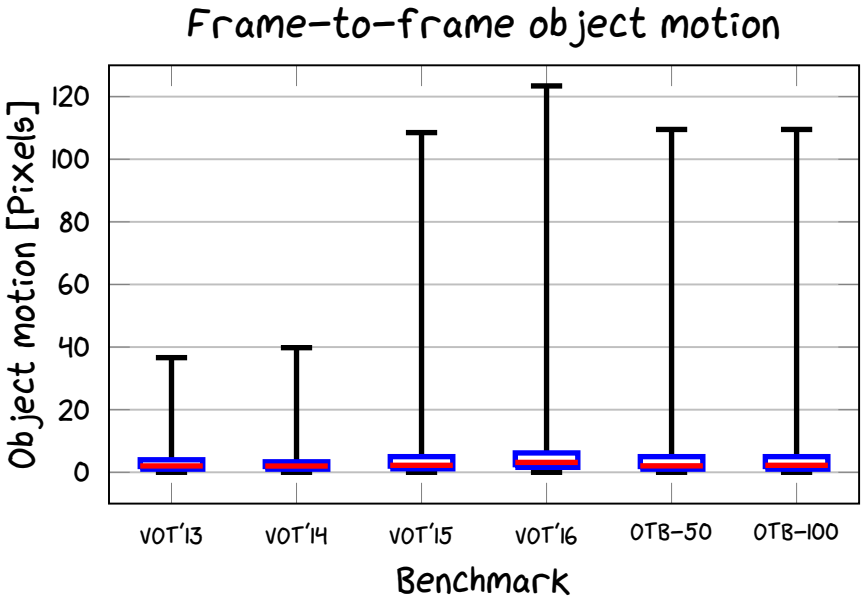


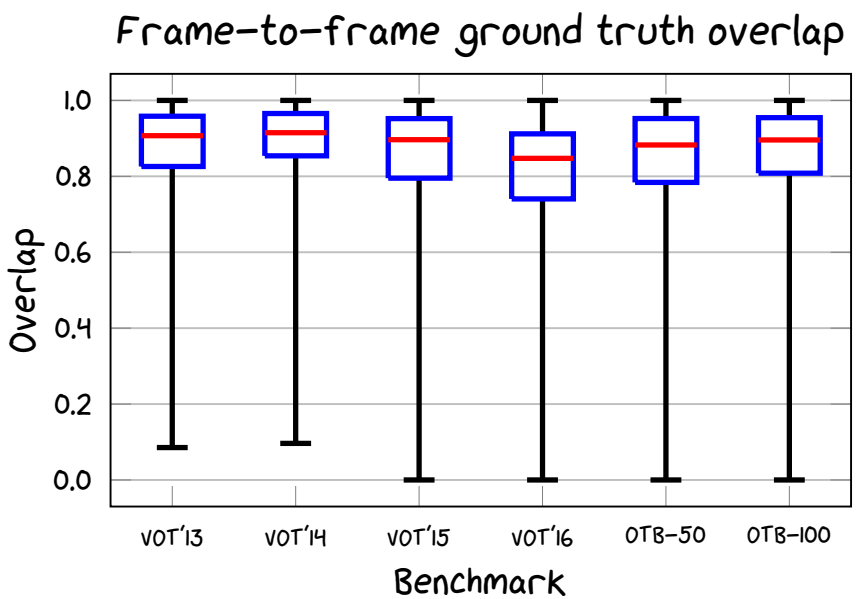
Evaluation

Benchmark Characteristics









Tracking Performance

Visual Object Tracking (VOT) Challenges

- Experiments:
  - Supervised.
    - Baseline.
    - Region noise.
  - Unsupervised.

Tracking Performance

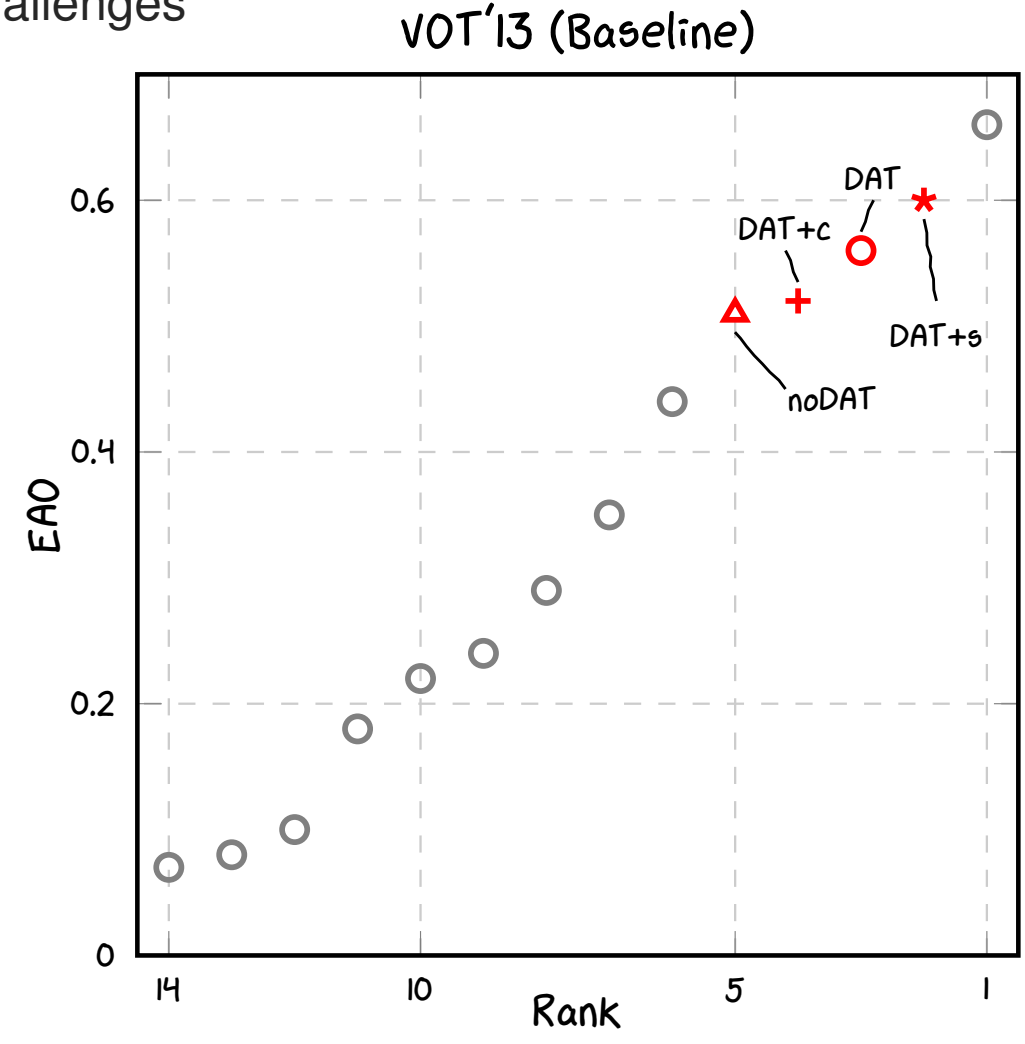
Visual Object Tracking (VOT) Challenges

- Experiments:
  - Supervised.
    - Baseline.
    - Region noise.
  - Unsupervised.
- Measures:
  - Accuracy $\in [0, 1]$.
  - Robustness $\in \mathbb{R}_0^+$.
  - Expected Average Overlap (EAO) $\in [0, 1]$.

Tracking Performance

Visual Object Tracking (VOT) Challenges

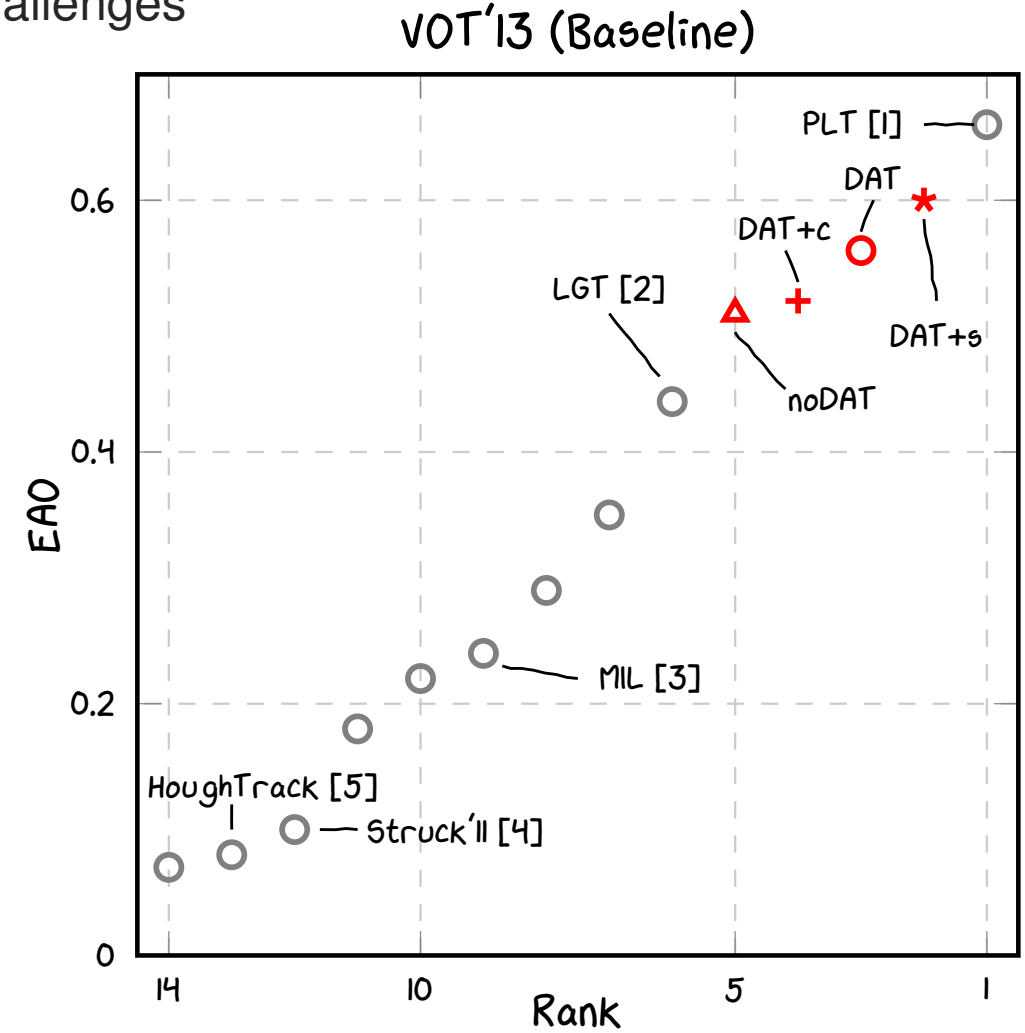
- Experiments:
  - Supervised.
    - Baseline.
    - Region noise.
  - Unsupervised.
- Measures:
  - Accuracy $\in [0, 1]$.
  - Robustness $\in \mathbb{R}_0^+$.
  - **Expected Average Overlap (EAO)** $\in [0, 1]$.



Tracking Performance

Visual Object Tracking (VOT) Challenges

- Experiments:
  - Supervised.
    - Baseline.
    - Region noise.
  - Unsupervised.
- Measures:
  - Accuracy $\in [0, 1]$.
  - Robustness $\in \mathbb{R}_0^+$.
  - **Expected Average Overlap (EAO)** $\in [0, 1]$.



[1] Heng *et al. Single Scale Pixel based LUT Tracker*. Presented at VOT'13.

[2] Čehovin *et al. Robust Visual Tracking using an Adaptive Coupled-layer Visual Model*. TPAMI 35(4), 2013.

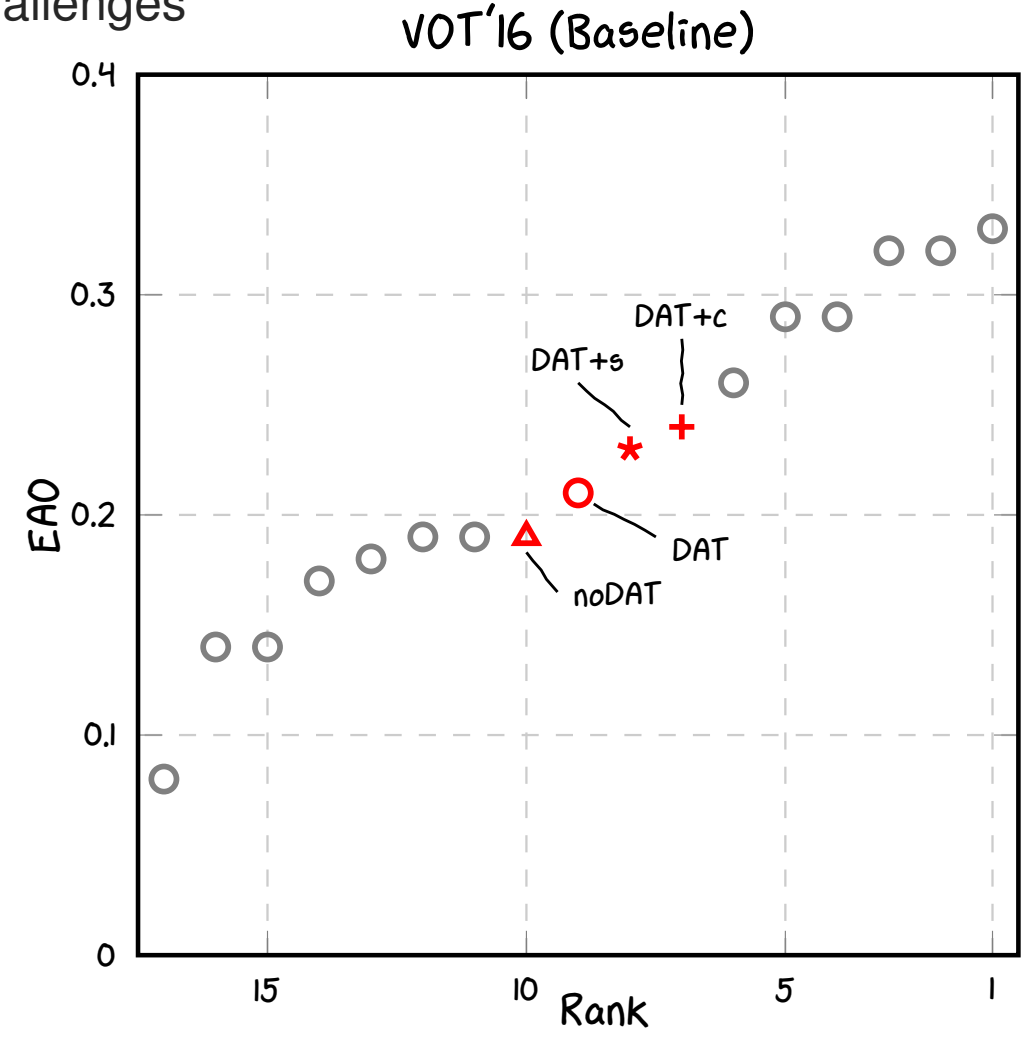
[3] Babenko *et al. Robust Object Tracking with Online Multiple Instance Learning*. TPAMI 33(7), 2011.

[4] Hare *et al. Struck: Structured Output Tracking with Kernels*. ICCV'11.

[5] Godec *et al. Hough-based Tracking of Non-rigid Objects*. CVIU 117(10), 2013.

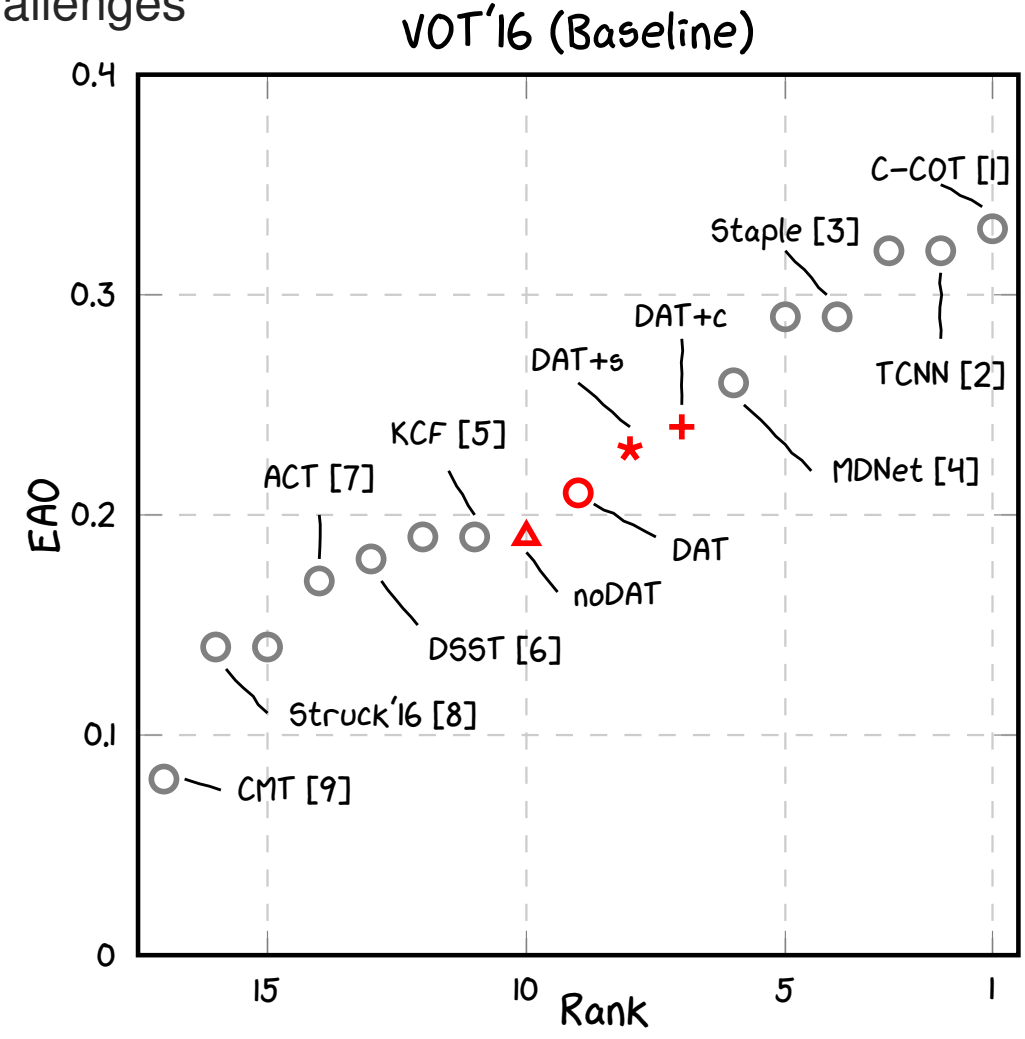
Tracking Performance

Visual Object Tracking (VOT) Challenges



Tracking Performance

Visual Object Tracking (VOT) Challenges



[1] Danelljan *et al. Beyond Correlation Filters: Learning Continuous Convolution Operators for Visual Tracking*. ECCV'16.

[2] Nam *et al. Modeling and Propagating CNNs in a Tree Structure for Visual Tracking*. arXiv abs/1704.06326, 2017.

[3] Bertinetto *et al. Staple: Complementary Learners for Real-Time Tracking*. CVPR'16.

[4] Nam and Han. *Learning Multi-Domain Convolutional Neural Networks for Visual Tracking*. CVPR'16.

[5] Henriques *et al. High-Speed Tracking with Kernelized Correlation Filters*. TPAMI 37(3), 2015.

[6] Danelljan *et al. Accurate Scale Estimation for Robust Visual Tracking*. BMVC'14.

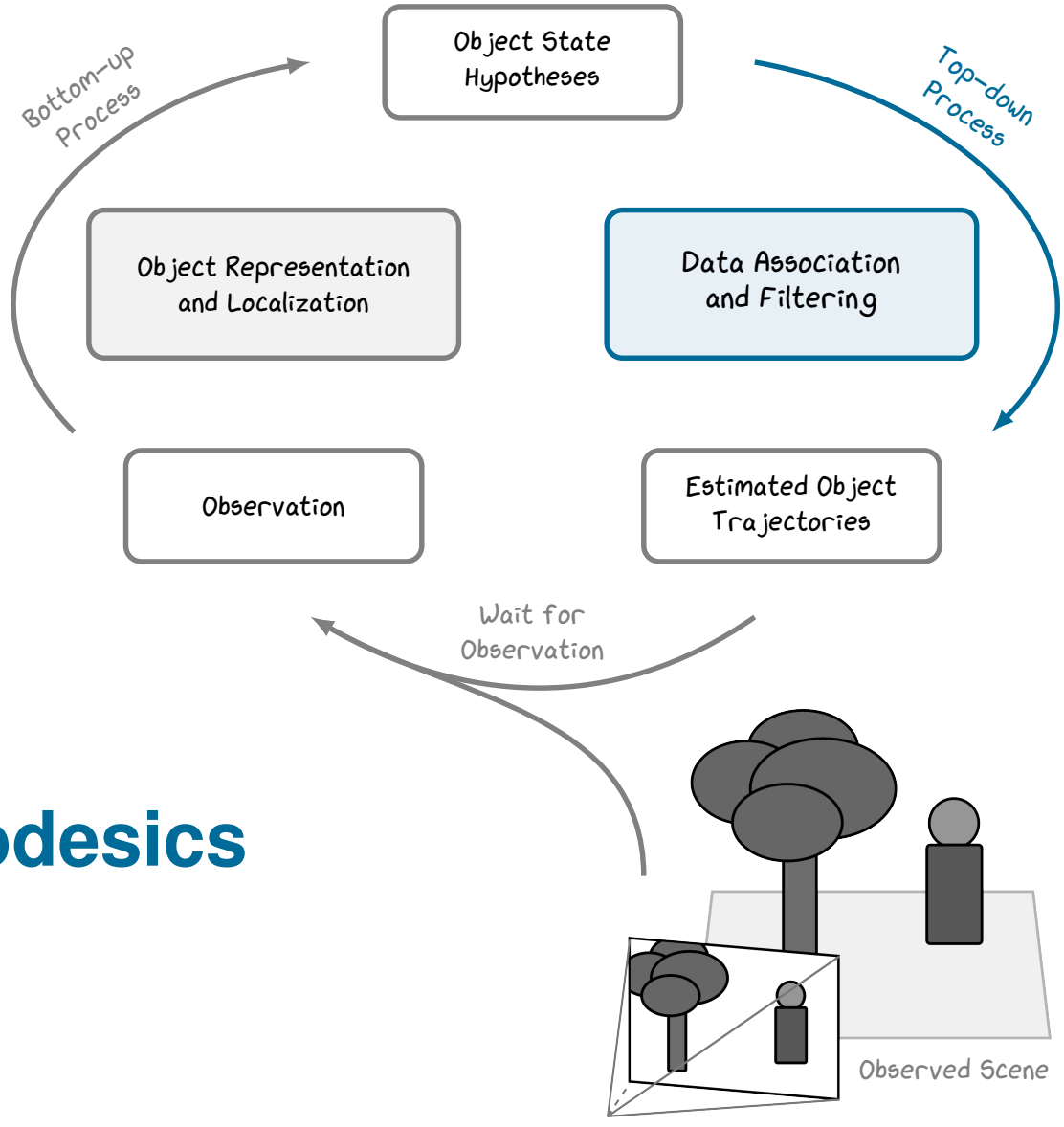
[7] Danelljan *et al. Adaptive Color Attributes for Real-Time Visual Tracking*. CVPR'14.

[8] Hare *et al. Struck: Structured Output Tracking with Kernels*. TPAMI 38(10), 2015.

[9] Nebehay and Pflugfelder. *Clustering of Static-Adaptive Correspondences for Deformable Object Tracking*. CVPR'15.

Exemplary Application

Outdoor Sports (DAT+s)

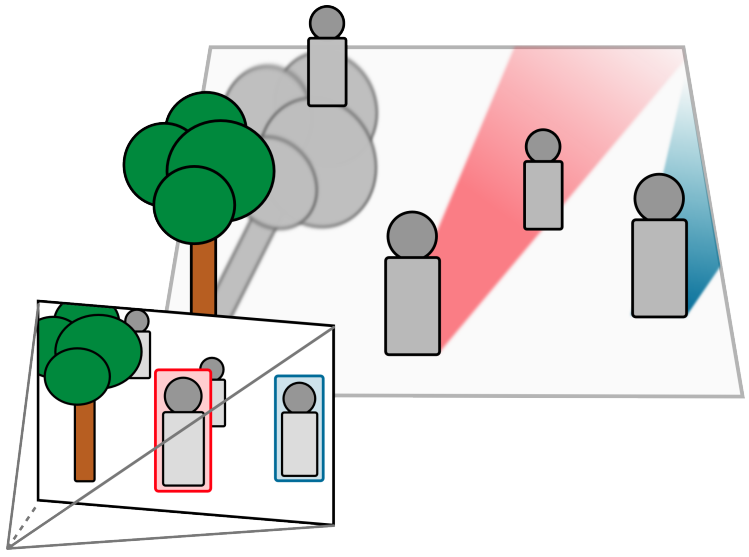


Occlusion Geodesics

Data Association

Exploiting Scene Context

- Closed-world assumptions [1,2].

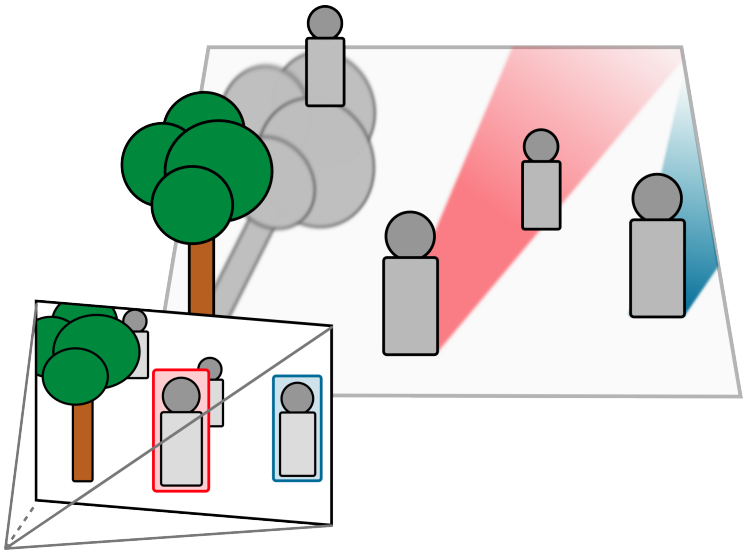


[1] Intille and Bobick. *Visual Tracking Using Closed-Worlds.* ICCV'95.

[2] Kristan *et al. Closed-World Tracking of Multiple Interacting Targets for Indoor-Sports Applications.* CVIU 113(5), 2009.

Exploiting Scene Context

- Closed-world assumptions [1,2].
- Objects fully visible:
  - Conservative tracking-by-detection.
  - Bipartite graph matching [3]:



$$\mathbf{A}^* = \arg\min_{\mathbf{A}} \sum_i \sum_j \underbrace{\psi_{i,j}^{(t)}}_{\text{Cost term}} \overbrace{a_{i,j}^{(t)}}^{\text{Binary assignment}},$$

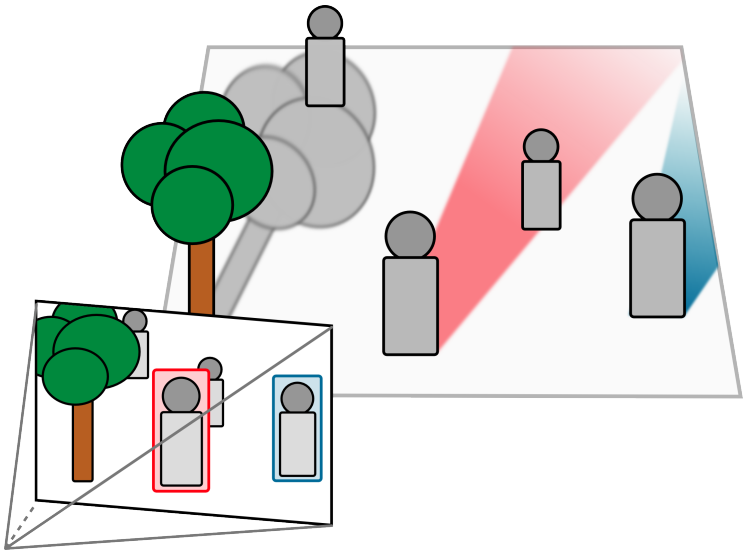
$$\text{s.t.} \sum_i a_{i,j}^{(t)} = 1, \quad \text{and} \quad \sum_j a_{i,j}^{(t)} = 1, \quad \text{with} \quad a_{i,j}^{(t)} \in \{0,1\}.$$

[1] Intille and Bobick. *Visual Tracking Using Closed-Worlds.* ICCV'95.

[2] Kristan *et al. Closed-World Tracking of Multiple Interacting Targets for Indoor-Sports Applications.* CVIU 113(5), 2009.

[3] Munkres. *Algorithms for the Assignment and Transportation Problems.* SIAM J. Appl. Math, 5(1), 1957.

Exploiting Scene Context



- Closed-world assumptions [1,2].
- Objects fully visible:
  - Conservative tracking-by-detection.
  - Bipartite graph matching [3]:

$$\mathbf{A}^* = \arg\min_{\mathbf{A}} \sum_i \sum_j \underbrace{\psi_{i,j}^{(t)}}_{\text{Cost term}} \overbrace{a_{i,j}^{(t)}}^{\text{Binary assignment}},$$

$$\text{s.t.} \sum_i a_{i,j}^{(t)} = 1, \quad \text{and} \quad \sum_j a_{i,j}^{(t)} = 1, \quad \text{with} \quad a_{i,j}^{(t)} \in \{0,1\}.$$

- Occluded/missed objects:
  - Let objects move within occluded regions.
  - Re-assignment based on occlusion geodesics (shortest paths).

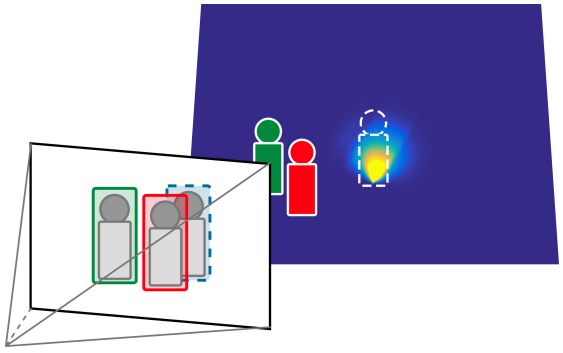
[1] Intille and Bobick. *Visual Tracking Using Closed-Worlds.* ICCV'95.

[2] Kristan *et al. Closed-World Tracking of Multiple Interacting Targets for Indoor-Sports Applications.* CVIU 113(5), 2009.

[3] Munkres. *Algorithms for the Assignment and Transportation Problems.* SIAM J. Appl. Math, 5(1), 1957.

Occlusion Geodesics

- Instance-specific confidence score at each time step:



$$\varphi_i^{(\delta_i)}(\mathbf{x}) = \underbrace{c_{o,i}^{(\delta_i)}(\mathbf{x})}_{\text{Occlusion term}} \overbrace{c_{p,i}^{(\delta_i)}(\mathbf{x})}^{\text{Plausible movement}} \underbrace{c_{d,i}^{(\delta_i)}(\mathbf{x})}_{\text{Directional similarity}}$$

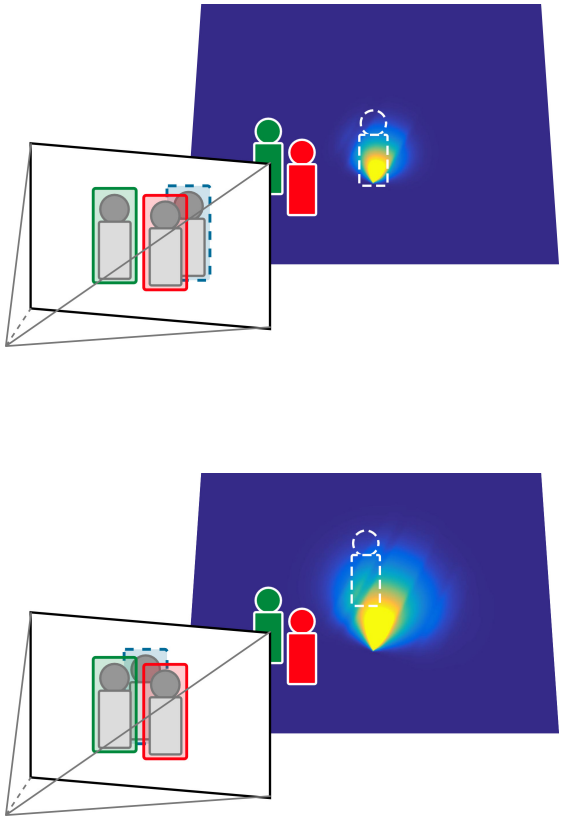
Occlusion Geodesics

- Instance-specific confidence score at each time step:

$$\varphi_i^{(\delta_i)}(\mathbf{x}) = \underbrace{c_{o,i}^{(\delta_i)}(\mathbf{x})}_{\text{Occlusion term}} \overbrace{c_{p,i}^{(\delta_i)}(\mathbf{x})}^{\text{Plausible movement}} \underbrace{c_{d,i}^{(\delta_i)}(\mathbf{x})}_{\text{Directional similarity}}$$

- Weight physically plausible paths:

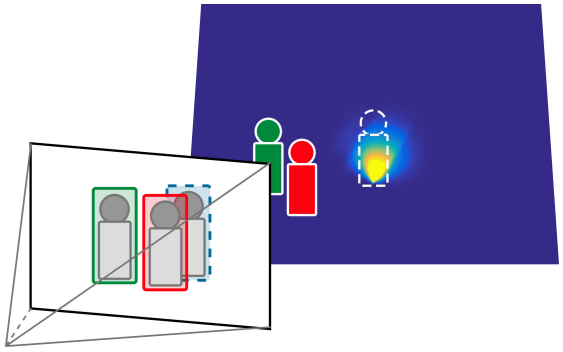
$$\Psi_i^{(\delta_i)}(\mathbf{x}) = 1 - \varphi_i^{(\delta_i)}(\mathbf{x}) + \underbrace{\inf_{\mathbf{z}} \Psi_i^{(\delta_i - 1)}(\mathbf{x}+\mathbf{z})}_{\text{Reachable neighborhood}} .$$



Occlusion Geodesics

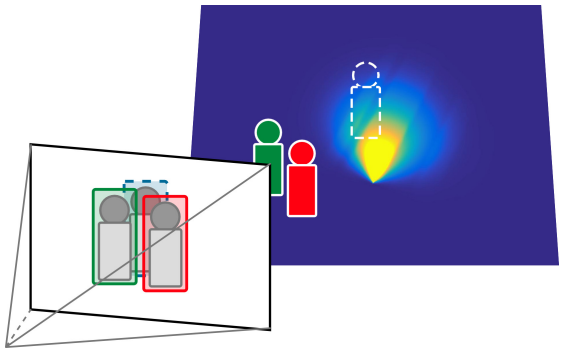
- Instance-specific confidence score at each time step:

$$\varphi_i^{(\delta_i)}(\mathbf{x}) = \underbrace{c_{o,i}^{(\delta_i)}(\mathbf{x})}_{\text{Occlusion term}} \overbrace{c_{p,i}^{(\delta_i)}(\mathbf{x})}^{\text{Plausible movement}} \underbrace{c_{d,i}^{(\delta_i)}(\mathbf{x})}_{\text{Directional similarity}}$$

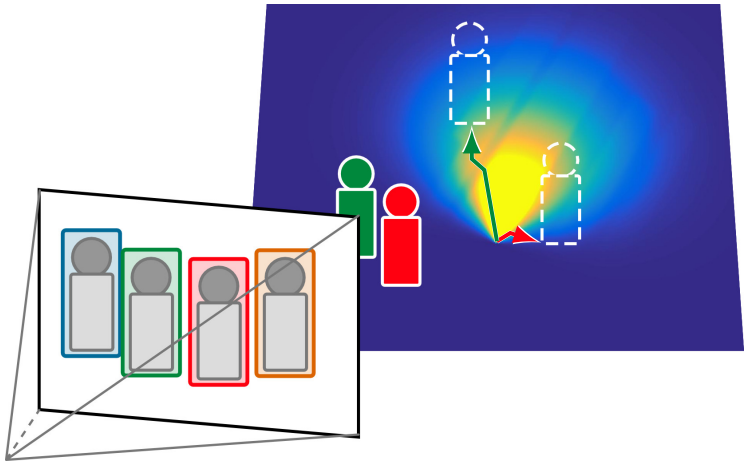


- Weight physically plausible paths:

$$\Psi_i^{(\delta_i)}(\mathbf{x}) = 1 - \varphi_i^{(\delta_i)}(\mathbf{x}) + \underbrace{\inf_{\mathbf{z}} \Psi_i^{(\delta_i - 1)}(\mathbf{x}+\mathbf{z})}_{\text{Reachable neighborhood}} .$$



- Re-assignment if feasible path exists.
- No need to compute the actual (hidden) trajectory – only costs/plausibility.



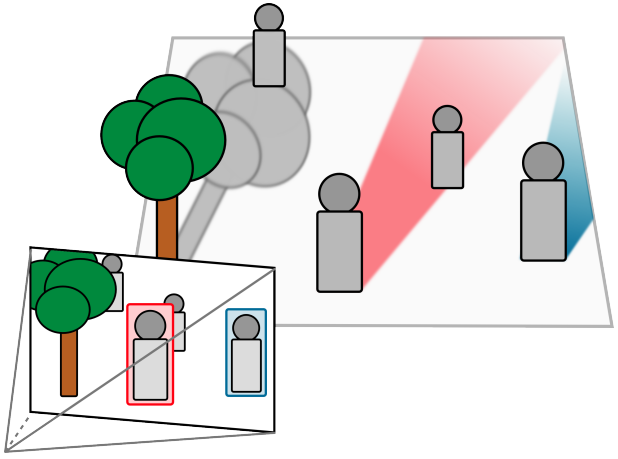
Confidence Scores

- Occlusion-based confidence:
  More likely to be occluded than to be missed.

Occlusion length

$$c_{o,i}^{(\delta_i)}(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{x} \in \mathcal{P}_{\text{static}} \cup \mathcal{P}_{\text{dynamic}}^{(t)} \\ 1 - \beta^{\delta_i} & \text{otherwise.} \end{cases}$$

Detector “belief“



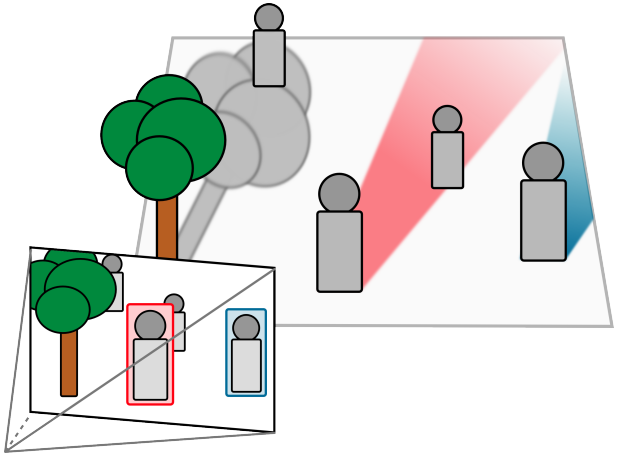
Confidence Scores

- Occlusion-based confidence:
  More likely to be occluded than to be missed.

Occlusion length

$$c_{o,i}^{(\delta_i)}(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{x} \in \mathcal{P}_{\text{static}} \cup \mathcal{P}_{\text{dynamic}}^{(t)} \\ 1 - \beta^{\delta_i} & \text{otherwise.} \end{cases}$$

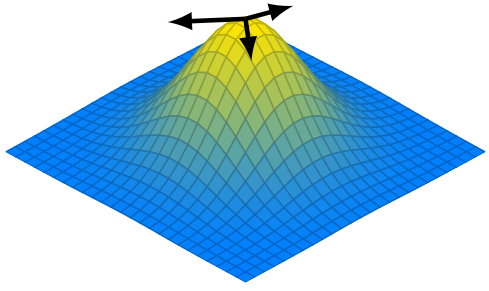
Detector "belief"



- Plausible motion confidence:

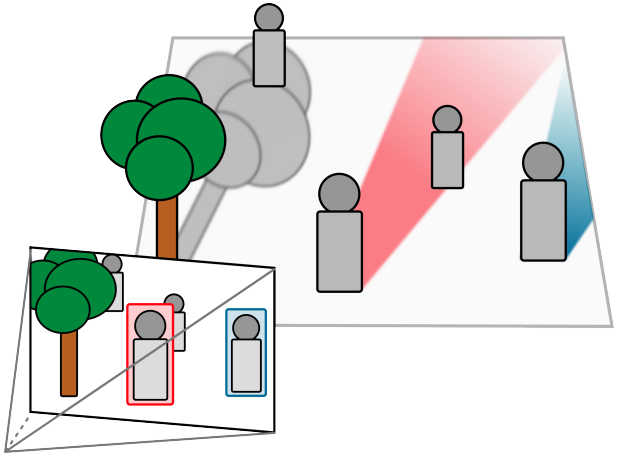
$$c_{p,i}^{(\delta_i)}(\mathbf{x}) = \exp\left(- \frac{\|\mathbf{x} - \hat{\mathbf{x}}_i\|_2^2}{2\,\sigma_p^2\,\delta_i^2\,\max\left(\|\hat{\mathbf{d}}_i\|_2, v_{\text{avg}}\right)^2} \right).$$

Predicted motion direction



Confidence Scores

- Occlusion-based confidence:
  More likely to be occluded than to be missed.

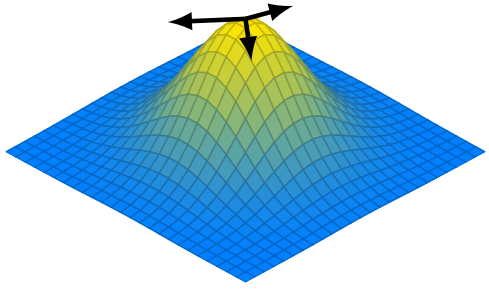


Occlusion length

$$c_{o,i}^{(\delta_i)}(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{x} \in \mathcal{P}_{\text{static}} \cup \mathcal{P}_{\text{dynamic}}^{(t)} \\ 1 - \beta^{\delta_i} & \text{otherwise.} \end{cases}$$

Detector "belief"

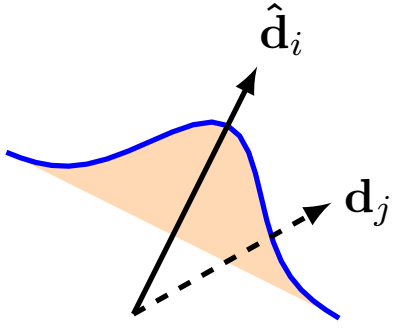
- Plausible motion confidence:



$$c_{p,i}^{(\delta_i)}(\mathbf{x}) = \exp\left(-\frac{\|\mathbf{x} - \hat{\mathbf{x}}_i\|_2^2}{2\,\sigma_p^2\,\delta_i^2\,\max\left(\|\hat{\mathbf{d}}_i\|_2, v_{\text{avg}}\right)^2} \right).$$

Predicted motion direction

- Directional confidence:

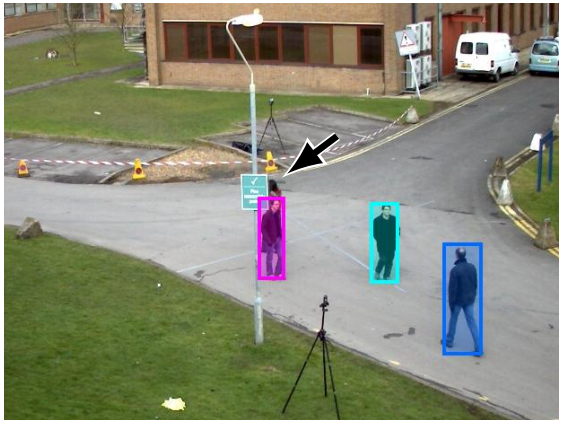


$$c_{d,i}^{(\delta_i)}(\mathbf{x}) = \exp\left(-\frac{\left(\left\langle \hat{\mathbf{d}}_i, \mathbf{d}_j \right\rangle - \|\hat{\mathbf{d}}_i\|\|\mathbf{d}_j\|\right)^2}{2\,\sigma_d^2\,\|\hat{\mathbf{d}}_i\|^2\,\|\mathbf{d}_j\|^2} \right).$$

Re-assignment Example

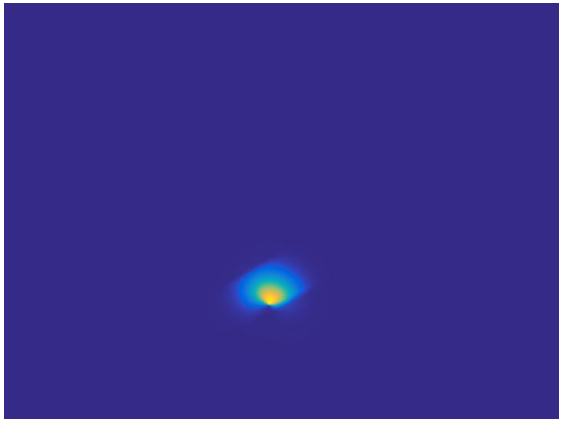
PETS'09 S2L1, step 1/4

Camera view.



Groundplane overlay.

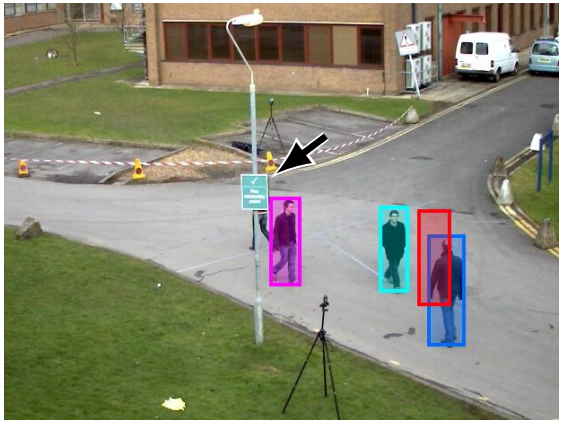
Confidence map (cropped).



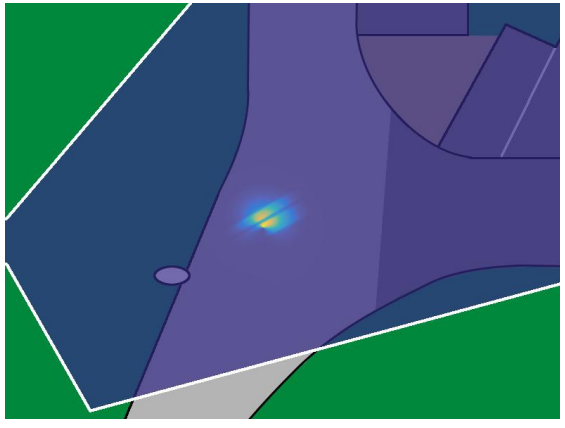
Re-assignment Example

PETS'09 S2L1, step 2/4

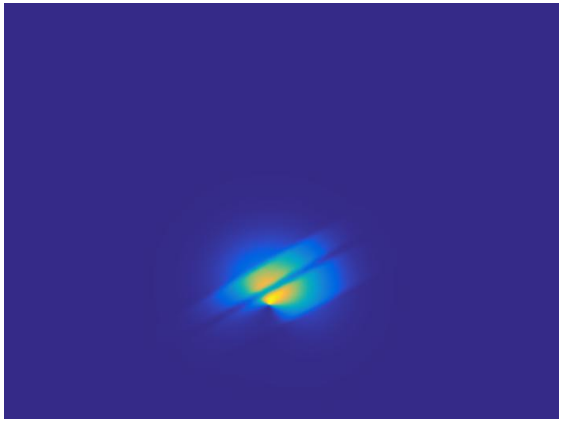
Camera view.



Groundplane overlay.



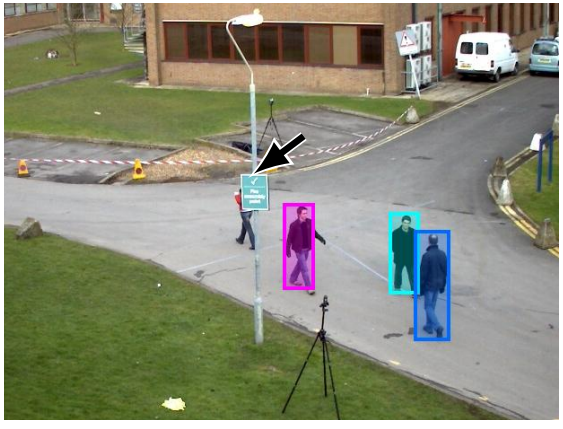
Confidence map (cropped).



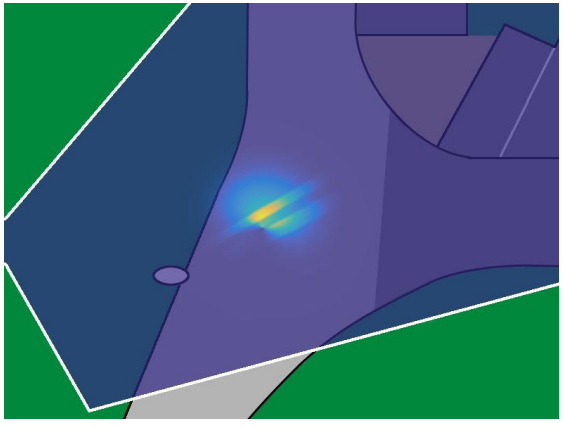
Re-assignment Example

PETS'09 S2L1, step 3/4

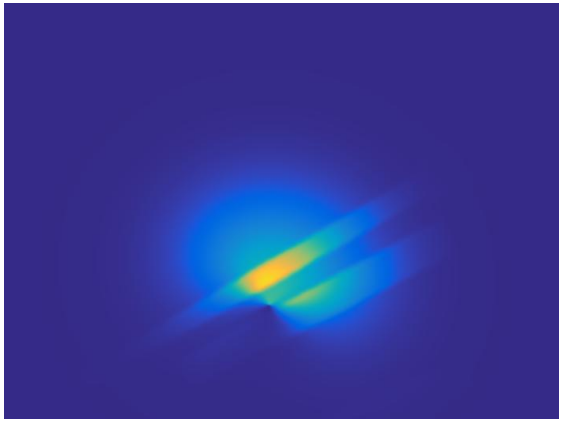
Camera view.



Groundplane overlay.



Confidence map (cropped).

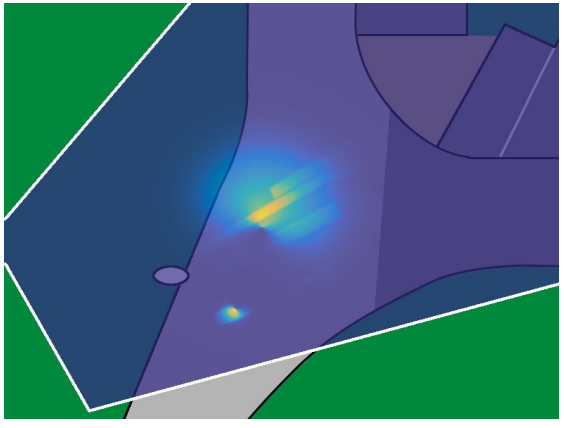


Re-assignment Example

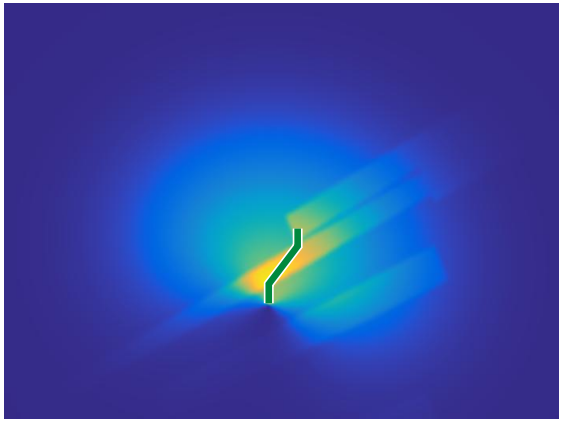
PETS'09 S2L1, step 4/4

Camera view.

Groundplane overlay.

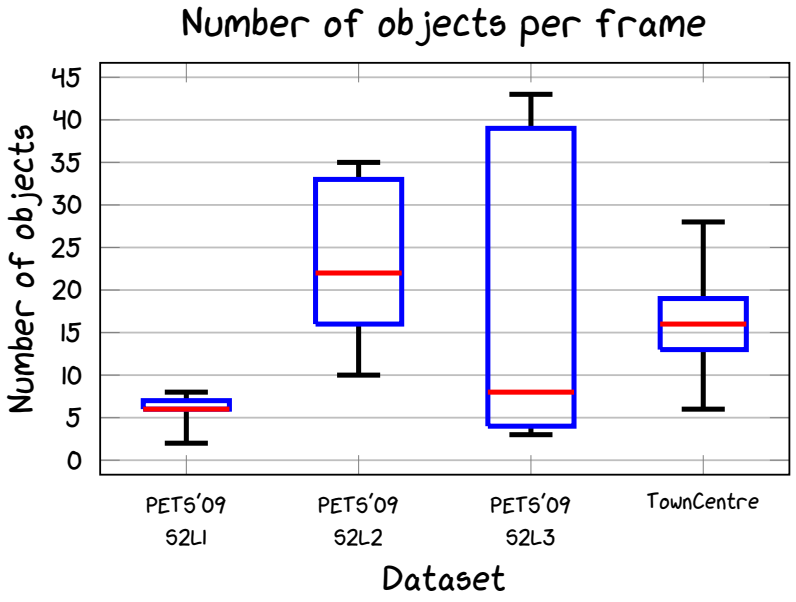


Confidence map (cropped).



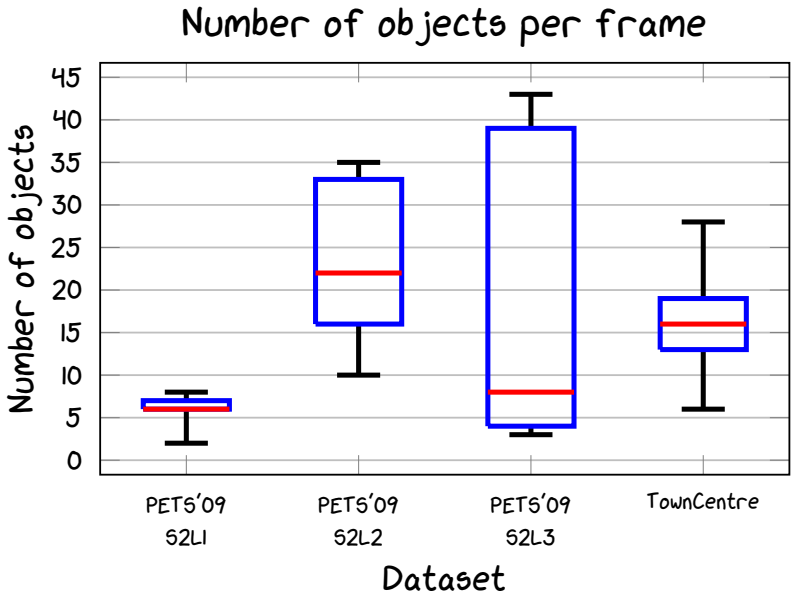
Evaluation

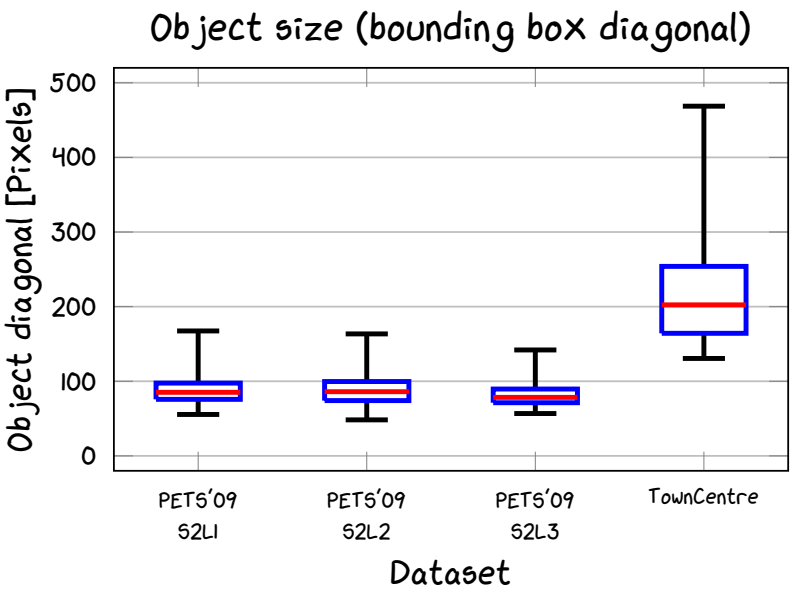
Sequence Characteristics



Evaluation

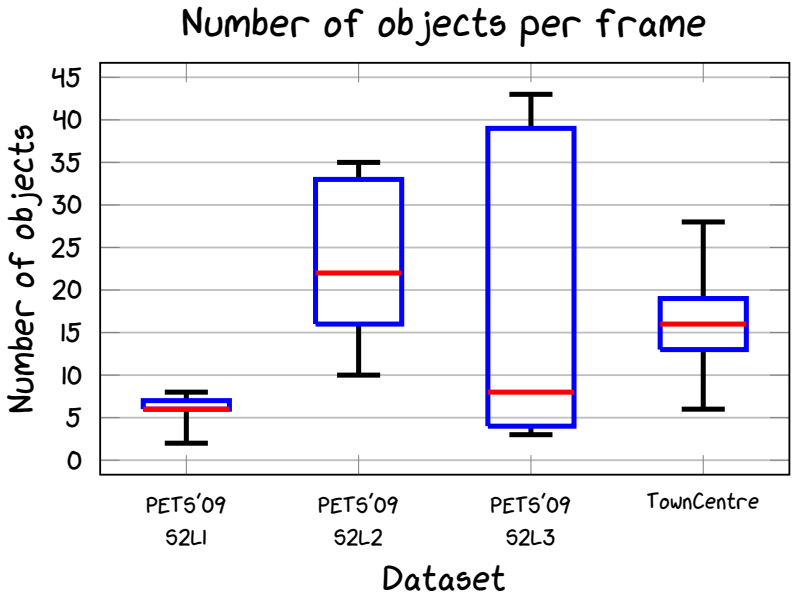
Sequence Characteristics

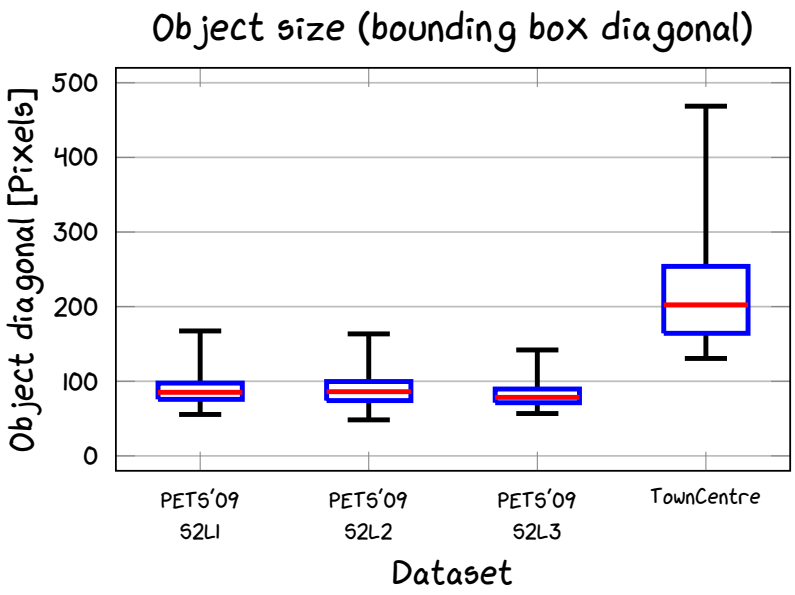


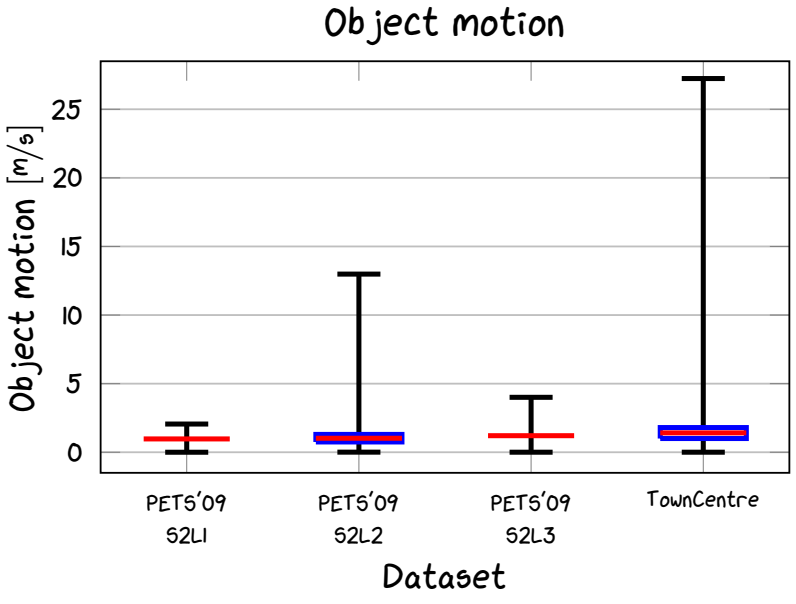


Evaluation

Sequence Characteristics

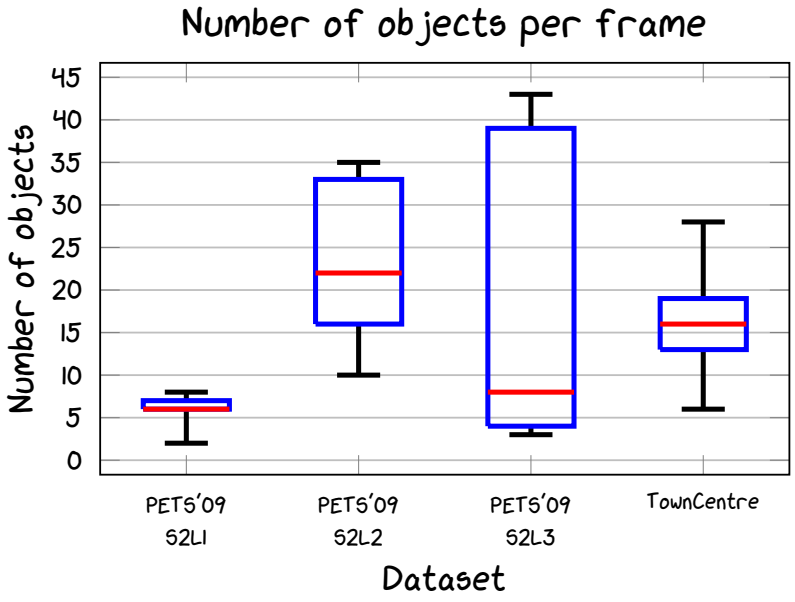


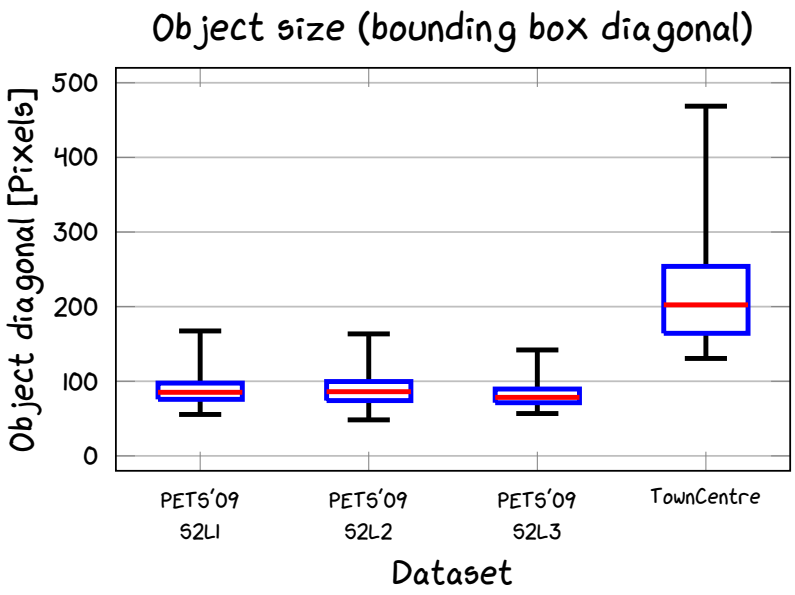


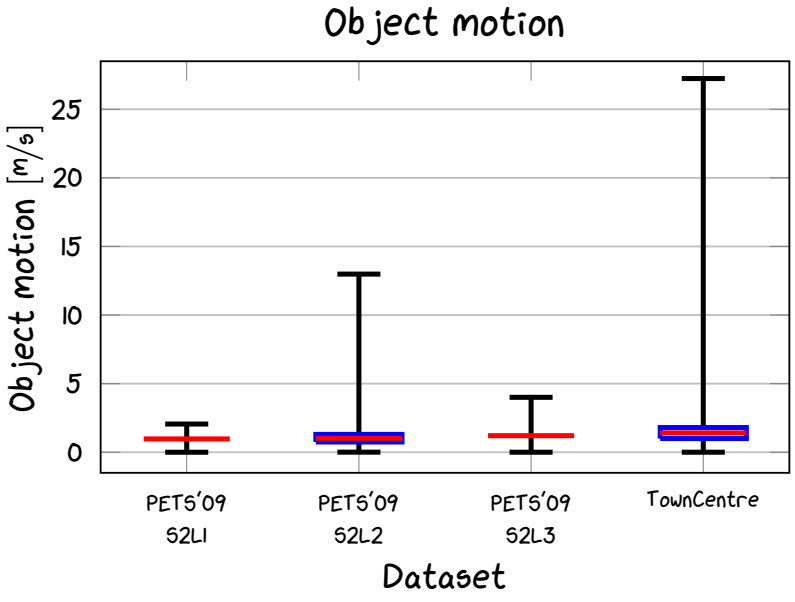


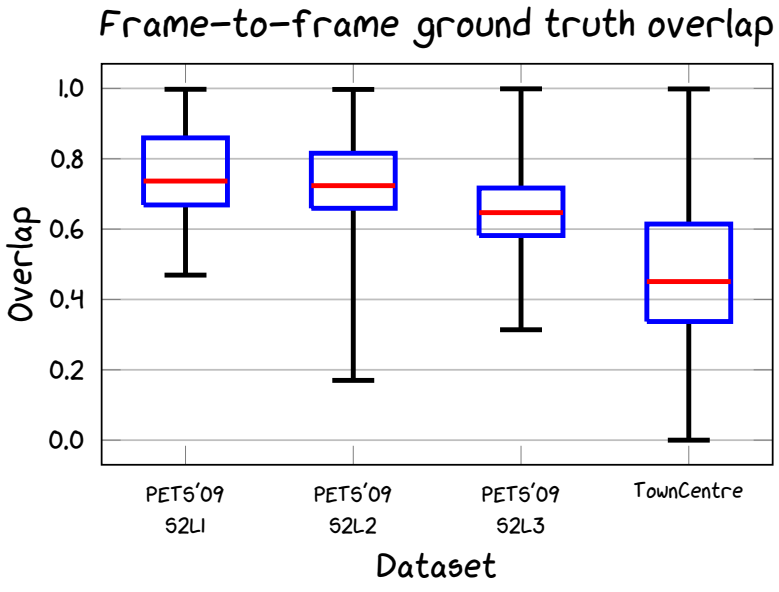
Evaluation

Sequence Characteristics









Tracking Performance

MOT'15 3D Challenge (PETS'09 S2L2 & TownCentre)

Highlights: Best, second best, third best.

Tracker	A	C
OccGeo (DPM)		✓
OccGeo (3D MOT'15)		✓
GPR-DBN [1]	✓	✓
LP-SFM [2]		
STV [3]	✓	
LP-3D [4]		
S-RNN [5]	✓	✓
K-SFM [6]		✓

- **A** explicitly models appearance.
- **C** indicates causal trackers.

[1] Klinger *et al. Probabilistic Multi-Person Localisation and Tracking in Image Sequences*. ISPRS JPRS 127, 2017.

[2] Leal-Taixé *et al. Everybody needs somebody: Modeling social and grouping behavior on a linear programming multiple people tracker*. ICCVW'11.

[3] Wen *et al. Multi-Camera Multi-Target Tracking with Space-Time-View Hyper-graph*. IJCV 122(2), 2017.

[4] Leal-Taixé *et al. Learning an Image-based Motion Context for Multiple People Tracking*. CVPR'14.

[5] Sadeghian *et al. Tracking the Untrackable: Learning to Track Multiple Cues with Long-Term Dependencies*. ICCV'17.

[6] Pellegrini *et al. You'll Never Walk Alone: Modeling Social Behavior for Multi-target Tracking*. ICCV'09.

Tracking Performance

MOT'15 3D Challenge (PETS'09 S2L2 & TownCentre)

Highlights: Best, second best, third best.

Tracker	A	C	MOTA↑	MOTP↑
OccGeo (DPM)		✓	0.51	0.62
OccGeo (3D MOT'15)		✓	0.31	0.59
GPR-DBN [1]	✓	✓	0.48	0.62
LP-SFM [2]			0.31	0.52
STV [3]	✓		0.31	0.55
LP-3D [4]			0.30	0.52
S-RNN [5]	✓	✓	0.22	0.54
K-SFM [6]		✓	0.21	0.52

- **A** explicitly models appearance.
- **C** indicates causal trackers.

[1] Klinger *et al. Probabilistic Multi-Person Localisation and Tracking in Image Sequences*. ISPRS JPRS 127, 2017.

[2] Leal-Taixé *et al. Everybody needs somebody: Modeling social and grouping behavior on a linear programming multiple people tracker*. ICCVW'11.

[3] Wen *et al. Multi-Camera Multi-Target Tracking with Space-Time-View Hyper-graph*. IJCV 122(2), 2017.

[4] Leal-Taixé *et al. Learning an Image-based Motion Context for Multiple People Tracking*. CVPR'14.

[5] Sadeghian *et al. Tracking the Untrackable: Learning to Track Multiple Cues with Long-Term Dependencies*. ICCV'17.

[6] Pellegrini *et al. You'll Never Walk Alone: Modeling Social Behavior for Multi-target Tracking*. ICCV'09.

Tracking Performance

MOT'15 3D Challenge (PETS'09 S2L2 & TownCentre)

Highlights: Best, second best, third best.

Tracker	A	C	MOTA↑	MOTP↑	MT/GT↑	ML/GT↓	IDs↓	FM↓
OccGeo (DPM)		✓	0.51	0.62	0.26	0.24	350	370
OccGeo (3D MOT'15)		✓	0.31	0.59	0.16	0.32	414	411
GPR-DBN [1]	✓	✓	0.48	0.62	0.33	0.21	181	270
LP-SFM [2]			0.31	0.52	0.16	0.22	396	467
STV [3]	✓		0.31	0.55	0.14	0.25	383	439
LP-3D [4]			0.30	0.52	0.24	0.14	487	542
S-RNN [5]	✓	✓	0.22	0.54	0.03	0.36	785	1053
K-SFM [6]		✓	0.21	0.52	0.07	0.14	1463	1322

- **A** explicitly models appearance.
- **C** indicates causal trackers.

[1] Klinger *et al. Probabilistic Multi-Person Localisation and Tracking in Image Sequences*. ISPRS JPRS 127, 2017.

[2] Leal-Taixé *et al. Everybody needs somebody: Modeling social and grouping behavior on a linear programming multiple people tracker*. ICCVW'11.

[3] Wen *et al. Multi-Camera Multi-Target Tracking with Space-Time-View Hyper-graph*. IJCV 122(2), 2017.

[4] Leal-Taixé *et al. Learning an Image-based Motion Context for Multiple People Tracking*. CVPR'14.

[5] Sadeghian *et al. Tracking the Untrackable: Learning to Track Multiple Cues with Long-Term Dependencies*. ICCV'17.

[6] Pellegrini *et al. You'll Never Walk Alone: Modeling Social Behavior for Multi-target Tracking*. ICCV'09.

Tracking Performance

MOT'15 3D Challenge (PETS'09 S2L2 & TownCentre)

Highlights: Best, second best, third best.

Tracker	A	C	MOTA↑	MOTP↑	MT/GT↑	ML/GT↓	IDs↓	FM↓	FPS↑
OccGeo (DPM)		✓	0.51	0.62	0.26	0.24	350	370	7.5
OccGeo (3D MOT'15)		✓	0.31	0.59	0.16	0.32	414	411	4.8
GPR-DBN [1]	✓	✓	0.48	0.62	0.33	0.21	181	270	0.1
LP-SFM [2]			0.31	0.52	0.16	0.22	396	467	8.4
STV [3]	✓		0.31	0.55	0.14	0.25	383	439	1.9
LP-3D [4]			0.30	0.52	0.24	0.14	487	542	83.5
S-RNN [5]	✓	✓	0.22	0.54	0.03	0.36	785	1053	1.2
K-SFM [6]		✓	0.21	0.52	0.07	0.14	1463	1322	30.6

- **A** explicitly models appearance.
- **C** indicates causal trackers.

[1] Klinger *et al. Probabilistic Multi-Person Localisation and Tracking in Image Sequences*. ISPRS JPRS 127, 2017.

[2] Leal-Taixé *et al. Everybody needs somebody: Modeling social and grouping behavior on a linear programming multiple people tracker*. ICCVW'11.

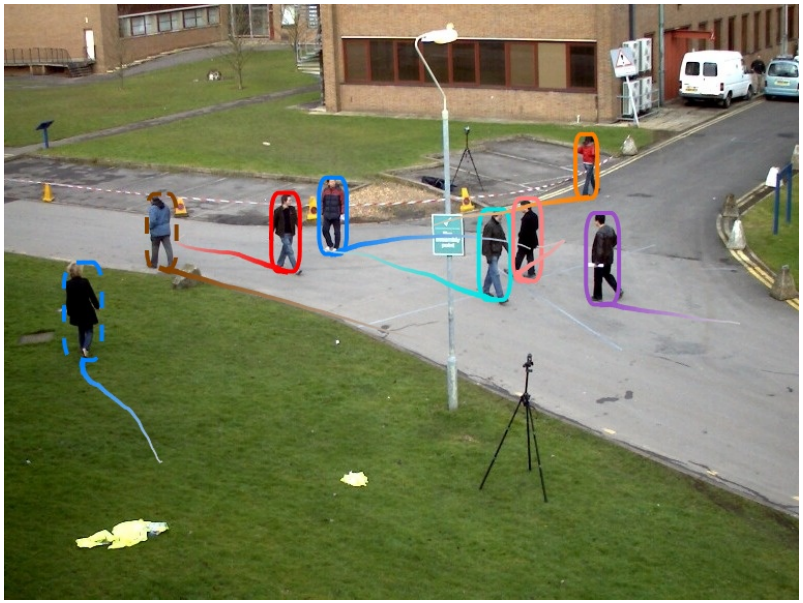
[3] Wen *et al. Multi-Camera Multi-Target Tracking with Space-Time-View Hyper-graph*. IJCV 122(2), 2017.

[4] Leal-Taixé *et al. Learning an Image-based Motion Context for Multiple People Tracking*. CVPR'14.

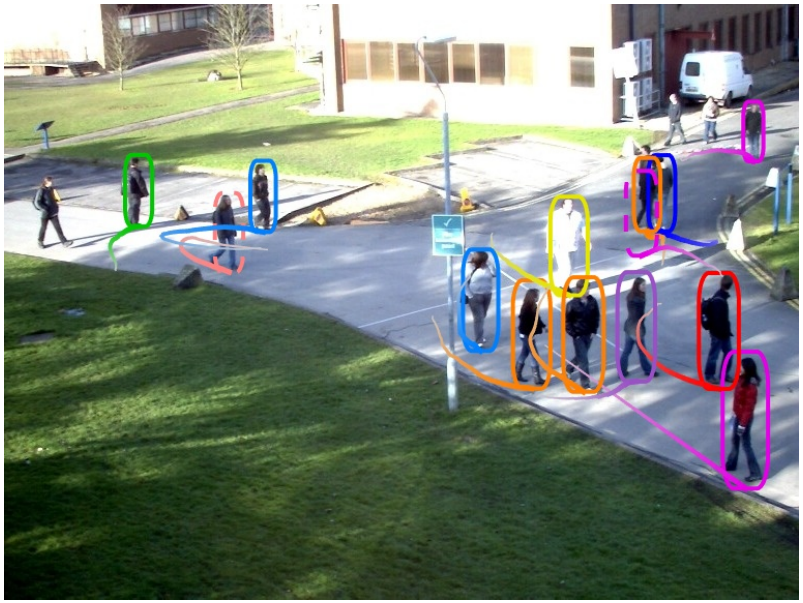
[5] Sadeghian *et al. Tracking the Untrackable: Learning to Track Multiple Cues with Long-Term Dependencies*. ICCV'17.

[6] Pellegrini *et al. You'll Never Walk Alone: Modeling Social Behavior for Multi-target Tracking*. ICCV'09.

Exemplary Results



PETS'09 S2L1.



PETS'09 S2L2.

TownCentre.

TownCentre.

Conclusion

- Take away:
  - Anticipate the unexpected.

    Investigating your requirements pays off.
  - Garbage in, garbage out.

Conclusion

- Take away:
  - Anticipate the unexpected.

    Investigating your requirements pays off.
  - Garbage in, garbage out.

- Contextual cues:
  - Often overlooked, but powerful.
  - Allow for intuitive models.
  - Enable real-time visual tracking at competitive performance.

Conclusion

- Take away:
  - Anticipate the unexpected.

    Investigating your requirements pays off.
  - Garbage in, garbage out.

- Contextual cues:
  - Often overlooked, but powerful.
  - Allow for intuitive models.
  - Enable real-time visual tracking at competitive performance.

- We made tracking algorithms see*.

  * Simple, efficient, yet effective.

Publications

Conferences & Journals

- Possegger, Rüther, Sternig, Mauthner, Klopschitz, Roth & Bischof. *Unsupervised Calibration of Camera Networks and Virtual PTZ Cameras*, **CVWW**'12.
- Possegger, Sternig, Mauthner, Roth & Bischof. *Robust Real-Time Tracking of Multiple Objects by Volumetric Mass Densities*, **CVPR**'13.
- Possegger, Mauthner, Roth & Bischof. *Occlusion Geodesics for Online Multi-Object Tracking*, **CVPR**'14.
- Rudelsdorfer, Schrapf, Possegger, Mauthner, Bischof & Tilp. *A novel method for the analysis of sequential actions in team handball*, **IJCSS** 13(1):69–84, 2014.
- Possegger, Mauthner & Bischof. *In Defense of Color-based Model-free Tracking*, **CVPR**'15.
- Mauthner, Possegger, Waltner & Bischof. *Encoding based Saliency Detection for Videos and Images*, **CVPR**'15.
- Opitz, Waltner, Poier, Possegger & Bischof. *Grid Loss: Detecting Occluded Faces*, **ECCV**'16.
- Opitz, Possegger & Bischof. *Efficient Model Averaging for Deep Neural Networks*, **ACCV**'16.
- Ertler, Possegger, Opitz & Bischof. *Pedestrian Detection in RGB-D Images from an Elevated Viewpoint*, **CVWW**'17.
- Opitz, Waltner, Possegger & Bischof. *BIER - Boosting Independent Embeddings Robustly*, **ICCV**'17.
- Aytekin, Possegger, Mauthner, Kiranyaz, Bischof & Gabbouj. *Spatiotemporal Saliency Estimation by Spectral Foreground Detection*, **TMM** 20(1):82–95, 2018.

Publications

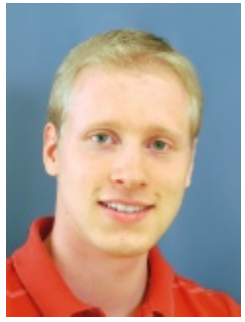
Visual Object Tracking Challenges

- Kristan *et al. The Visual Object Tracking VOT2014 Challenge Results*, **VOT**'14 (ECCV-WS).
  Appearance-Based Shape Filter (ABS)
- Kristan *et al. The Visual Object Tracking VOT2015 Challenge Results*, **VOT**'15 (ICCV-WS).
  Distractor-Aware Tracker (DAT)
- Kristan *et al. The Visual Object Tracking VOT2016 Challenge Results*, **VOT**'16 (ECCV-WS).
  Distractor-Aware Tracker (DAT)
- Felsberg *et al. The Thermal Infrared Visual Object Tracking VOT-TIR2016 Challenge Results*, **VOT**'16 (ECCV-WS).
  Distractor-Aware Tracker (DAT), monochrome

Special Thanks!



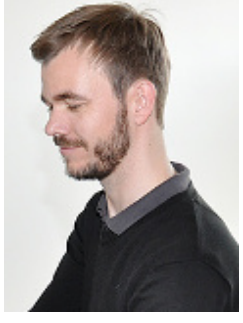










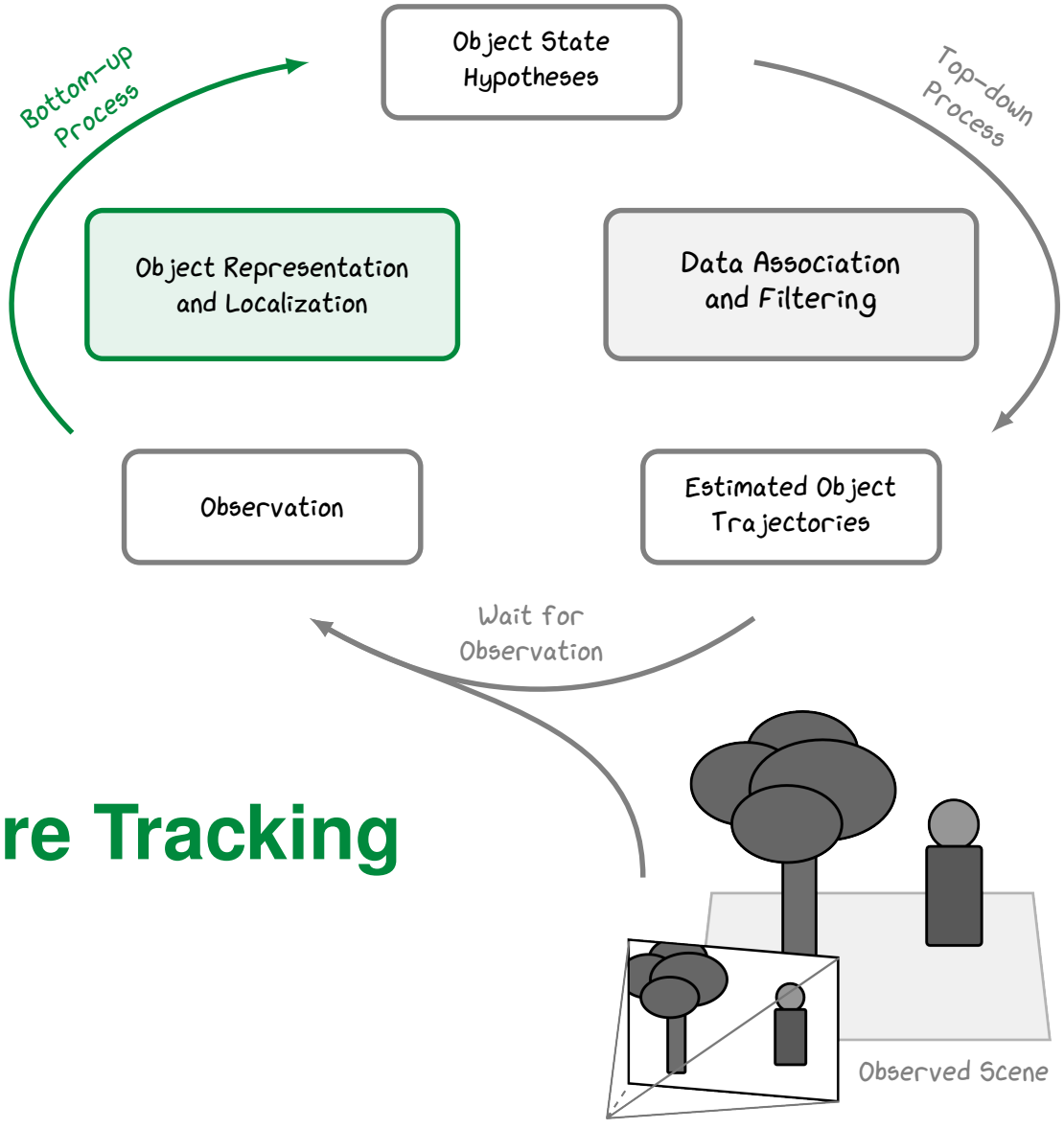


Thank You!

LEARNING, RECOGNITION & SURVEILLANCE

Key Challenges

- **Appearance variations**
  Non-rigid deformations, scale changes, illumination.
- **Dynamics**
  Target, scene and camera.
- **Illumination conditions**
  Easy to record low quality imagery, hard to get it right.
- **Occlusions**
  Obstacles, objects and self-occlusions (deformations).

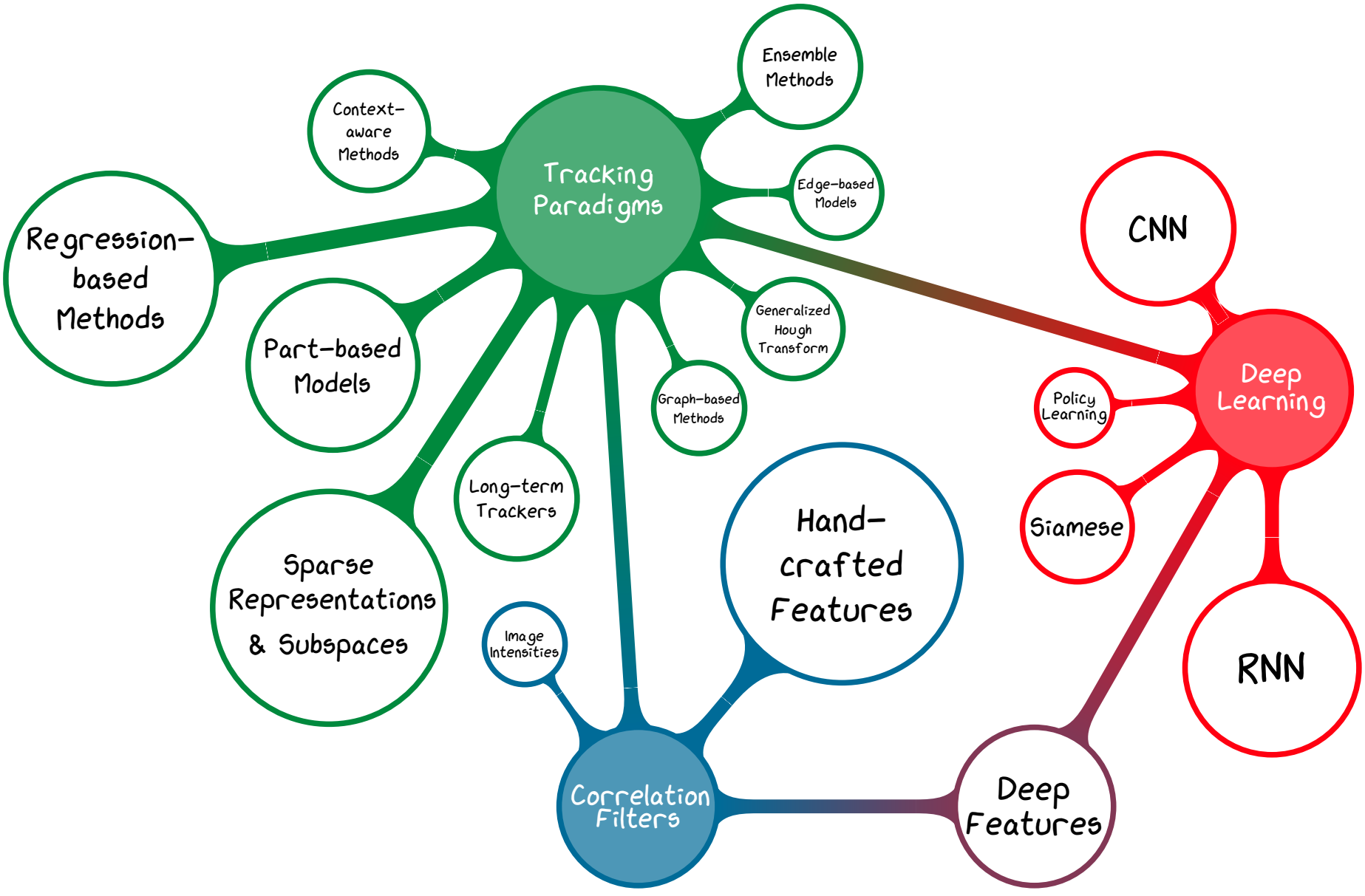


Distractor-aware Tracking

Details & Extras

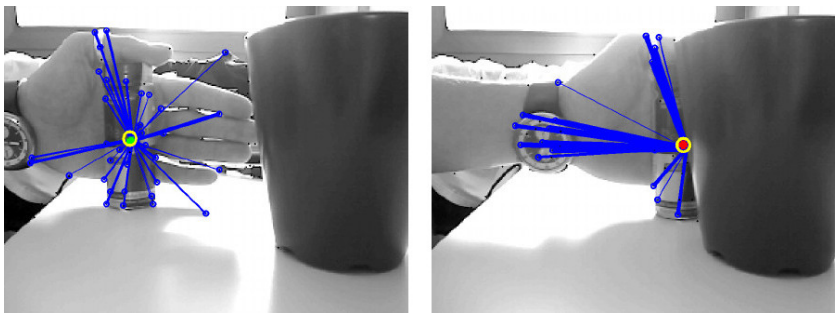
Single Object Tracking (SOT) Paradigms

Top-performing Approaches (2013-today)

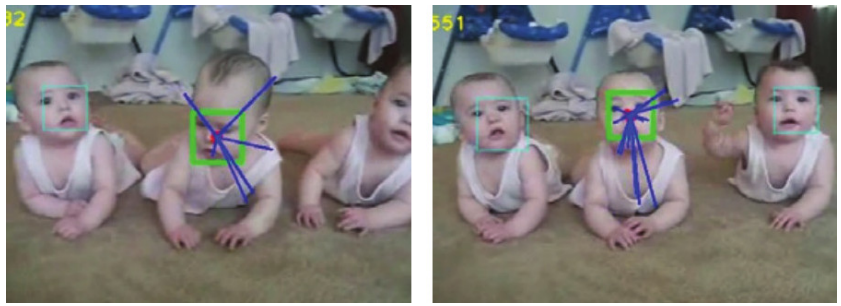


Related Generic Tracking Approaches

- Color features:
  - Expressive and efficient, popular in the early 2000's, *e.g.* [1,2].
  - For example: online feature selection [3], segmentation [4].
  - Tend to drift towards visually similar regions.
- Context-aware trackers:
  - Distinguish **supporters** and **distractors**, *e.g.* [5,6].
  - Often requires explicit tracking of such regions.



Grabner *et al.* [5].



Dinh *et al.* [6].

[1] Comaniciu *et al. Kernel-Based Object Tracking.* TPAMI 25(5), 2003.

[2] Pérez *et al. Color-Based Probabilistic Tracking.* ECCV'02.

[3] Collins *et al. Online Selection of Discriminative Tracking Features.* TPAMI 27(10), 2005.

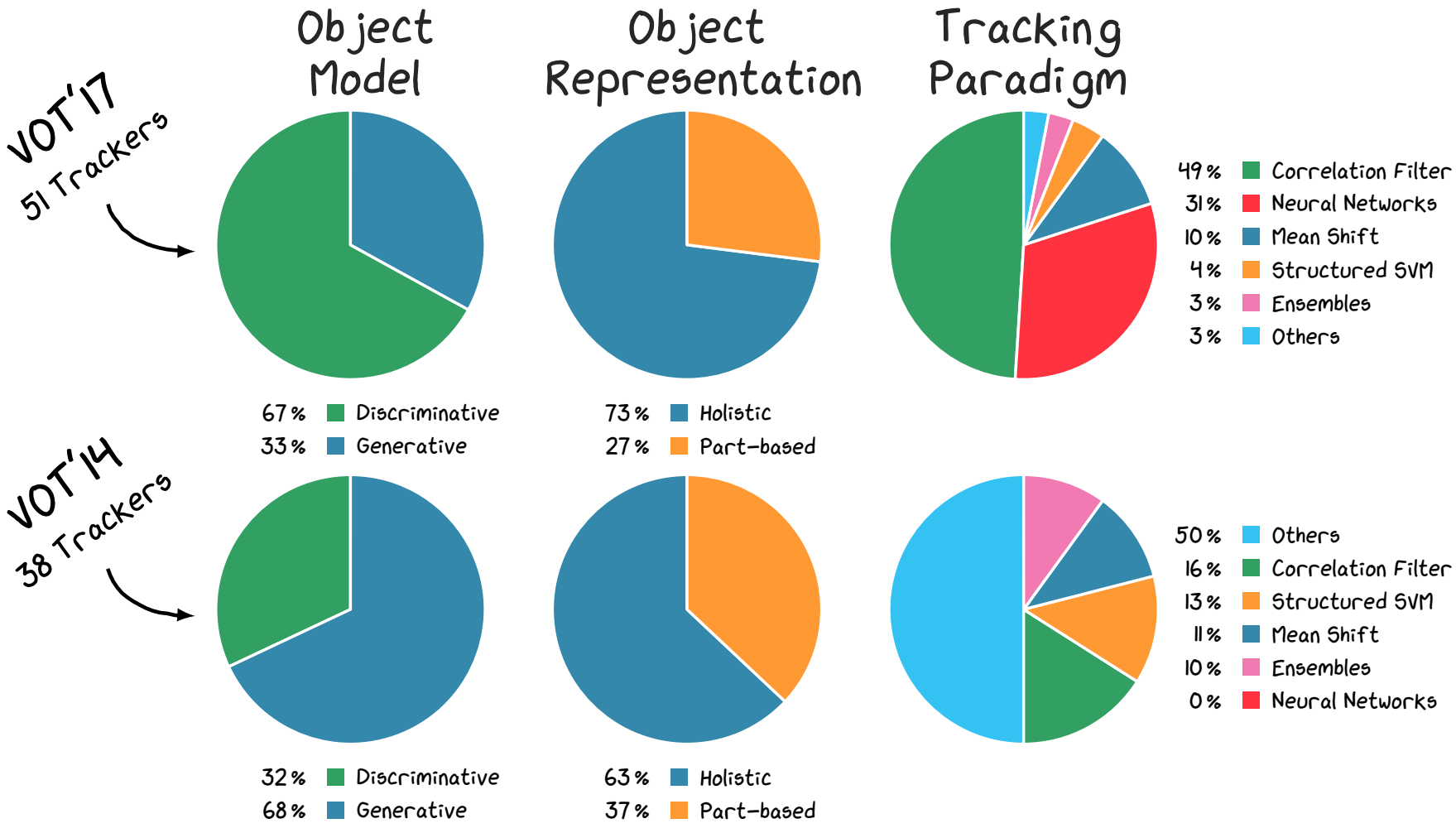
[4] Bibby and Reid. *Robust Real-Time Visual Tracking using Pixel-Wise Posteriors.* ECCV'08.

[5] Grabner *et al. Tracking the Invisible: Learning Where the Object Might be.* CVPR'10.

[6] Dinh *et al. Context Tracker: Exploring Supporters and Distracters in Unconstrained Environments.* CVPR'11.

Paradigm Shift

The Past 3 Years @ Visual Object Tracking (VOT) Challenges



Characteristics of SOT Benchmarks

Benchmark	Num. Videos	Number of Frames				Experiments		
		Total	Min	Mean	Max	Sup.	Unsup.	Pert.
VOT'13	16	5681	172	355 ± 158	770	✓		✓
VOT'14	25	10213	164	409 ± 248	1210	✓		✓
VOT'15	60	21455	41	358 ± 266	1500	✓		
VOT'16	60	21455	41	358 ± 266	1500	✓	✓	
OTB-50	49	26499	71	541 ± 433	1918		✓	✓
OTB-100	98	58260	71	595 ± 603	3872		✓	✓

- Publications:
  - VOTs: Kristan *et al.* VOT Workshops in conjunction with ECCV/ICCV.
  - OTBs: Wu *et al.* CVPR'13 and PAMI'15.
- Experiments:
  - Supervised - reset upon failure.
  - Unsupervised - init once without reset.
  - Perturbed - randomly perturbed initialization.

SOT Evaluation Metrics

- VOT Challenges [1] framework & protocol.
  - Accuracy:
    Average overlap (intersection over union).
  - Robustness:
    Average number of tracking failures.
  - Expected average overlap (EAO):
    Estimator of average region overlap on similar sequences.

- Online Tracking Benchmarks [2] framework & protocol.
  - Measure-threshold plots (similar to ROC curves).
    - Overlap success plot:
      Based on region overlap.
    - Distance precision plot:
      Based on center distance error.
  - Rank by AUC of overlap success plot.

---

[1] Kristan *et al.* The Visual Object Tracking Challenges. Workshops in conj. with ECCV/ICCV, 2013–2017.

[2] Wu *et al. Online Tracking Benchmark*, TPAMI 37(9), 2015.

Runtime Performance

Visual Object Tracking Challenges

	Tracker	Publication	Implementation	GPU	EFO
Ours	DAT		MATLAB		17.2
	DAT+s		MATLAB		17.0
	DAT+c		MATLAB		9.9
	noDAT		MATLAB		20.1
Major literature	ACT	CVPR'14	MATLAB/MEX		18.3
	C-COT	ECCV'16	MATLAB/MEX	✓	0.5
	DSST	BMVC'14	MATLAB/MEX		12.7
	FoT	CVWW'11	C/C++		114.6
	HoughTrack	CVIU'13	C/C++		0.9
	LGT	TPAMI'13	MATLAB/MEX		4.1
	MIL	TPAMI'11	C/C++		1.9
	PLT	VOT'13	C/C++		75.9
	SAMF	VOT'14	MATLAB/MEX		4.0
	Staple	CVPR'16	MATLAB/MEX		11.1
	Struck'16	TPAMI'16	C/C++		14.6
	TCNN	–	MATLAB/MEX	✓	1.0

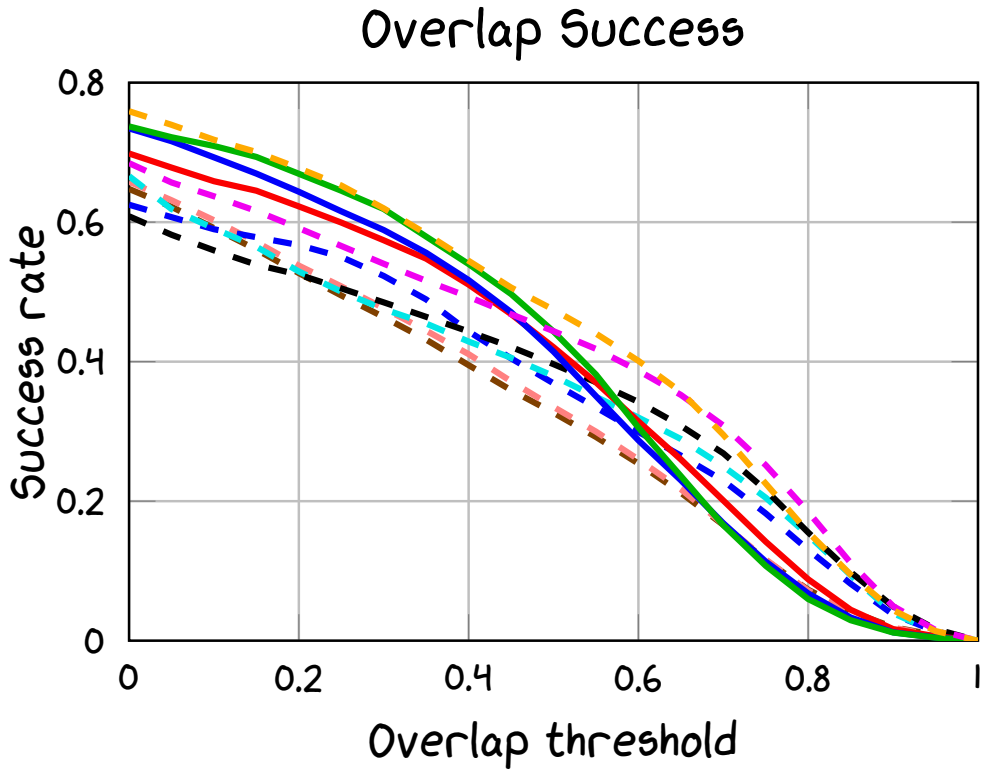
Runtime Performance

Online Tracking Benchmark

	Tracker	Publication	Implementation	FPS
Ours	DAT		MATLAB	143.1
	DAT+s		MATLAB	132.8
	DAT+c		MATLAB	70.2
	noDAT		MATLAB	180.2
Major literature	ASLA	CVPR'12	MATLAB/MEX	7.1
	CSK	ECCV'12	MATLAB/MEX	229.6
	CXT	CVPR'11	C/C++	14.3
	SCM	TIP'14	MATLAB/MEX	0.4
	Struck'11	ICCV'11	C/C++	10.0
	VTD	CVPR'10	MATLAB/MEX	3.3
	VTS	ICCV'11	MATLAB/MEX	3.1

Tracking Performance

OTB – Results over all 76 color sequences



Tracker	AUC	RDP
Struck [2]	0.43	0.57
DAT+s	0.39	0.54
SCM [7]	0.39	0.50
DAT+c	0.38	0.52
DAT	0.37	0.50
CXT [1]	0.35	0.47
CSK [3]	0.35	0.47
ASLA [4]	0.35	0.44
VTD [5]	0.32	0.44
VTS [6]	0.31	0.43

[1] Dinh *et al. Context Tracker: Exploring Supporters and Distracters in Unconstrained Environments*. CVPR'11.

[2] Hare *et al. Struck: Structured Output Tracking with Kernels*. ICCV'11.

[3] Henriques *et al. Exploiting the Circulant Structure of Tracking-by-detection with Kernels*. ECCV'12.

[4] Jia *et al. Visual Tracking via Adaptive Structural Local Sparse Appearance Model*. CVPR'12.

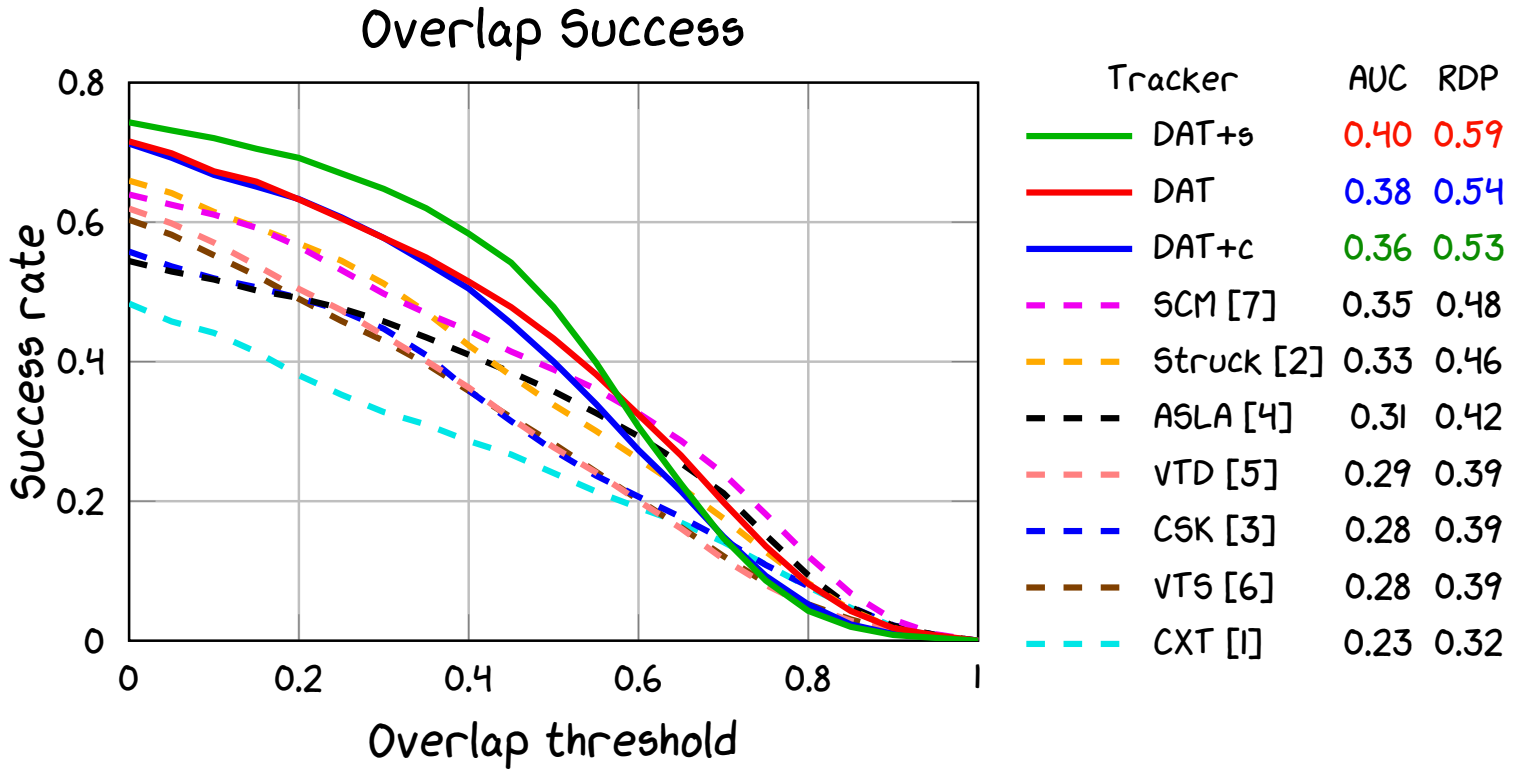
[5] Kwon and Lee. *Visual Tracking Decomposition*. CVPR'10.

[6] Kwon and Lee. *Tracking by Sampling Trackers*. ICCV'11.

[7] Zhong *et al. Robust Object Tracking via Sparse Collaborative Appearance Model*. TIP 23(5), 2014.

Tracking Performance

OTB – Non-rigid Deformations ✓



[1] Dinh *et al. Context Tracker: Exploring Supporters and Distracters in Unconstrained Environments*. CVPR'11.

[2] Hare *et al. Struck: Structured Output Tracking with Kernels*. ICCV'11.

[3] Henriques *et al. Exploiting the Circulant Structure of Tracking-by-detection with Kernels*. ECCV'12.

[4] Jia *et al. Visual Tracking via Adaptive Structural Local Sparse Appearance Model*. CVPR'12.

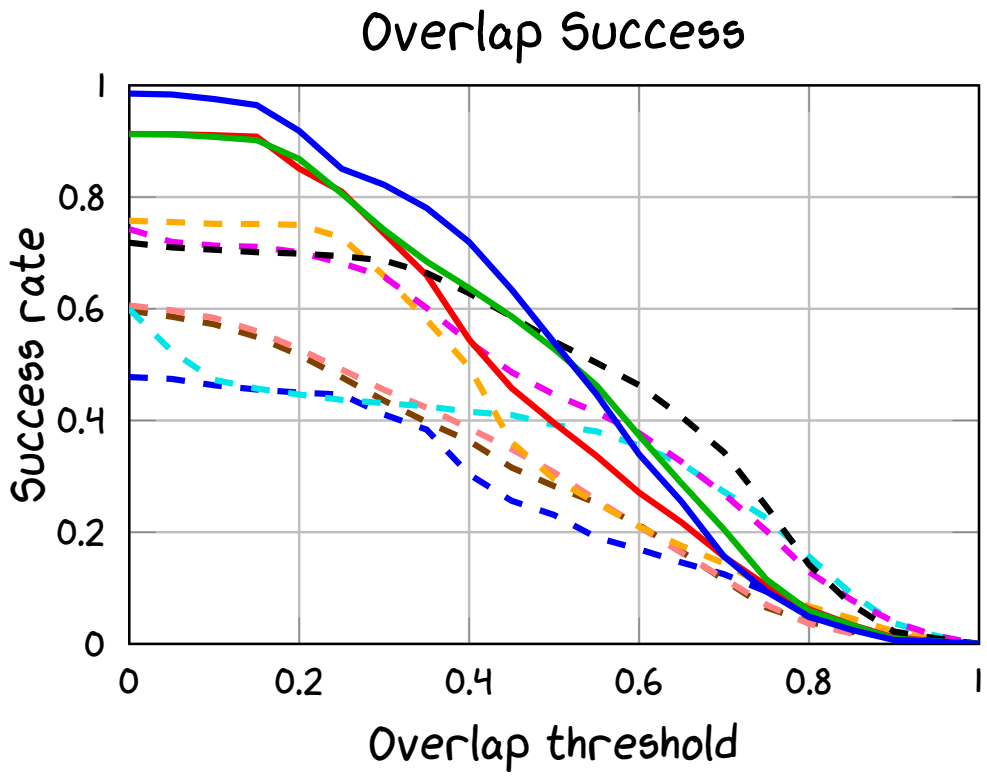
[5] Kwon and Lee. *Visual Tracking Decomposition*. CVPR'10.

[6] Kwon and Lee. *Tracking by Sampling Trackers*. ICCV'11.

[7] Zhong *et al. Robust Object Tracking via Sparse Collaborative Appearance Model*. TIP 23(5), 2014.

Tracking Performance

OTB – Low Resolution ✓



Tracker	AUC	RDP
DAT+c	0.50	0.97
DAT+s	0.48	0.91
ASLA [4]	0.45	0.71
DAT	0.44	0.91
SCM [7]	0.42	0.72
Struck [2]	0.38	0.76
CXT [1]	0.33	0.48
VTD [5]	0.29	0.59
VTS [6]	0.29	0.58
CSK [3]	0.25	0.46

[1] Dinh *et al. Context Tracker: Exploring Supporters and Distracters in Unconstrained Environments*. CVPR'11.

[2] Hare *et al. Struck: Structured Output Tracking with Kernels*. ICCV'11.

[3] Henriques *et al. Exploiting the Circulant Structure of Tracking-by-detection with Kernels*. ECCV'12.

[4] Jia *et al. Visual Tracking via Adaptive Structural Local Sparse Appearance Model*. CVPR'12.

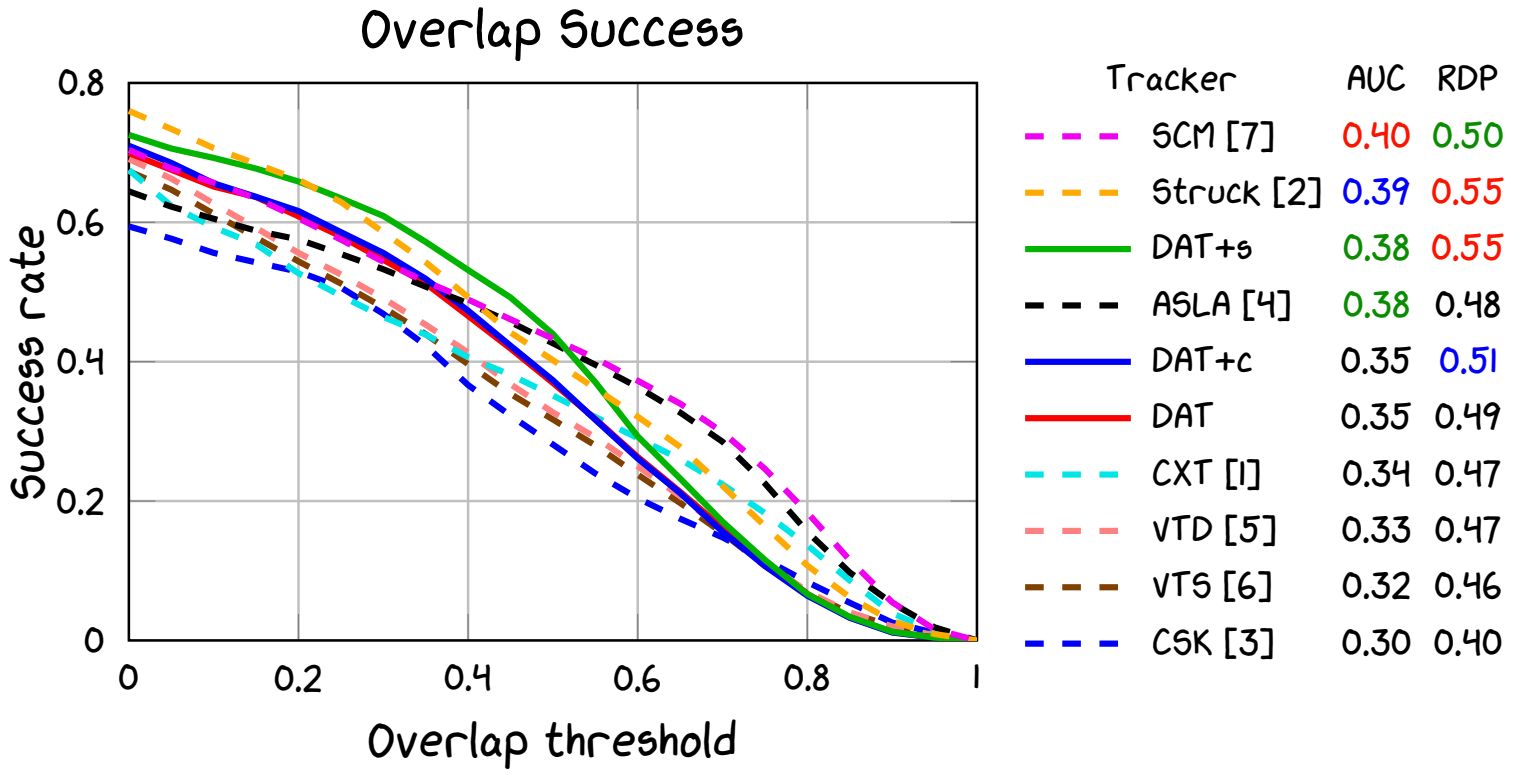
[5] Kwon and Lee. *Visual Tracking Decomposition*. CVPR'10.

[6] Kwon and Lee. *Tracking by Sampling Trackers*. ICCV'11.

[7] Zhong *et al. Robust Object Tracking via Sparse Collaborative Appearance Model*. TIP 23(5), 2014.

Tracking Performance

OTB – Scale Variation ✗



[1] Dinh *et al. Context Tracker: Exploring Supporters and Distracters in Unconstrained Environments*. CVPR'11.

[2] Hare *et al. Struck: Structured Output Tracking with Kernels*. ICCV'11.

[3] Henriques *et al. Exploiting the Circulant Structure of Tracking-by-detection with Kernels*. ECCV'12.

[4] Jia *et al. Visual Tracking via Adaptive Structural Local Sparse Appearance Model*. CVPR'12.

[5] Kwon and Lee. *Visual Tracking Decomposition*. CVPR'10.

[6] Kwon and Lee. *Tracking by Sampling Trackers*. ICCV'11.

[7] Zhong *et al. Robust Object Tracking via Sparse Collaborative Appearance Model*. TIP 23(5), 2014.

Default Parameter Settings

Distractor-aware Tracking

Parameter		Value
Color space		RGB
Histogram bins		$16 \times 16 \times 16$
Learning rate for $p^t_{O,S}(\mathbf{x} \in \mathcal{O} \mid b_{\mathbf{x}})$	$\eta_S \in [0, 1]$	0.05
Learning rate for $p^t_{O,D}(\mathbf{x} \in \mathcal{O} \mid b_{\mathbf{x}})$	$\eta_D \in [0, 1]$	0.20
Scaling factor for surrounding region S	$\lambda_S \in (1, \lambda_W)$	2.00
Scaling factor for search window W	$\lambda_W \in (\lambda_S, \infty)$	4.00
NMS patch overlap	$o_\nu \in [0, 1)$	0.90
NMS reporting threshold	$\tau_\nu \in (0, 1)$	0.50

Ablation Study - Example

Color Spaces

- Standard (s)RGB.
- rg chromaticity.

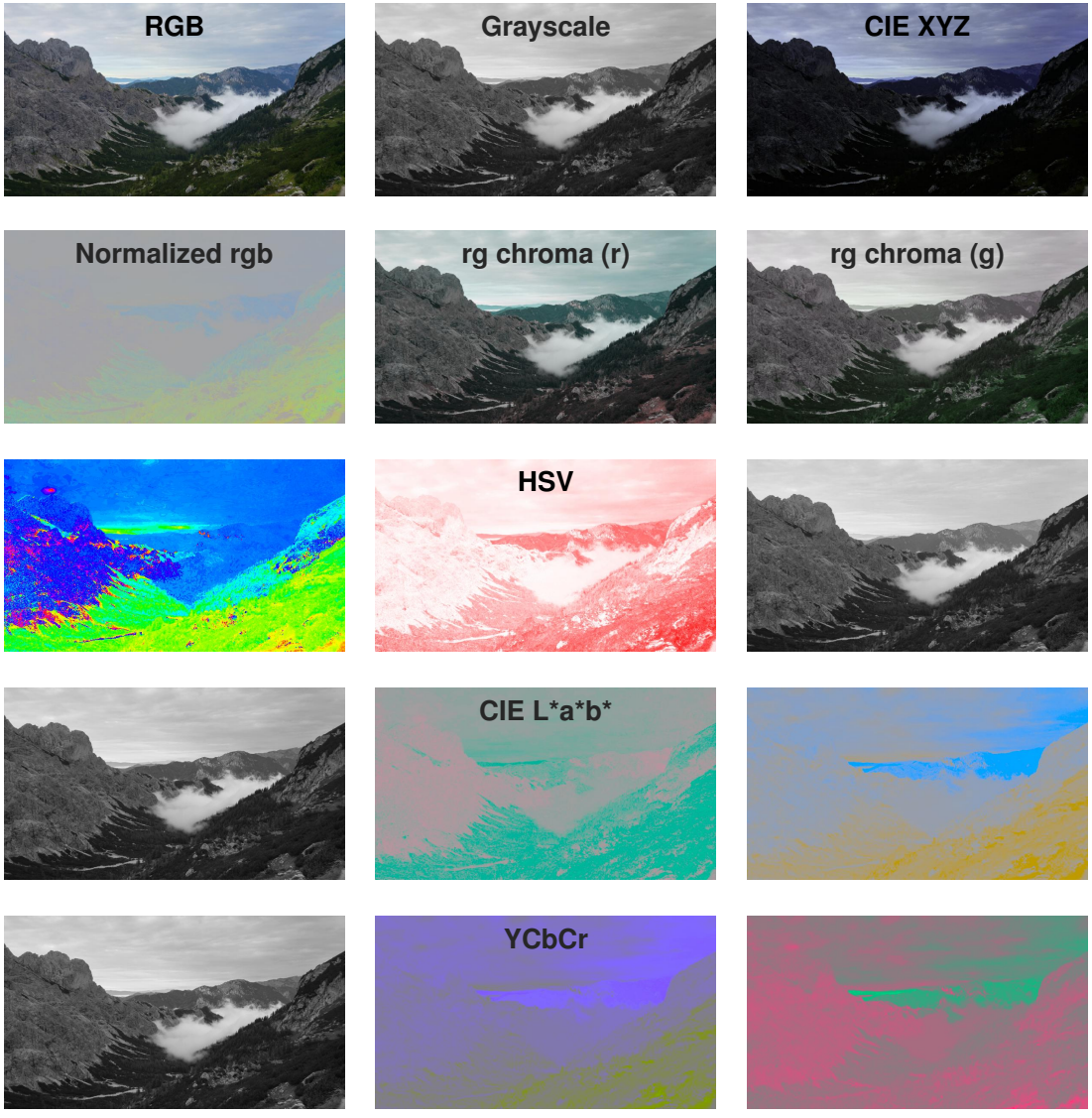
  Red/green proportions.
- Grayscale.
- CIE XYZ.

  Device independent.
- HSV.

  “The best for tracking.”
- CIE L*a*b*.

  “The best for tracking”, again.
- YCbCr.

  Video transmission.



Ablation Study

Color Spaces

Highlights: Best, second best, third best.

Tracker	Color Space	Experiment *baseline*	Experiment *region noise*	Overall	
DAT	RGB				
noDAT	RGB				
DAT	HSV				
noDAT	HSV				
DAT	L*a*b*				
noDAT	L*a*b*				
DAT	YCbCr				
noDAT	YCbCr				
DAT	XYZ				
noDAT	XYZ				
DAT	HS				
noDAT	HS				
DAT	rg chroma				
noDAT	rg chroma				
DAT	Gray				
noDAT	Gray				

Ablation Study

Color Spaces

Highlights: Best, second best, third best.

Tracker	Color Space	Experiment *baseline* Acc.↑	Experiment *region noise*	Overall
DAT	RGB	0.60		
noDAT	RGB	0.60		
DAT	HSV	0.61		
noDAT	HSV	0.61		
DAT	L*a*b*	0.59		
noDAT	L*a*b*	0.59		
DAT	YCbCr	0.58		
noDAT	YCbCr	0.58		
DAT	XYZ	0.53		
noDAT	XYZ	0.54		
DAT	HS	0.59		
noDAT	HS	0.58		
DAT	rg chroma	0.57		
noDAT	rg chroma	0.57		
DAT	Gray	0.53		
noDAT	Gray	0.52		

Ablation Study

Color Spaces

Highlights: Best, second best, third best.

Tracker	Color Space	Experiment *baseline* Acc.↑	Experiment *baseline* Rob.↓	Experiment *region noise*	Overall
DAT	RGB	0.60	0.08		
noDAT	RGB	0.60	0.19		
DAT	HSV	0.61	0.34		
noDAT	HSV	0.61	0.42		
DAT	L*a*b*	0.59	0.19		
noDAT	L*a*b*	0.59	0.32		
DAT	YCbCr	0.58	0.23		
noDAT	YCbCr	0.58	0.15		
DAT	XYZ	0.53	1.38		
noDAT	XYZ	0.54	2.76		
DAT	HS	0.59	0.48		
noDAT	HS	0.58	0.63		
DAT	rg chroma	0.57	1.39		
noDAT	rg chroma	0.57	1.83		
DAT	Gray	0.53	3.70		
noDAT	Gray	0.52	4.51		

Ablation Study

Color Spaces

Highlights: Best, second best, third best.

Tracker	Color Space	Experiment *baseline* Acc.↑	Experiment *baseline* Rob.↓	Experiment *region noise* Acc.↑	Experiment *region noise* Rob.↓	Overall
DAT	RGB	0.60	0.08	0.59	0.12	
noDAT	RGB	0.60	0.19	0.59	0.21	
DAT	HSV	0.61	0.34	0.60	0.28	
noDAT	HSV	0.61	0.42	0.60	0.35	
DAT	L*a*b*	0.59	0.19	0.58	0.22	
noDAT	L*a*b*	0.59	0.32	0.58	0.30	
DAT	YCbCr	0.58	0.23	0.57	0.18	
noDAT	YCbCr	0.58	0.15	0.57	0.22	
DAT	XYZ	0.53	1.38	0.53	1.26	
noDAT	XYZ	0.54	2.76	0.54	2.30	
DAT	HS	0.59	0.48	0.57	0.43	
noDAT	HS	0.58	0.63	0.57	0.61	
DAT	rg chroma	0.57	1.39	0.56	1.29	
noDAT	rg chroma	0.57	1.83	0.56	1.75	
DAT	Gray	0.53	3.70	0.52	3.39	
noDAT	Gray	0.52	4.51	0.53	4.66	

Ablation Study

Color Spaces

Highlights: Best, second best, third best.

Tracker	Color Space	Experiment *baseline*		Experiment *region noise*		Overall	
		Acc.↑	Rob.↓	Acc.↑	Rob.↓	EAO↑	FPS↑
DAT	RGB	0.60	0.08	0.59	0.12	0.55	113.0
noDAT	RGB	0.60	0.19	0.59	0.21	0.51	160.5
DAT	HSV	0.61	0.34	0.60	0.28	0.46	71.9
noDAT	HSV	0.61	0.42	0.60	0.35	0.43	89.1
DAT	L*a*b*	0.59	0.19	0.58	0.22	0.46	36.5
noDAT	L*a*b*	0.59	0.32	0.58	0.30	0.42	39.3
DAT	YCbCr	0.58	0.23	0.57	0.18	0.45	83.3
noDAT	YCbCr	0.58	0.15	0.57	0.22	0.46	106.2
DAT	XYZ	0.53	1.38	0.53	1.26	0.25	31.9
noDAT	XYZ	0.54	2.76	0.54	2.30	0.18	34.5
DAT	HS	0.59	0.48	0.57	0.43	0.39	79.0
noDAT	HS	0.58	0.63	0.57	0.61	0.37	95.9
DAT	rg chroma	0.57	1.39	0.56	1.29	0.20	115.0
noDAT	rg chroma	0.57	1.83	0.56	1.75	0.16	143.3
DAT	Gray	0.53	3.70	0.52	3.39	0.14	169.1
noDAT	Gray	0.52	4.51	0.53	4.66	0.11	217.4

Ablation Study

Model Size (Number of Bins) – Scale Adaptations

Model Size	Experiment *baseline*		Experiment *region noise*		Overall	
	Acc.↑	Rob.↓	Acc.↑	Rob.↓	EAO↑	FPS↑
$8 \times 8 \times 8$	0.58	0.19	0.57	0.21	0.47	117.9
$10 \times 10 \times 10$	0.59	0.23	0.58	0.24	0.44	124.1
$16 \times 16 \times 16$	0.60	0.08	0.59	0.16	0.53	112.1
$32 \times 32 \times 32$	0.60	0.38	0.59	0.38	0.45	105.9
$64 \times 64 \times 64$	0.59	1.17	0.57	1.05	0.29	91.2

Tracker	Experiment *baseline*		Experiment *region noise*		Overall	
	Acc.↑	Rob.↓	Acc.↑	Rob.↓	EAO↑	FPS↑
DAT	0.60	0.08	0.59	0.12	0.55	113.0
DAT+s	0.57	0.00	0.56	0.07	0.56	108.8
DAT+c	0.58	0.09	0.57	0.12	0.51	56.7
DAT+r	0.51	0.45	0.49	0.56	0.44	90.7

Ablation Study

Learning Rates

Learning Rate		Experiment *baseline*		Experiment *region noise*		Overall	
η_S	η_D	Acc.↑	Rob.↓	Acc.↑	Rob.↓	EAO↑	FPS↑
0.01	0.20	0.60	0.26	0.59	0.29	0.49	111.5
0.05	0.20	0.60	0.08	0.59	0.15	0.54	113.8
0.10	0.20	0.59	0.14	0.59	0.11	0.53	113.6
0.15	0.20	0.60	0.49	0.58	0.30	0.43	113.0
0.20	0.20	0.57	0.69	0.58	0.42	0.42	113.2
0.25	0.20	0.58	0.73	0.57	0.58	0.40	115.4
0.05	0.01	0.60	0.15	0.59	0.16	0.52	111.2
0.05	0.05	0.60	0.08	0.59	0.16	0.53	112.9
0.05	0.10	0.60	0.08	0.59	0.14	0.54	110.8
0.05	0.15	0.60	0.08	0.59	0.12	0.54	108.7
0.05	0.20	0.60	0.08	0.59	0.15	0.54	113.8
0.05	0.25	0.60	0.12	0.59	0.15	0.51	114.1

Ablation Study

Window Sizes

Window Scale Parameter		Experiment *baseline*		Experiment *region noise*		Overall	
λ_W	λ_S	Acc.↑	Rob.↓	Acc.↑	Rob.↓	EAO↑	FPS↑
2.0	1.5	0.56	0.94	0.56	0.87	0.36	212.9
4.0	2.0	0.60	0.15	0.58	0.22	0.51	114.1
4.0	3.0	0.58	0.43	0.57	0.37	0.44	99.5
8.0	2.0	0.59	0.08	0.58	0.24	0.51	46.6
8.0	3.0	0.57	0.53	0.57	0.37	0.45	44.7
8.0	4.0	0.57	0.39	0.56	0.52	0.44	39.8
8.0	5.0	0.57	0.39	0.56	0.50	0.42	37.0
8.0	6.0	0.57	0.43	0.56	0.66	0.39	33.8
8.0	7.0	0.57	0.67	0.55	0.70	0.33	32.3

Ablation Study

Non-maximum Suppression

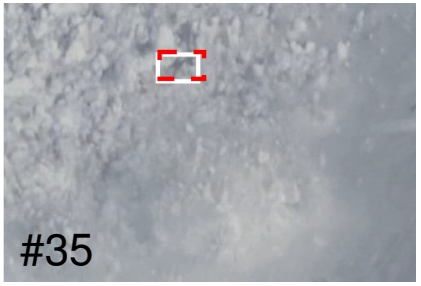
NMS Parameter		Experiment *baseline*		Experiment *region noise*		Overall	
o_ν	τ_ν	Acc.$^\uparrow$	Rob.$^\downarrow$	Acc.$^\uparrow$	Rob.$^\downarrow$	EAO$^\uparrow$	FPS$^\uparrow$
0.95	0.50	0.60	0.19	0.59	0.16	0.51	104.8
0.90	0.50	0.60	0.08	0.59	0.16	0.53	113.2
0.85	0.50	0.59	0.08	0.59	0.16	0.52	109.7
0.75	0.50	0.58	0.29	0.57	0.19	0.47	147.1
0.50	0.50	0.50	0.10	0.50	0.13	0.45	115.8
0.90	0.75	0.60	0.15	0.60	0.15	0.53	118.8
0.90	0.50	0.60	0.08	0.59	0.16	0.53	113.2
0.90	0.25	0.60	0.08	0.59	0.14	0.54	105.5

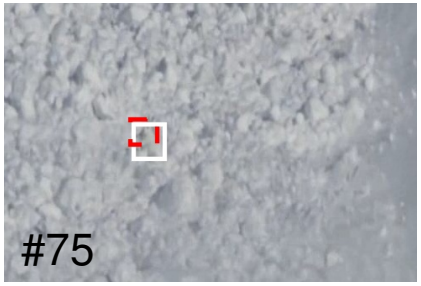
Limitations

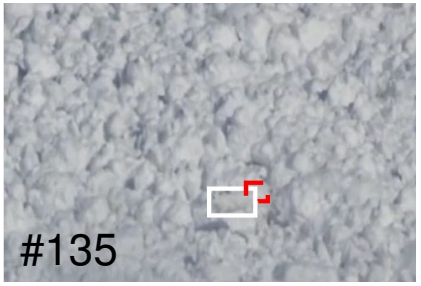
Distractor-aware Tracking

- Indistinguishable color distributions for object and surroundings.

rabbit





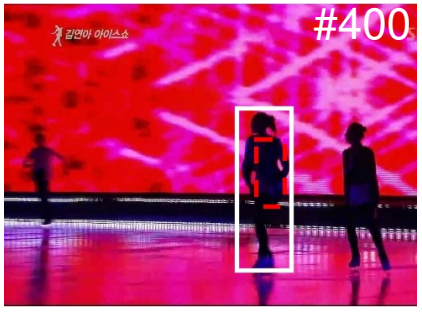


- Scale changes in combination with partial occlusions.

graduate

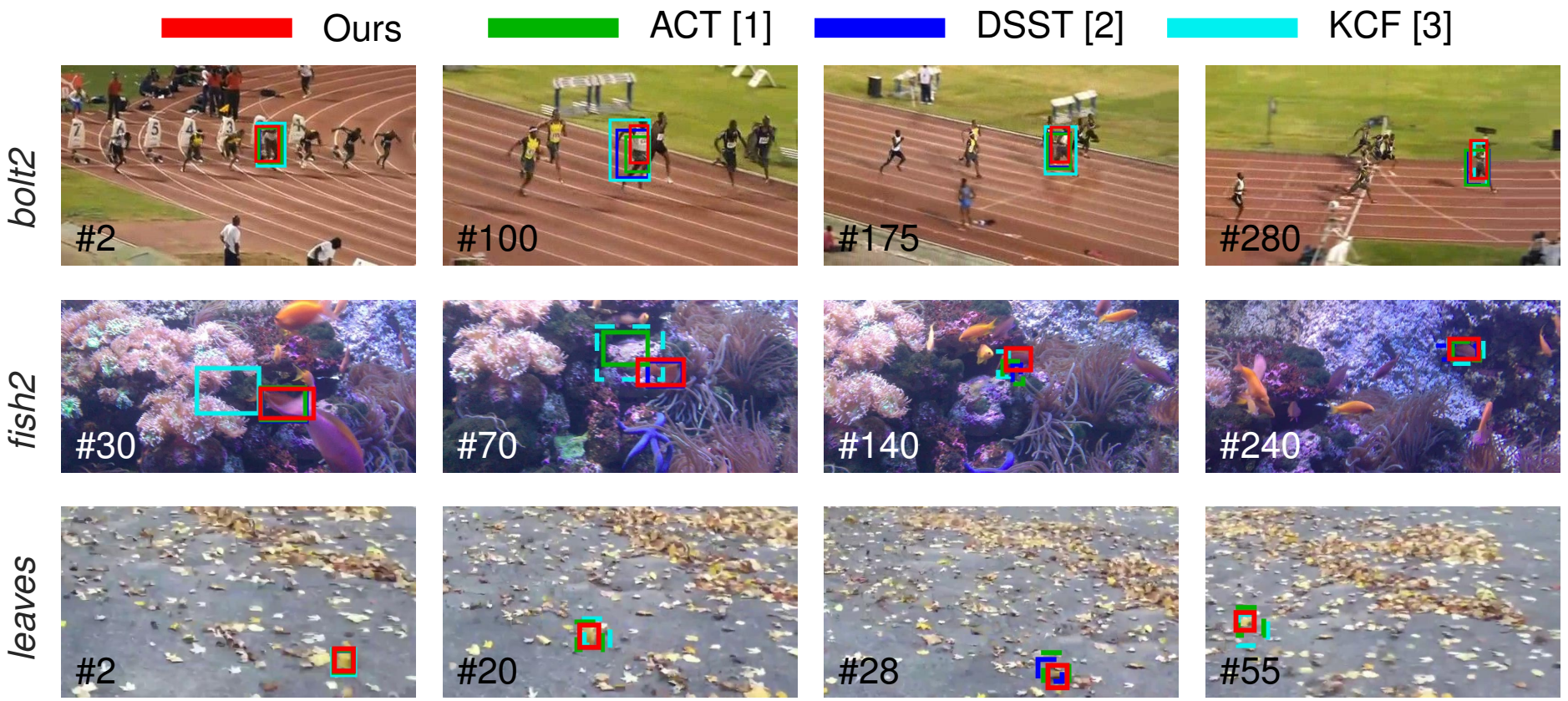
- Immediate illumination changes, occlusions and low contrast.

skating



Exemplary Results – 1/3

Distractor-aware Tracking



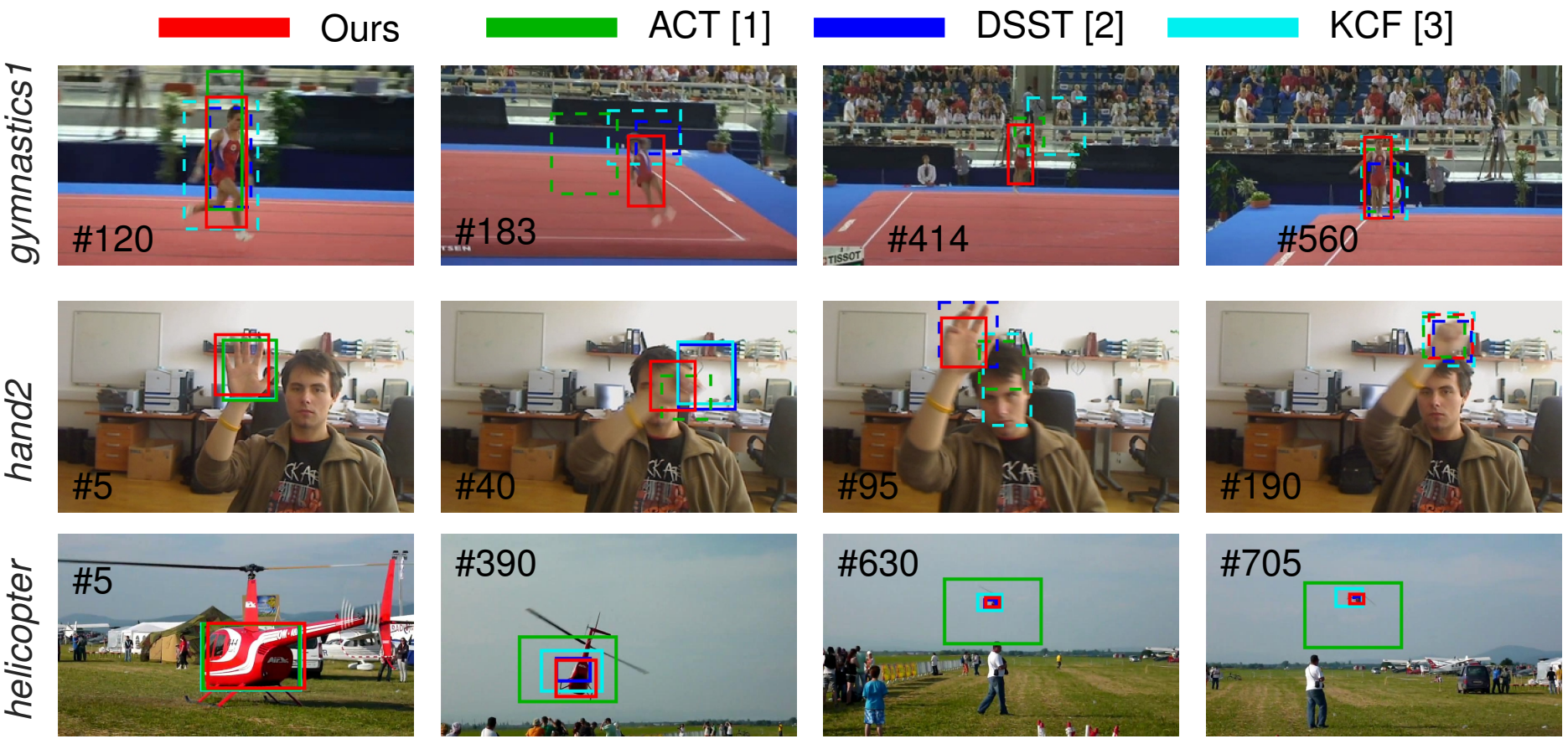
[1] Danelljan *et al. Adaptive Color Attributes for Real-Time Visual Tracking*. CVPR'14.

[2] Danelljan *et al. Accurate Scale Estimation for Robust Visual Tracking*. BMVC'14.

[3] Henriques *et al. High-Speed Tracking with Kernelized Correlation Filters*. TODO PAMI'15.

Exemplary Results – 2/3

Distractor-aware Tracking



[1] Danelljan *et al. Adaptive Color Attributes for Real-Time Visual Tracking*. CVPR'14.

[2] Danelljan *et al. Accurate Scale Estimation for Robust Visual Tracking*. BMVC'14.

[3] Henriques *et al. High-Speed Tracking with Kernelized Correlation Filters*.TODO PAMI'15.

Exemplary Results - 3/3

Distractor-aware Tracking

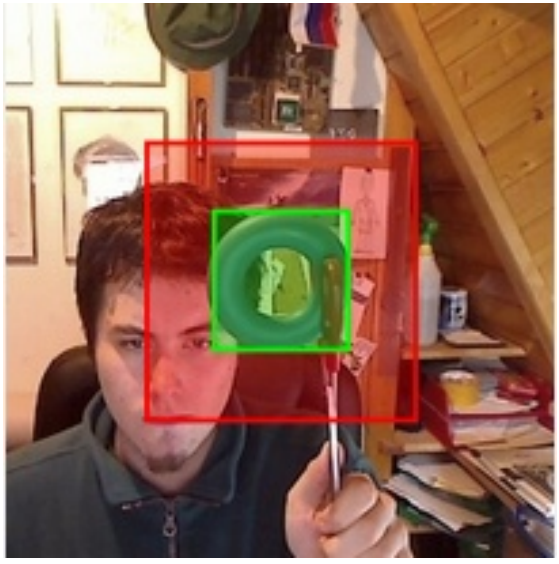
[1] Danelljan *et al. Adaptive Color Attributes for Real-Time Visual Tracking*. CVPR'14.

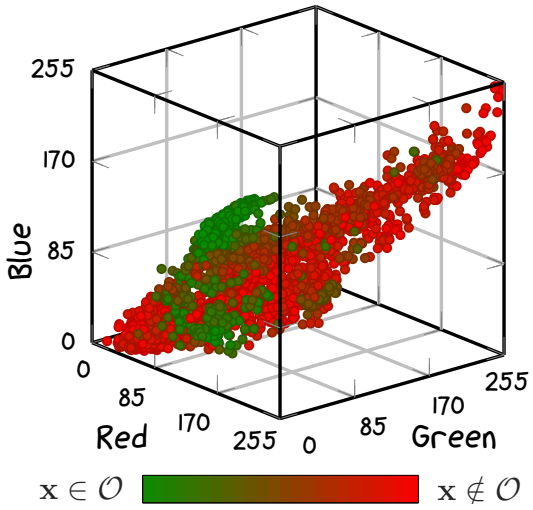
[2] Danelljan *et al. Accurate Scale Estimation for Robust Visual Tracking*. BMVC'14.

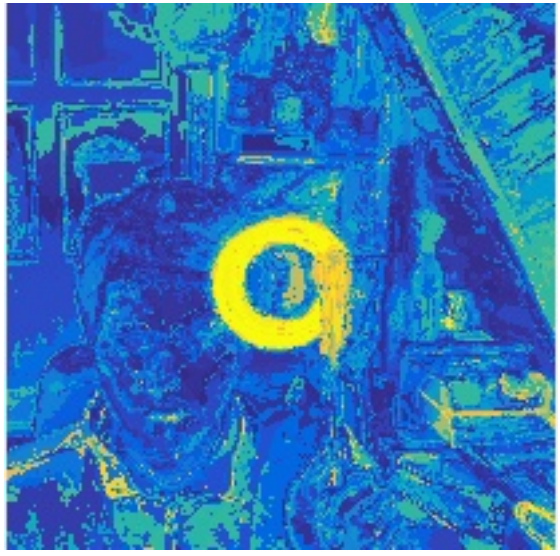
[3] Henriques *et al. High-Speed Tracking with Kernelized Correlation Filters*.TODO PAMI'15.

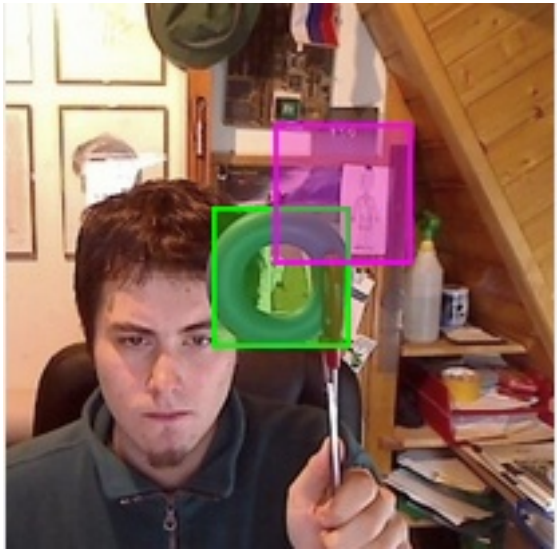
Color Distributions

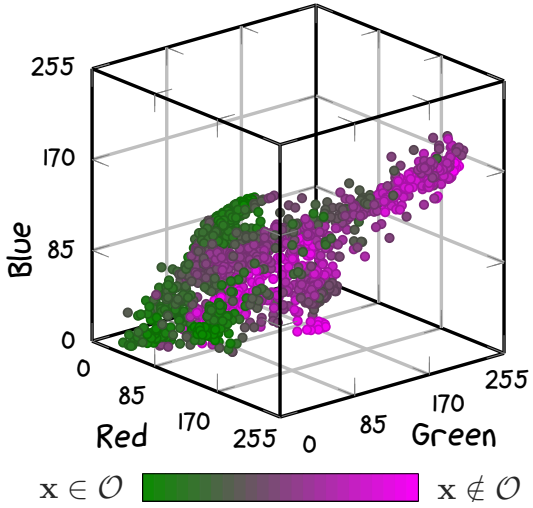
Sequence `torus`

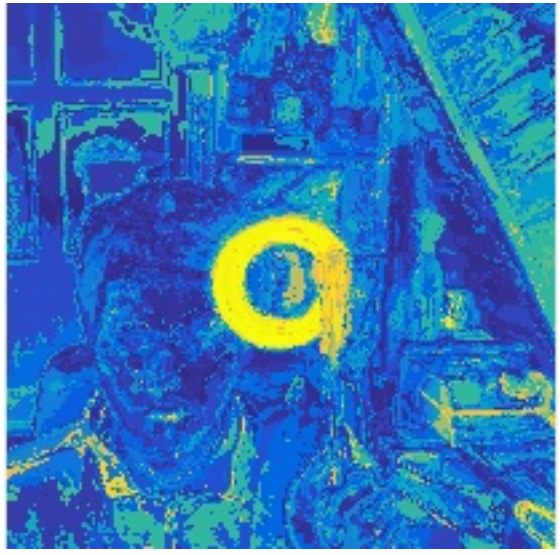






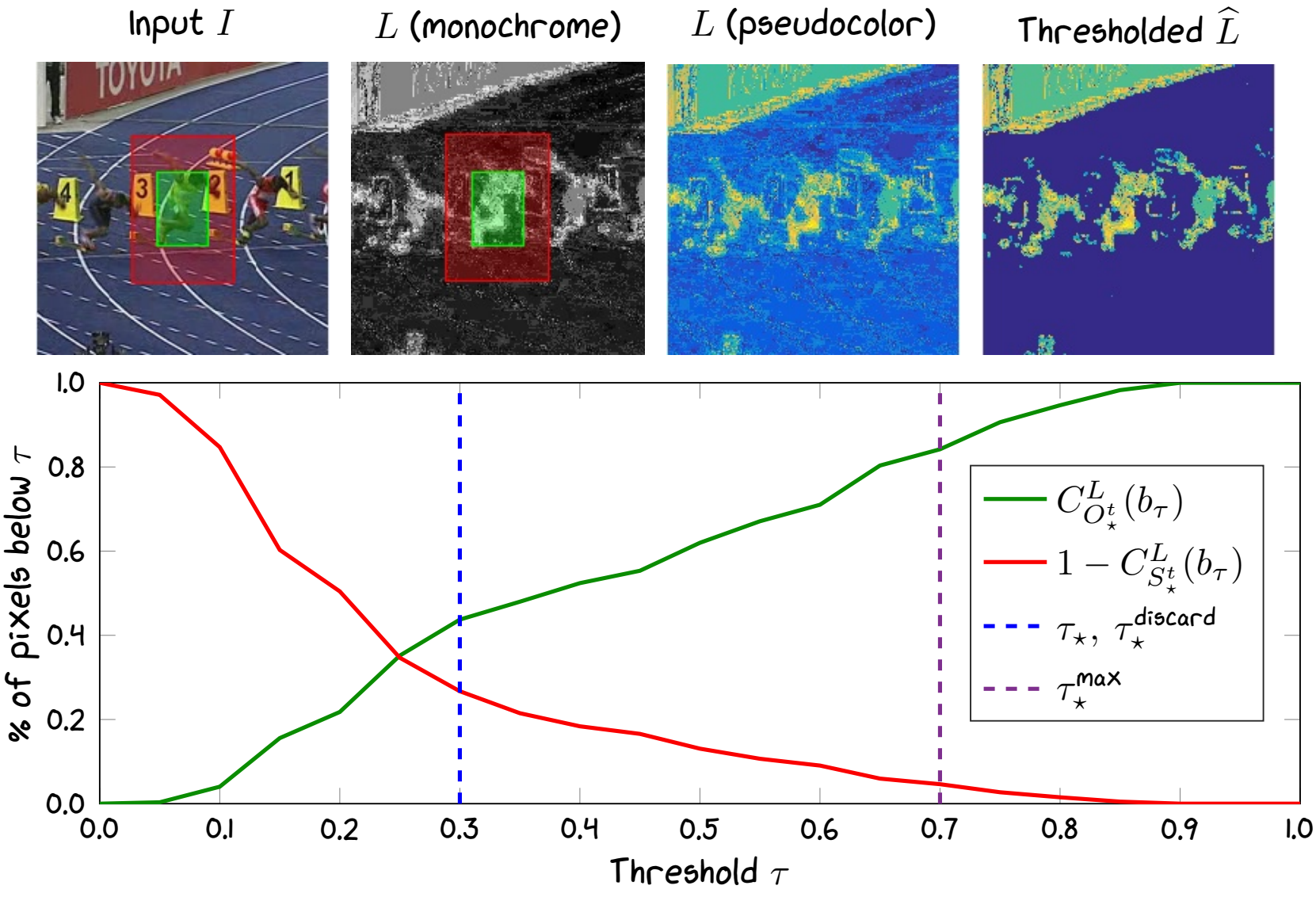






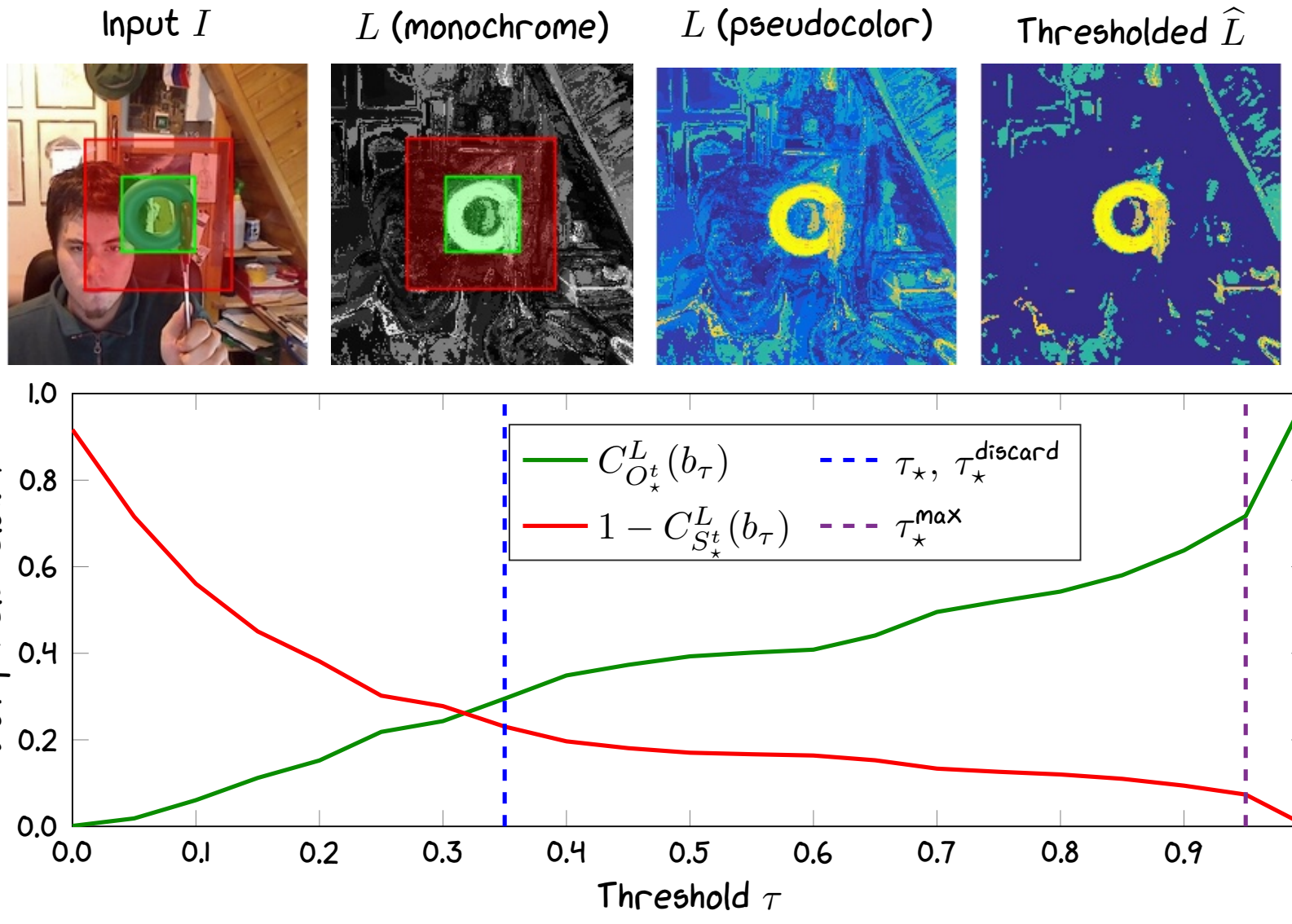
Scale Adaptation

Adaptive Pre-Segmentation (Sequence `bolt`)



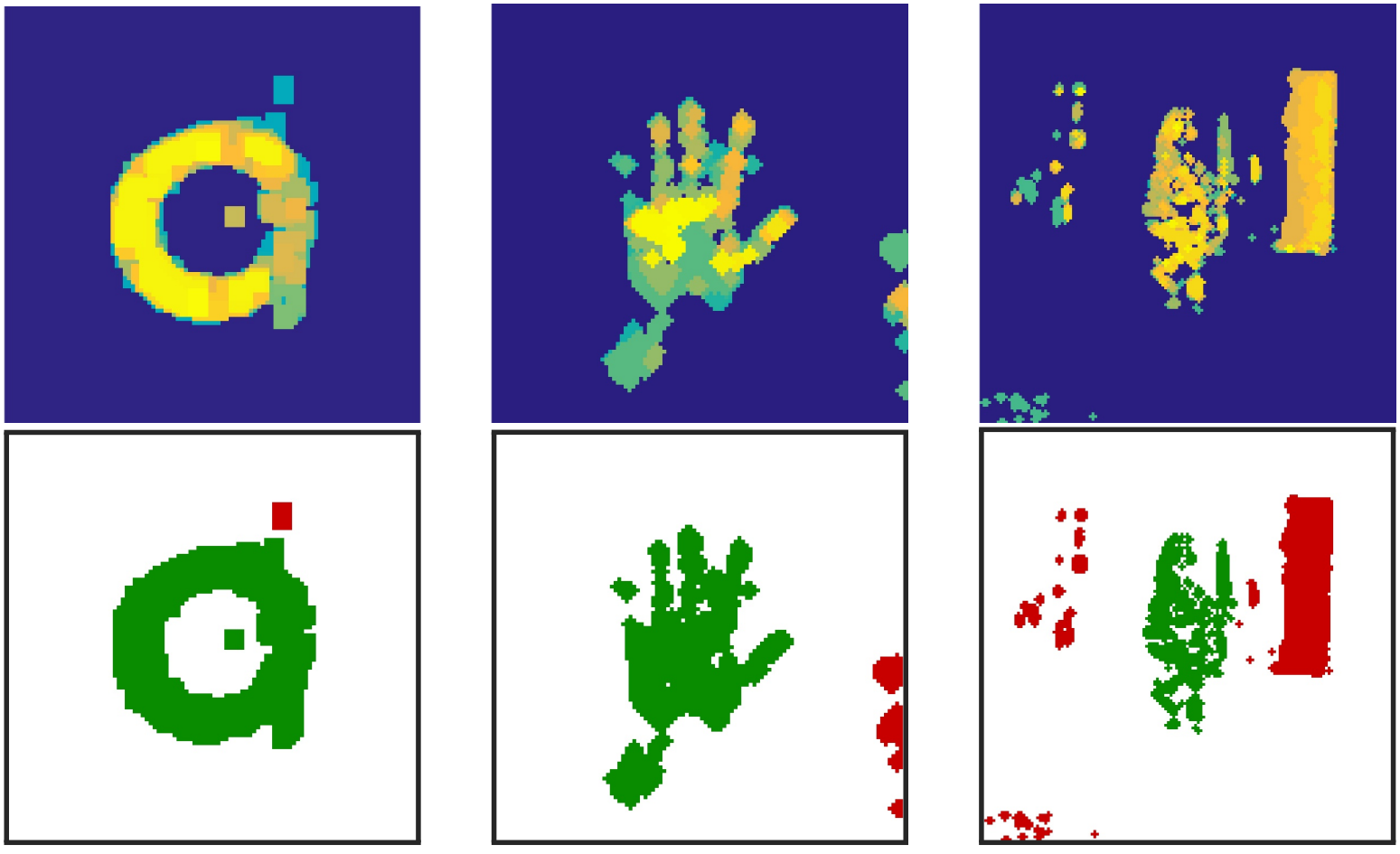
Adaptive Pre-Segmentation

Sequence `torus`



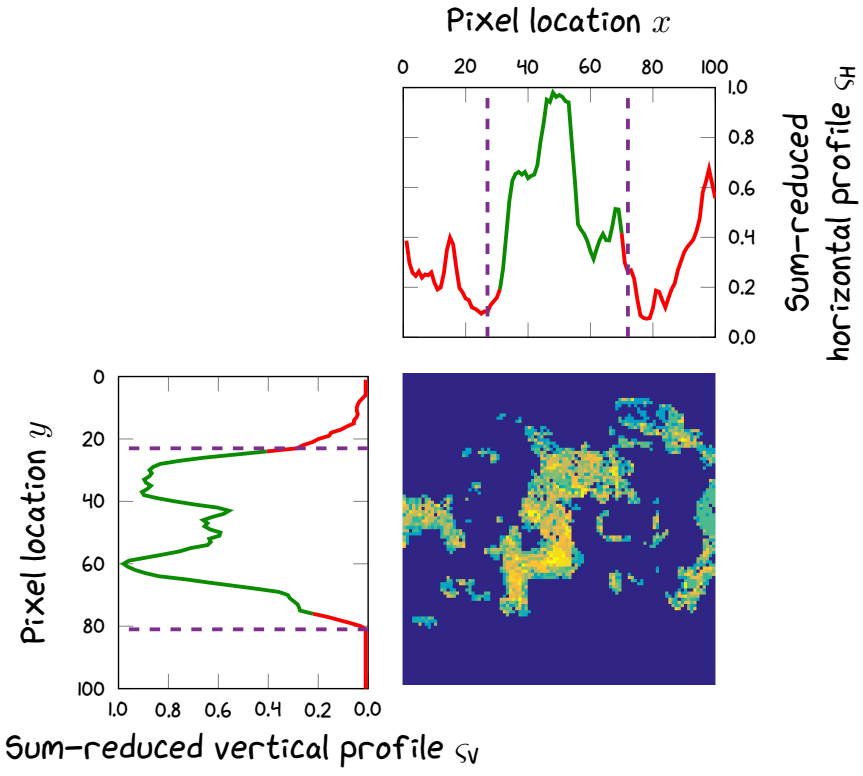
Connected Components

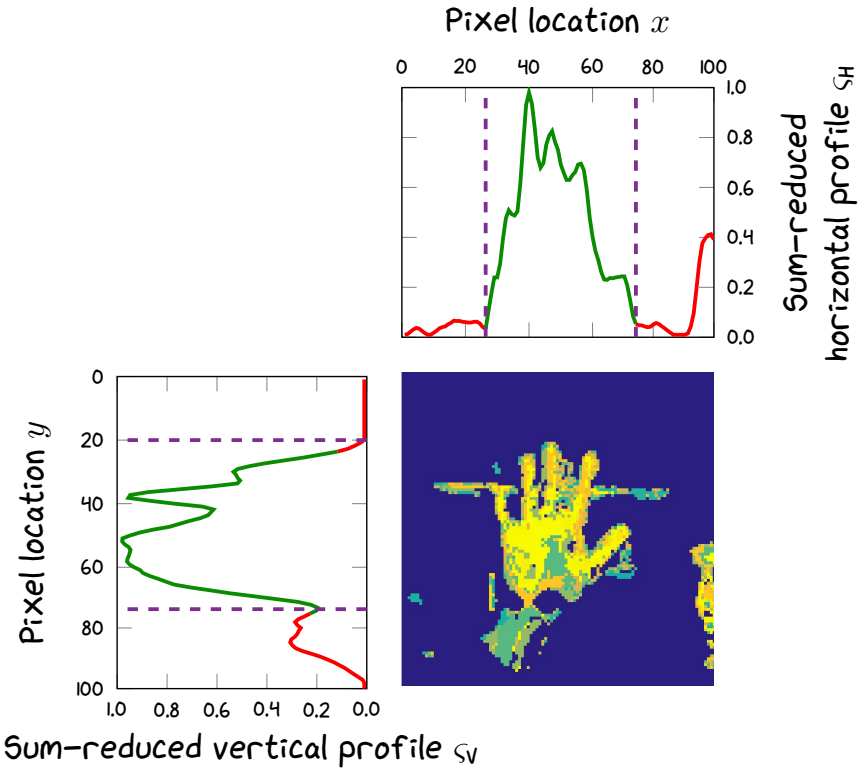
Sequences `torus`, `hand2`, `bicycle`



Scaling via Sum Reduction Profiles

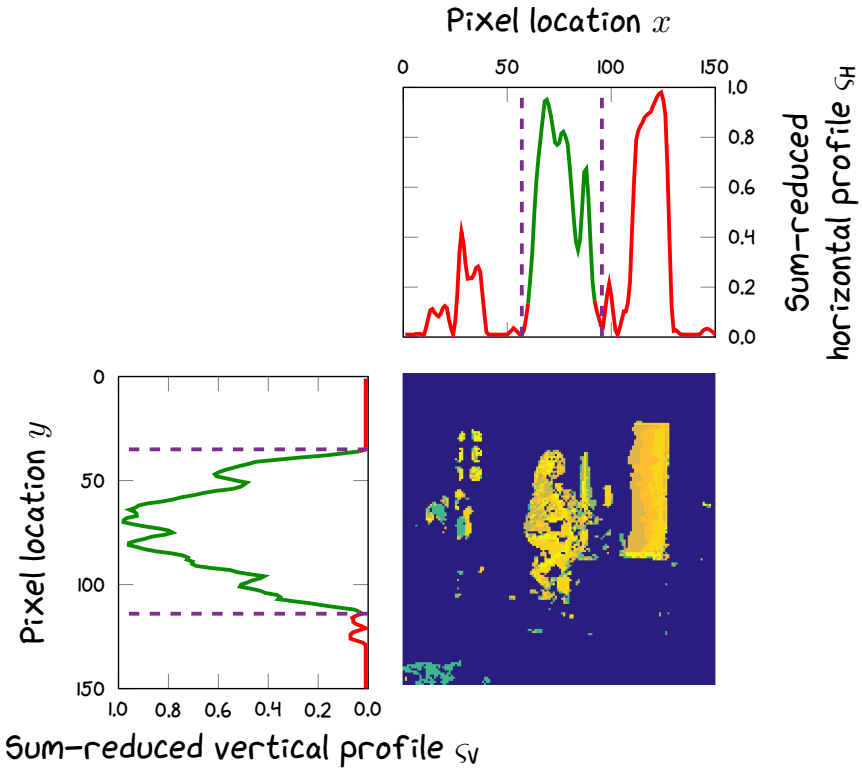
Sequences `bolt`, `hand2`

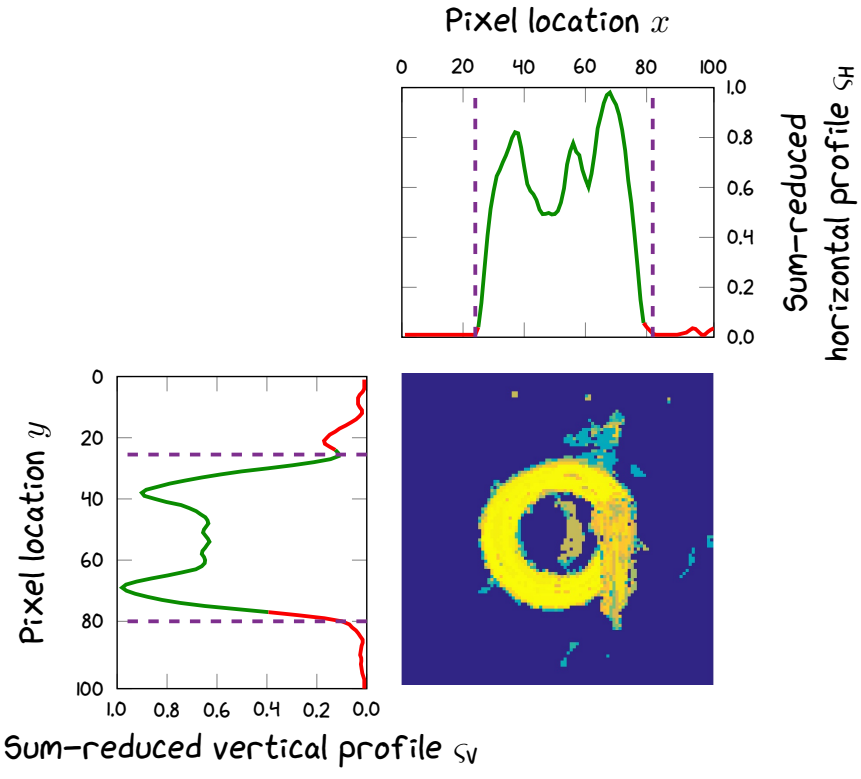




Scaling via Sum Reduction Profiles

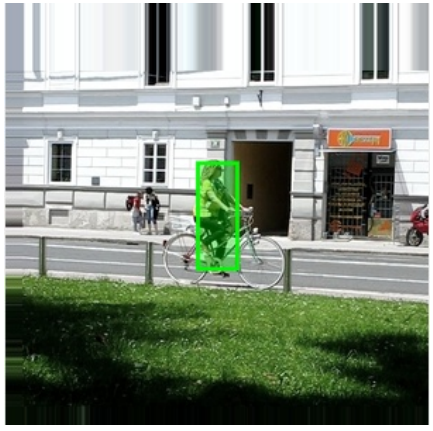
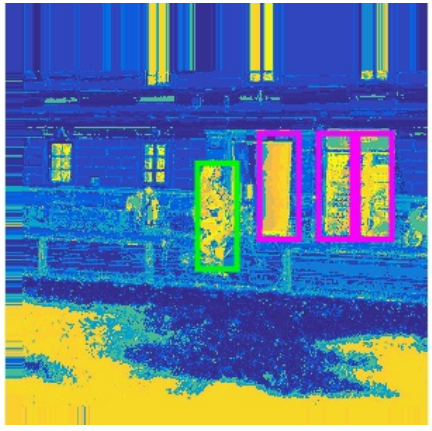
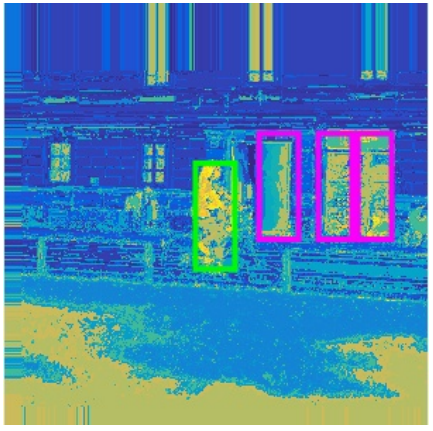
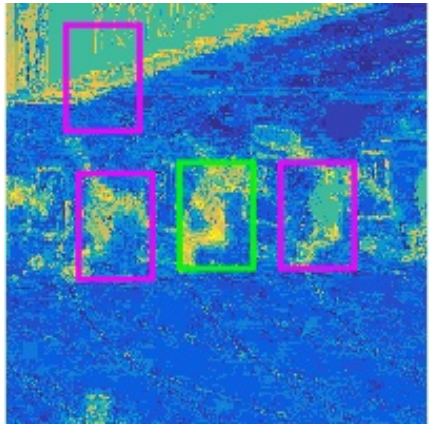
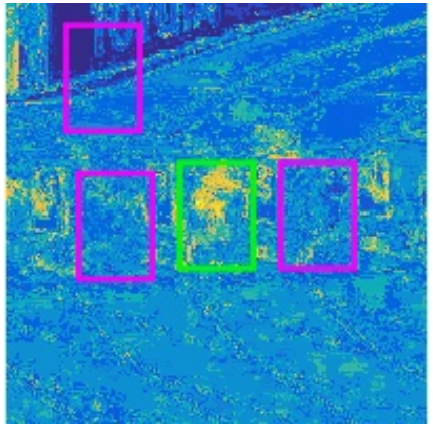
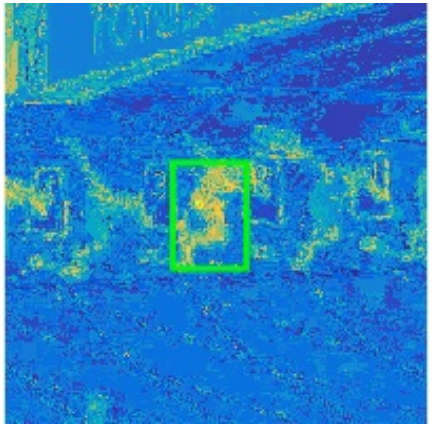
Sequences `bicycle`, `torus`





Localization

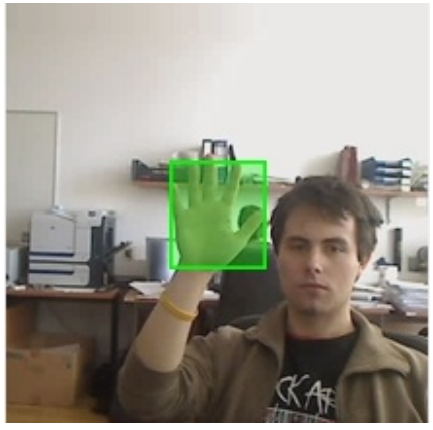
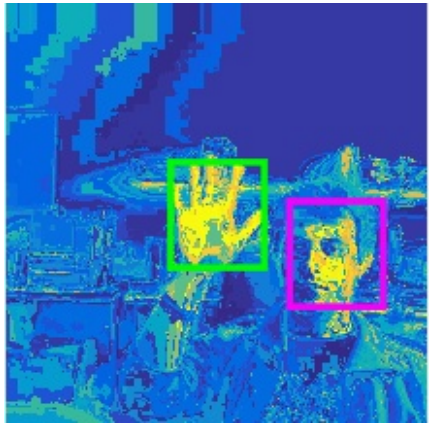
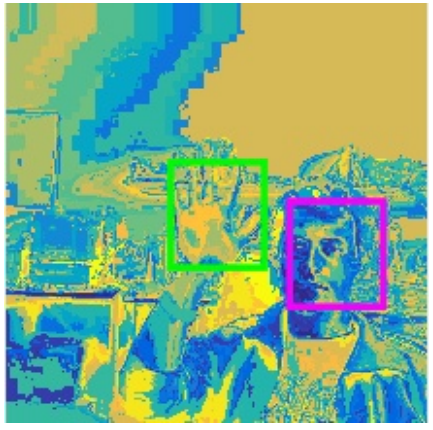
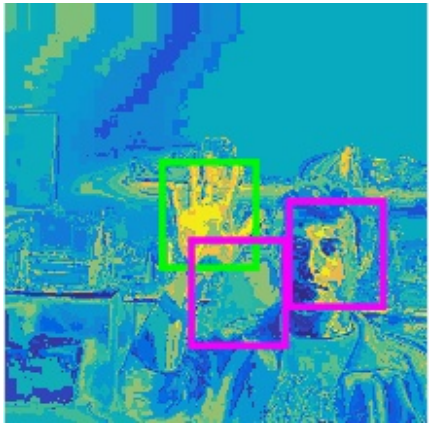
Sequences `bicycle`, `bolt`

Input I	$p^{1:t}_{O,S}(\mathbf{x} \in \mathcal{O} \mid b_{\mathbf{x}})$	$p^{1:t}_{O,D}(\mathbf{x} \in \mathcal{O} \mid b_{\mathbf{x}})$	Combined, CVPR'15
			
Input I	$p^{1:t}_{O,S}(\mathbf{x} \in \mathcal{O} \mid b_{\mathbf{x}})$	$p^{1:t}_{O,D}(\mathbf{x} \in \mathcal{O} \mid b_{\mathbf{x}})$	Combined, CVPR'15
			

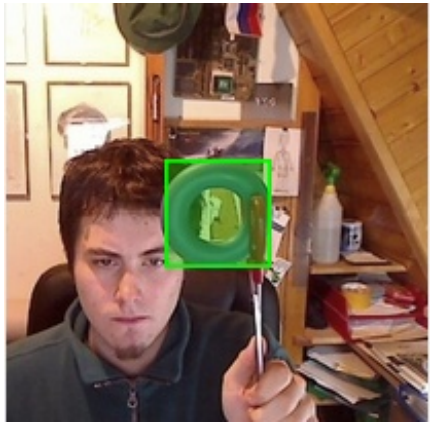
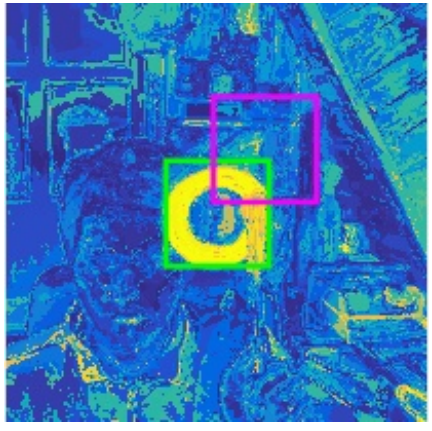
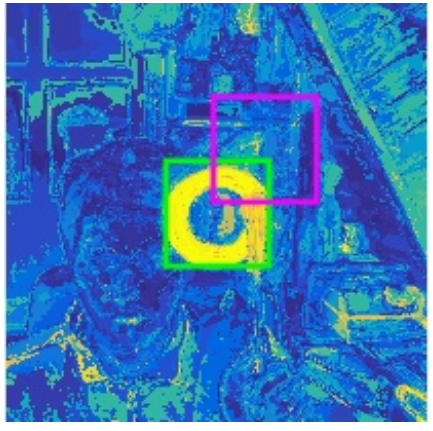
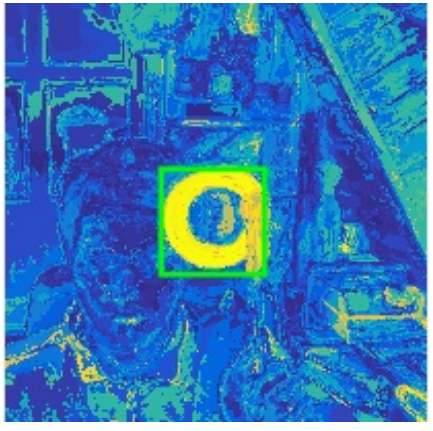
Localization

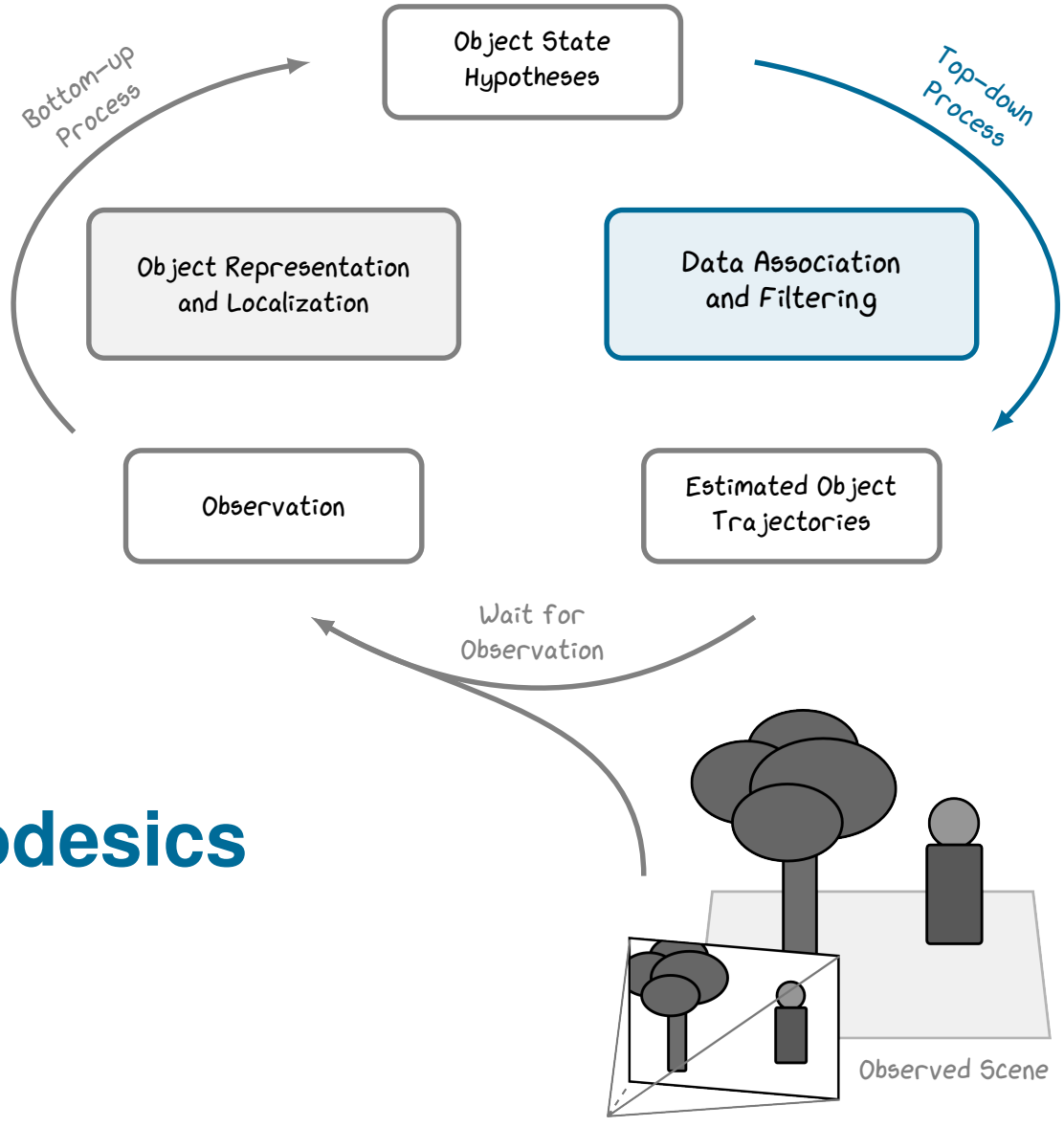
Sequences `hand2`, `torus`

Input I | $p^{1:t}_{O,S}(\mathbf{x} \in \mathcal{O} \mid b_{\mathbf{x}})$ | $p^{1:t}_{O,D}(\mathbf{x} \in \mathcal{O} \mid b_{\mathbf{x}})$ | Combined, CVPR'15

Input I | $p^{1:t}_{O,S}(\mathbf{x} \in \mathcal{O} \mid b_{\mathbf{x}})$ | $p^{1:t}_{O,D}(\mathbf{x} \in \mathcal{O} \mid b_{\mathbf{x}})$ | Combined, CVPR'15

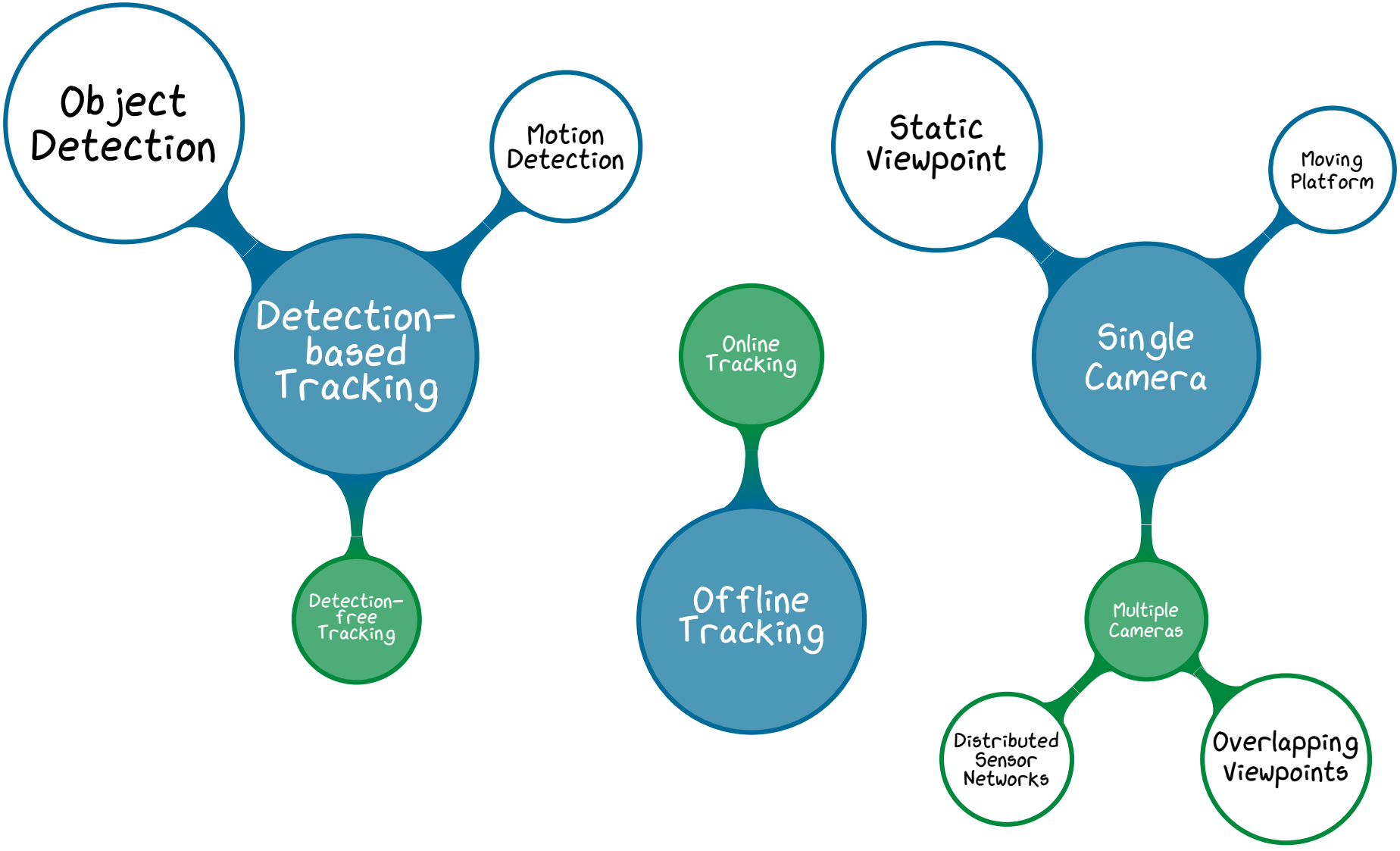


Occlusion Geodesics

Details & Extras

Multi-Object Tracking Paradigms

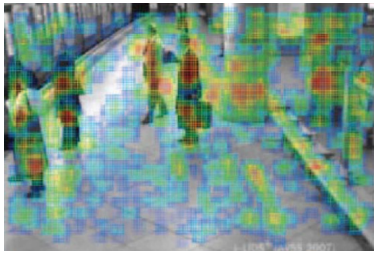
The Past Decade



Related Multi-Object Tracking Approaches

Tracking-by-Detection

- Traditionally causal:
  - High uncertainty $\rightarrow$ leverage probabilistic inference.
  - MHT, JPDAF, Kalman and particle filters.
  - Multiple hypotheses $\rightarrow$ often exponential state space.
- Recent focus on offline approaches:
  - Optimize assignments over batches.
  - Graph problems (KSP, MWIS, cuts, network flow, *etc.*)
- Closely related:
  - Leverage detector confidence [1].
  - Modeling group interactions, social force model [2,3].



Breitenstein [1]

Pellegrini [3]

[1] Breitenstein *et al. Online Multiperson Tracking-by-Detection from a Single, Uncalibrated Camera*. TPAMI 33(9), 2011.

[2] Leal-Taixé *et al. Everybody needs somebody: Modeling social and grouping behavior on a linear programming multiple people tracker*. ICCVW'11.

[3] Pellegrini *et al. You'll Never Walk Alone: Modeling Social Behavior for Multi-target Tracking*. ICCV'09.

MOT Benchmark Sequences

Sequence	Image Resolution	Frame Rate	Num. Frames	Num. Trajectories	Tracking Area
PETS'09 S2L1	768 × 576	7.0	795	19	19.1 × 16.0
PETS'09 S2L2	768 × 576	7.0	436	68	19.1 × 16.0
PETS'09 S2L3	768 × 576	7.0	240	44	19.1 × 16.0
TownCentre	1920 × 1080	2.5	450	227	36.0 × 19.0

- Publications:
  - PETS'09: Ferryman and Shahrokni, Winter-PETS'09.
  - TownCentre: Benfold and Reid, CVPR'11.

MOT Evaluation Measures

- MOT Challenge [1] framework & protocol.
- CLEAR MOT measures [2]:
  - Multiple Object Tracking Accuracy (MOTA $\in (-\infty, 1]$)

$$\text{MOTA} = 1 - \frac{\sum_{t=1}^{N} \text{FN}^t + \text{FP}^t + \text{IDS}^t}{\sum_{t=1}^{N} \text{GT}^t}$$

  - Multiple Object Tracking Precision (MOTP $\in [0, 1]$)

$$\text{MOTP} = 1 - \frac{\sum_{t=1}^{N} \sum_{i=1}^{\text{TP}^t} \|\mathbf{x}_{\text{G},i}^t - \mathbf{x}_{\text{T},m}^t\|_2}{\tau_d \sum_{t=1}^{N} \text{TP}^t}$$

- Trajectory quality measures [3]:
  - Mostly tracked (MT, $\geq 80\%$), partially tracked (PT) or mostly lost (ML, $< 20\%$).
  - Trajectory fragmentation (FM $\in \mathbb{Z}_0^+$).
  - Identity switches (IDS $\in \mathbb{Z}_0^+$).

---

[1] Leal-Taixé *et al. MOTChallenge 2015: Towards a Benchmark for Multi-Target Tracking*. arXiv abs/1504.01942, 2015.

[2] Bernardin and Stiefelhagen. *Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics*. EURASIP JIVP 2008(1), 2008.

[3] Wu and Nevatia. *Tracking of Multiple, Partially Occluded Humans based on Static Body Part Detection*. CVPR'06.

Default Parameter Settings

Parameter		Value
Conservative association threshold in [m/s]	$\tau_c \in (0, \infty)$	2.00
Physically feasible motion cut-off	$\tau_p \in [0, 1]$	10^{-4}
Plausible motion variance	$\sigma_p^2 \in (0, \infty)$	1.30
Directional motion variance	$\sigma_d^2 \in (0, 1]$	0.40
Detector belief factor	$\beta_d \in [0, 1]$	0.70

Ablation Study

Conservative Association Threshold (All Sequences)

τ_c	MOTA↑	MOTP↑	MT/GT↑	ML/GT↓	IDS↓	FM↓	FPS↑
0.50	0.49	0.65	0.39	0.16	657	571	6.1
1.00	0.49	0.65	0.41	0.16	640	562	8.2
1.50	0.48	0.65	0.40	0.15	599	544	9.6
2.00	0.48	0.65	0.38	0.16	576	561	12.8
2.50	0.47	0.65	0.38	0.16	599	536	15.3
3.00	0.47	0.65	0.37	0.16	612	532	16.9
3.50	0.47	0.65	0.36	0.16	640	554	19.1
4.00	0.47	0.66	0.36	0.16	622	553	19.9
4.50	0.48	0.65	0.38	0.15	640	570	20.3

Ablation Study

Feasible Movement Threshold (All Sequences)

τ_p	MOTA↑	MOTP↑	MT/GT↑	ML/GT↓	IDS↓	FM↓	FPS↑
10^{-7}	0.49	0.66	0.35	0.14	688	679	8.4
10^{-6}	0.49	0.66	0.35	0.14	688	679	8.4
10^{-5}	0.46	0.66	0.35	0.16	589	589	9.5
10^{-4}	0.48	0.65	0.38	0.16	576	561	12.8
10^{-3}	0.45	0.65	0.38	0.15	577	507	12.1
10^{-2}	0.40	0.64	0.38	0.15	505	471	12.2
10^{-1}	0.30	0.63	0.33	0.19	517	418	9.4

Ablation Study

Plausible Motion Variance (All Sequences)

σ_p^2	MOTA↑	MOTP↑	MT/GT↑	ML/GT↓	IDS↓	FM↓	FPS↑
0.10	0.23	0.62	0.26	0.24	525	331	10.4
0.30	0.29	0.63	0.33	0.20	512	410	8.4
0.50	0.36	0.64	0.37	0.16	535	463	9.4
0.70	0.41	0.64	0.39	0.15	537	495	12.4
0.90	0.46	0.65	0.39	0.15	547	505	12.4
1.10	0.47	0.66	0.40	0.14	559	502	12.8
1.30	0.48	0.65	0.38	0.16	576	561	12.8
1.50	0.47	0.65	0.35	0.15	609	576	12.6
1.70	0.48	0.66	0.35	0.15	603	613	9.7
1.90	0.48	0.65	0.36	0.14	651	645	8.4

Ablation Study

Directional Variance (All Sequences)

σ_d^2	MOTA↑	MOTP↑	MT/GT↑	ML/GT↓	IDS↓	FM↓	FPS↑
0.10	0.49	0.65	0.38	0.16	575	556	12.7
0.20	0.48	0.65	0.38	0.16	569	550	12.5
0.30	0.48	0.65	0.37	0.15	581	560	12.9
0.40	0.48	0.65	0.38	0.16	576	561	12.8
0.50	0.49	0.65	0.39	0.15	589	552	13.1
0.60	0.48	0.65	0.40	0.15	595	551	13.0
0.70	0.48	0.65	0.40	0.16	596	550	13.0
0.80	0.48	0.65	0.39	0.16	590	554	12.9
0.90	0.48	0.65	0.39	0.15	587	549	13.0
1.00	0.47	0.66	0.38	0.16	597	552	13.1

Ablation Study

Detector Belief Factor (All Sequences)

- ACF [1] detections.
- AUC $= 0.73$ across all sequences.

β_d	MOTA$^{\uparrow}$	MOTP$^{\uparrow}$	MT/GT$^{\uparrow}$	ML/GT$^{\downarrow}$	IDS$^{\downarrow}$	FM$^{\downarrow}$	FPS$^{\uparrow}$
0.00	0.47	0.65	0.38	0.15	560	547	12.8
0.10	0.47	0.65	0.38	0.15	558	543	12.7
0.20	0.46	0.65	0.38	0.15	575	545	12.7
0.30	0.46	0.65	0.37	0.15	570	545	12.8
0.40	0.47	0.65	0.38	0.15	574	547	12.9
0.50	0.46	0.65	0.38	0.15	580	545	12.8
0.60	0.47	0.65	0.38	0.16	584	553	12.9
0.70	0.48	0.65	0.38	0.16	576	561	12.8
0.80	0.48	0.65	0.38	0.16	581	557	13.3
0.90	0.48	0.65	0.37	0.15	575	555	13.2
1.00	0.46	0.65	0.31	0.15	756	637	13.3

[1] Dollár *et al. Fast Feature Pyramids for Object Detection*. TPAMI 36(8), 2014.

Ablation Study

Detector Influence (All Sequences)

Detector	MOTA↑	MOTP↑
DPM [1]	0.58	0.65
F-RCNN [2]	0.50	0.62
ACF [3]	0.48	0.65
R-FCN [4]	0.48	0.62
IKSVM [5]	0.46	0.61
Poselets [6]	0.46	0.64
LDCF [7]	0.35	0.64
SSD [8]	0.29	0.60
YOLO [9]	0.29	0.58

[1] Felzenszwalb *et al. Object Detection with Discriminatively Trained Part Based Models*. TPAMI 32(9), 2010.

[2] Ren *et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks*. NIPS'15.

[3] Dollár *et al. Fast Feature Pyramids for Object Detection*. TPAMI 36(8), 2014.

[4] Dai *et al. R-FCN: Object Detection via Region-based Fully Convolutional Networks*. NIPS'16.

[5] Maji *et al. Classification using Intersection Kernel Support Vector Machines is efficient*. CVPR'08.

[6] Bourdev and Malik. *Poselets: Body Part Detectors Trained Using 3D Human Pose Annotations*. ICCV'09.

[7] Nam *et al. Local Decorrelation for Improved Detection*. NIPS'14.

[8] Liu *et al. SSD: Single Shot MultiBox Detector*. ECCV'16.

[9] Redmon *et al. You Only Look Once: Unified, Real-Time Object Detection*. CVPR'16.

Ablation Study

Detector Influence (All Sequences)

Detector	MOTA↑	MOTP↑	MT/GT↑	ML/GT↓	IDS↓	FM↓
DPM [1]	0.58	0.65	0.28	0.23	429	489
F-RCNN [2]	0.50	0.62	0.24	0.34	326	459
ACF [3]	0.48	0.65	0.38	0.16	576	561
R-FCN [4]	0.48	0.62	0.25	0.23	550	556
IKSVM [5]	0.46	0.61	0.18	0.30	321	410
Poselets [6]	0.46	0.64	0.24	0.20	485	549
LDCF [7]	0.35	0.64	0.15	0.34	482	479
SSD [8]	0.29	0.60	0.08	0.41	465	512
YOLO [9]	0.29	0.58	0.08	0.39	442	618

[1] Felzenszwalb *et al. Object Detection with Discriminatively Trained Part Based Models*. TPAMI 32(9), 2010.

[2] Ren *et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks*. NIPS'15.

[3] Dollár *et al. Fast Feature Pyramids for Object Detection*. TPAMI 36(8), 2014.

[4] Dai *et al. R-FCN: Object Detection via Region-based Fully Convolutional Networks*. NIPS'16.

[5] Maji *et al. Classification using Intersection Kernel Support Vector Machines is efficient*. CVPR'08.

[6] Bourdev and Malik. *Poselets: Body Part Detectors Trained Using 3D Human Pose Annotations*. ICCV'09.

[7] Nam *et al. Local Decorrelation for Improved Detection*. NIPS'14.

[8] Liu *et al. SSD: Single Shot MultiBox Detector*. ECCV'16.

[9] Redmon *et al. You Only Look Once: Unified, Real-Time Object Detection*. CVPR'16.

Ablation Study

Detector Influence (All Sequences)

Detector	MOTA↑	MOTP↑	MT/GT↑	ML/GT↓	IDS↓	FM↓	FPS↑
DPM [1]	0.58	0.65	0.28	0.23	429	489	17.4
F-RCNN [2]	0.50	0.62	0.24	0.34	326	459	16.9
ACF [3]	0.48	0.65	0.38	0.16	576	561	12.7
R-FCN [4]	0.48	0.62	0.25	0.23	550	556	16.6
IKSVM [5]	0.46	0.61	0.18	0.30	321	410	17.5
Poselets [6]	0.46	0.64	0.24	0.20	485	549	16.0
LDCF [7]	0.35	0.64	0.15	0.34	482	479	10.1
SSD [8]	0.29	0.60	0.08	0.41	465	512	13.0
YOLO [9]	0.29	0.58	0.08	0.39	442	618	9.1

[1] Felzenszwalb *et al. Object Detection with Discriminatively Trained Part Based Models*. TPAMI 32(9), 2010.

[2] Ren *et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks*. NIPS'15.

[3] Dollár *et al. Fast Feature Pyramids for Object Detection*. TPAMI 36(8), 2014.

[4] Dai *et al. R-FCN: Object Detection via Region-based Fully Convolutional Networks*. NIPS'16.

[5] Maji *et al. Classification using Intersection Kernel Support Vector Machines is efficient*. CVPR'08.

[6] Bourdev and Malik. *Poselets: Body Part Detectors Trained Using 3D Human Pose Annotations*. ICCV'09.

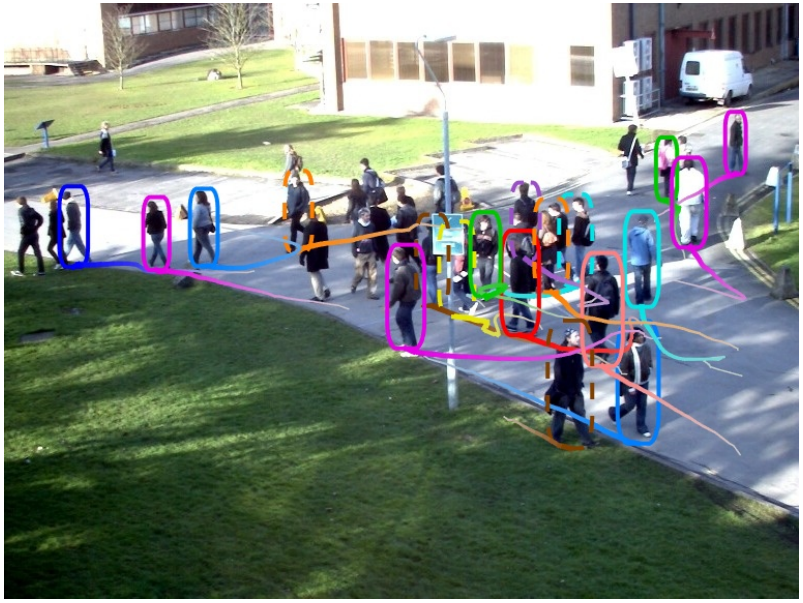
[7] Nam *et al. Local Decorrelation for Improved Detection*. NIPS'14.

[8] Liu *et al. SSD: Single Shot MultiBox Detector*. ECCV'16.

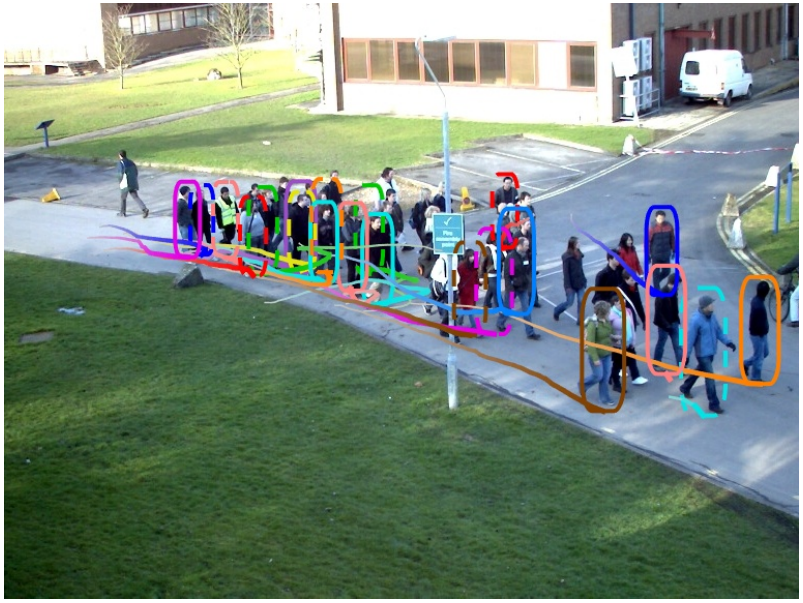
[9] Redmon *et al. You Only Look Once: Unified, Real-Time Object Detection*. CVPR'16.

Limitations

Crowded Scenes & Frequent Detector Failures

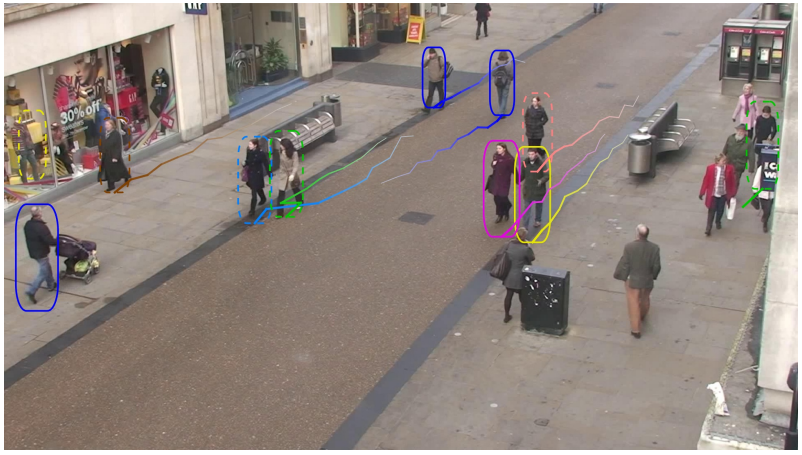


PETS'09 S2L2.



PETS'09 S2L3.

TownCentre.



TownCentre.