Video Annotation and Tracking with Active Learning

Carl Vondrick   Deva Ramanan   UC Irvine

How much does labeling 1 day of video cost?

Before seeing this poster:

- $15,000
- 8 man months

After seeing this poster:

- $1,500
- 24 man days

A novel algorithm that saves you $10,000 every day!

Big Idea: Not all frames are created equal. By only labeling the most important frames, we can economically annotate massive videos.

Contribution: We introduce Maximum Expected Label Change. ECL queries for labels on frames that could change the path the most.

Why not off-the-shelf active learning? 1) Video frames are structured, i.e. *non-i.i.d.* 2) Instead of right *model*, we want right *labels*.

Active Learning for Tracking

Score a path with appearance (HOG+color trained by SVM) and motion model:

$$E(b_{0:T}) = \sum_{t=0}^{T} U_t(b_t) + S(b_t, b_{t-1})$$
$$U_t(b_t) = \min\left(-w \cdot \phi_t(b_t), \alpha_1\right)$$
$$S(b_t, b_{t-1}) = \alpha_2 ||b_t - b_{t-1}||^2$$

Optimize efficiently with dynamic programming:

$$C_0^{\rightarrow}(b_0) = U_0(b_0)$$
$$C_t^{\rightarrow}(b_t) = U_t(b_t) + \min_{b_{t-1}} C_{t-1}^{\rightarrow}(b_{t-1}) + S(b_t, b_{t-1})$$
$$\pi_t^{\rightarrow}(b_t) = \operatorname*{argmin}_{b_{t-1}} C_{t-1}^{\rightarrow}(b_{t-1}) + S(b_t, b_{t-1})$$

Select the frame that would induce the largest expected change from the current best path:

$$t^* = \operatorname*{argmax}_{0 \le t \le T} \sum_{i=0}^{K} P(b_t^i) \cdot \Delta I(b_t^i)$$

Probability of annotating at a location is the cost of the path constrained by that point:

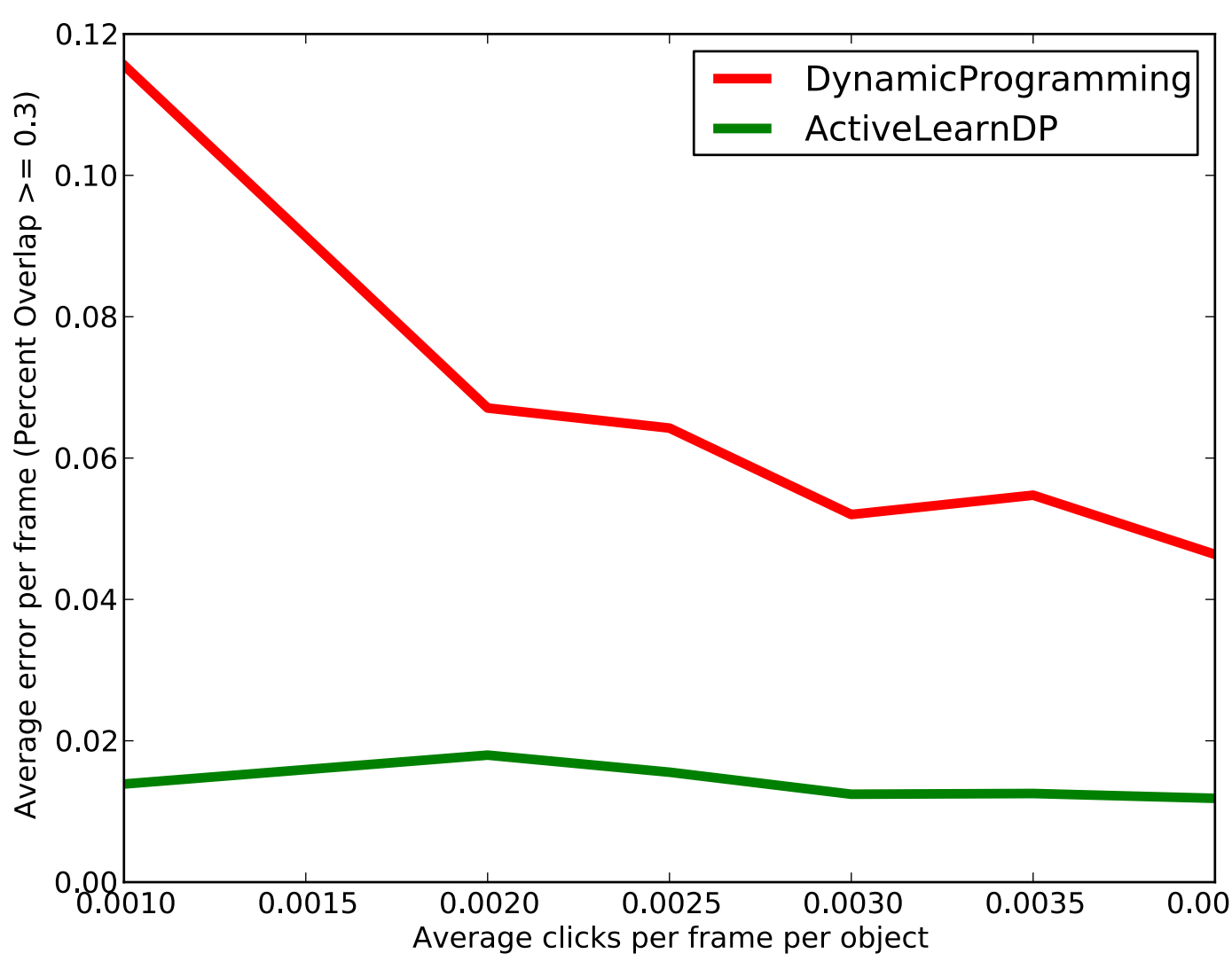
$$P(b_t^i) \propto \exp\left(\frac{-\Psi(b_t^i)}{\sigma^2}\right)$$
$$\Psi(b_t^i) = C_t^{\rightarrow}(b_t^i) + C_t^{\leftarrow}(b_t^i) - U(b_t^i)$$

Calculate how much the path would change:

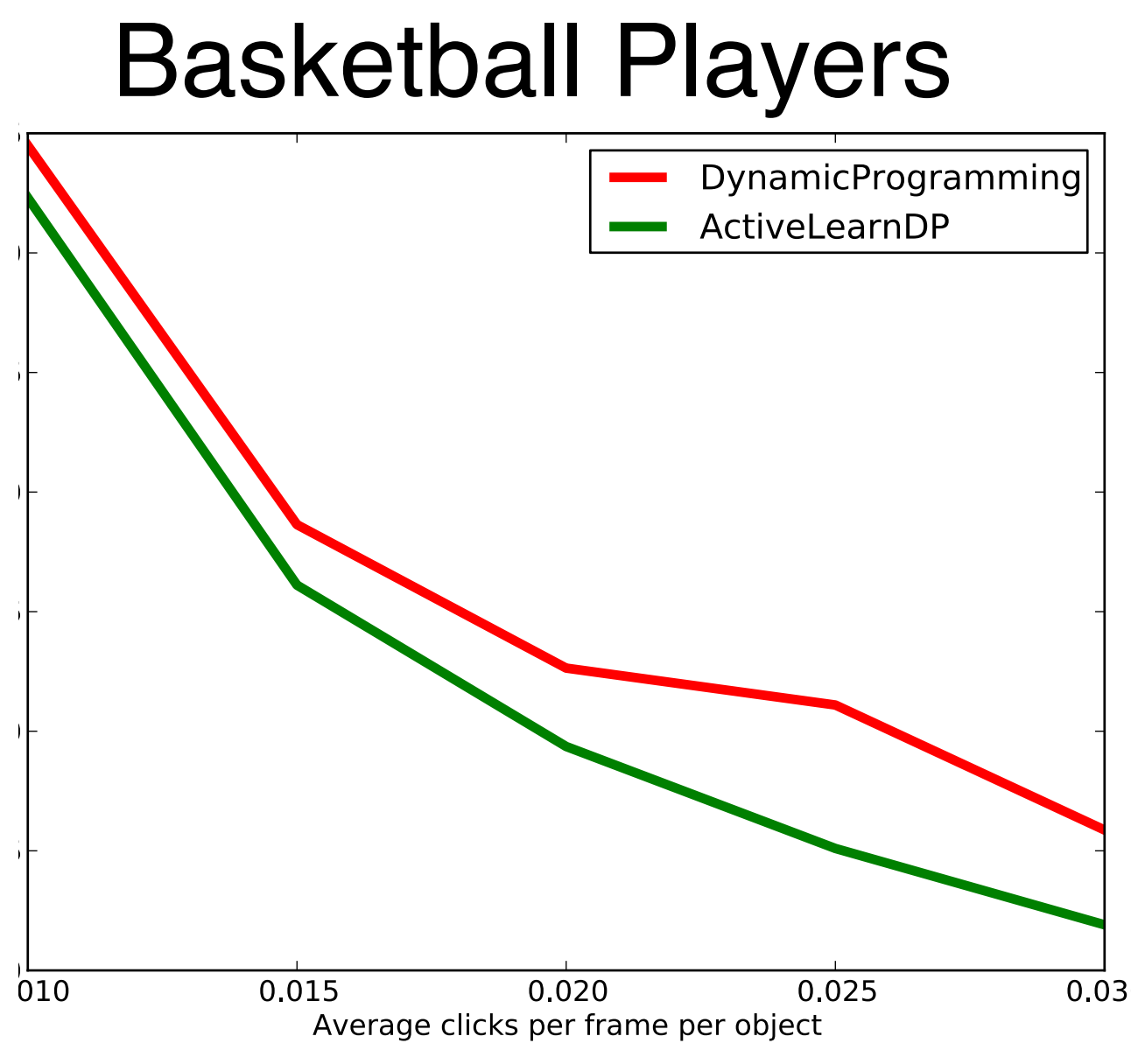
$$\Delta I(b_t^i) = \Theta_t^{\rightarrow}(b_t^i) + \Theta_t^{\leftarrow}(b_t^i) - \mathrm{err}(\mathrm{curr}_t, \mathrm{next}_t(b_t^i))$$
$$\Theta_0^{\rightarrow}(b_0) = \mathrm{err}(\mathrm{curr}_0, \mathrm{next}_0(b_0))$$
$$\Theta_t^{\rightarrow}(b_t) = \mathrm{err}(\mathrm{curr}_t, \mathrm{next}_t(b_t)) + \Theta_{t-1}^{\rightarrow}(\pi_t^{\rightarrow}(b_t))$$

Benchmark Evaluation

VIRAT Cars



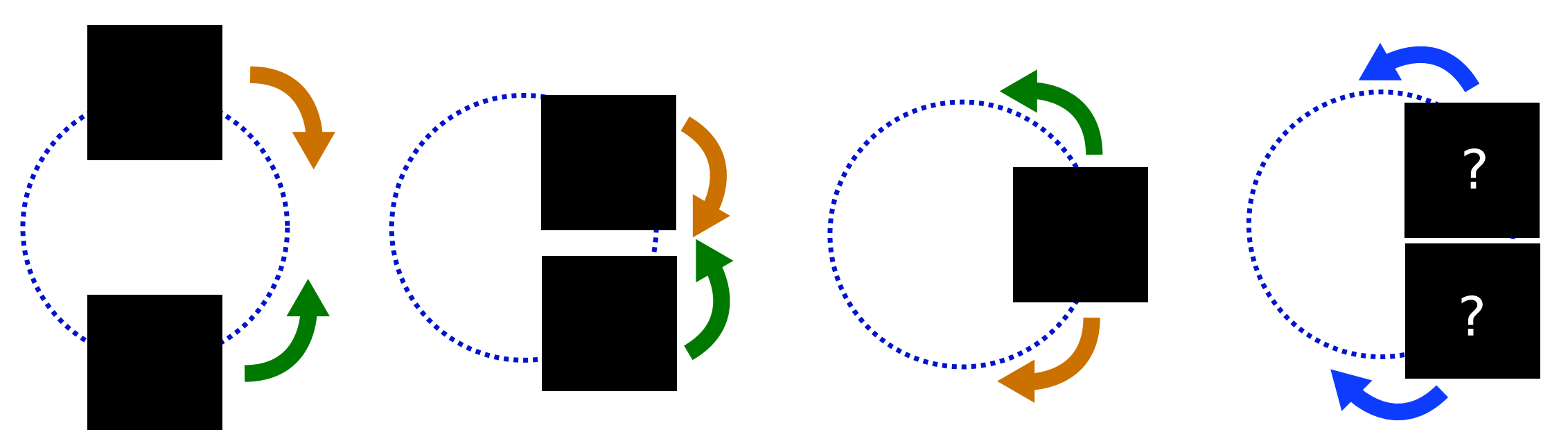
Basketball Players

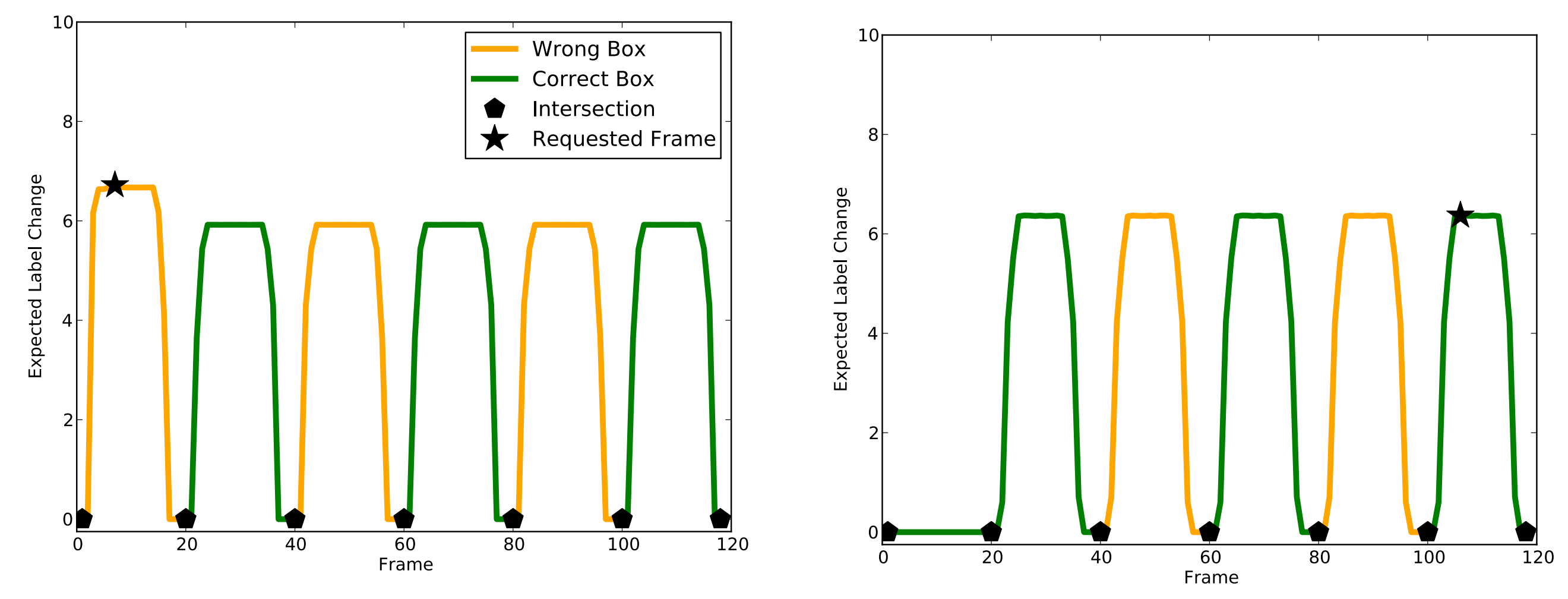


VIRAT

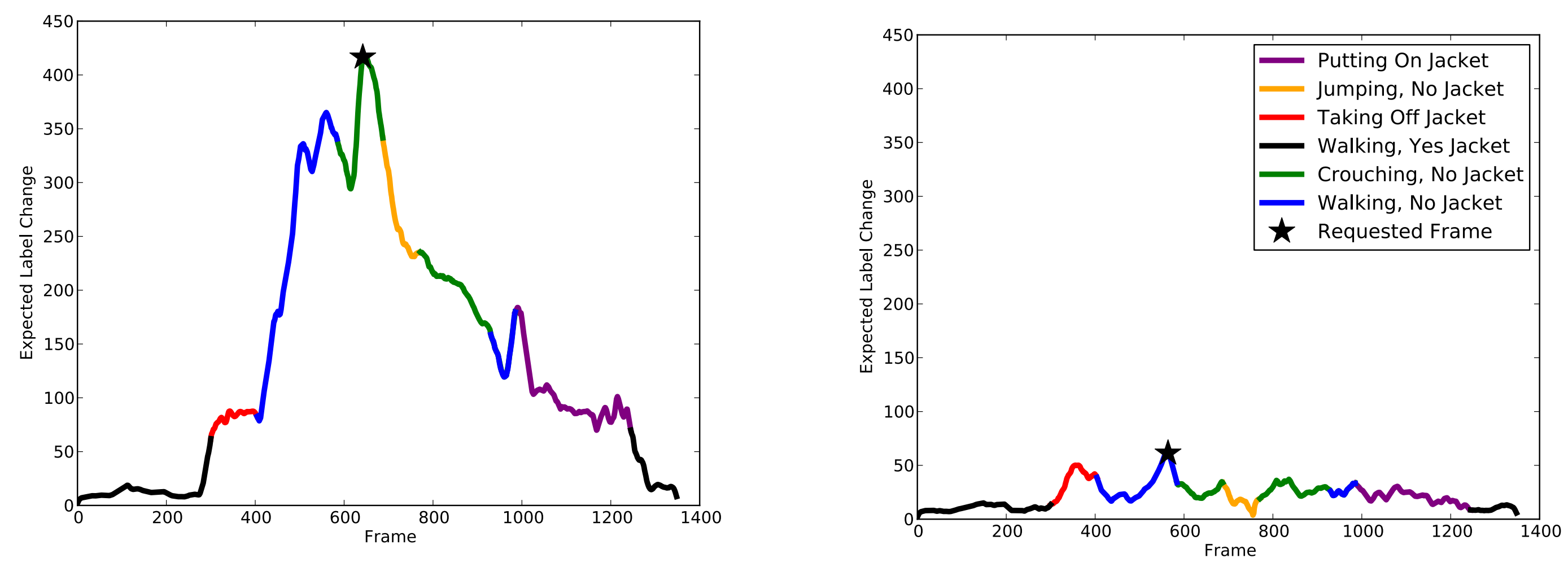
Basketball

Resolving Ambiguous Motion

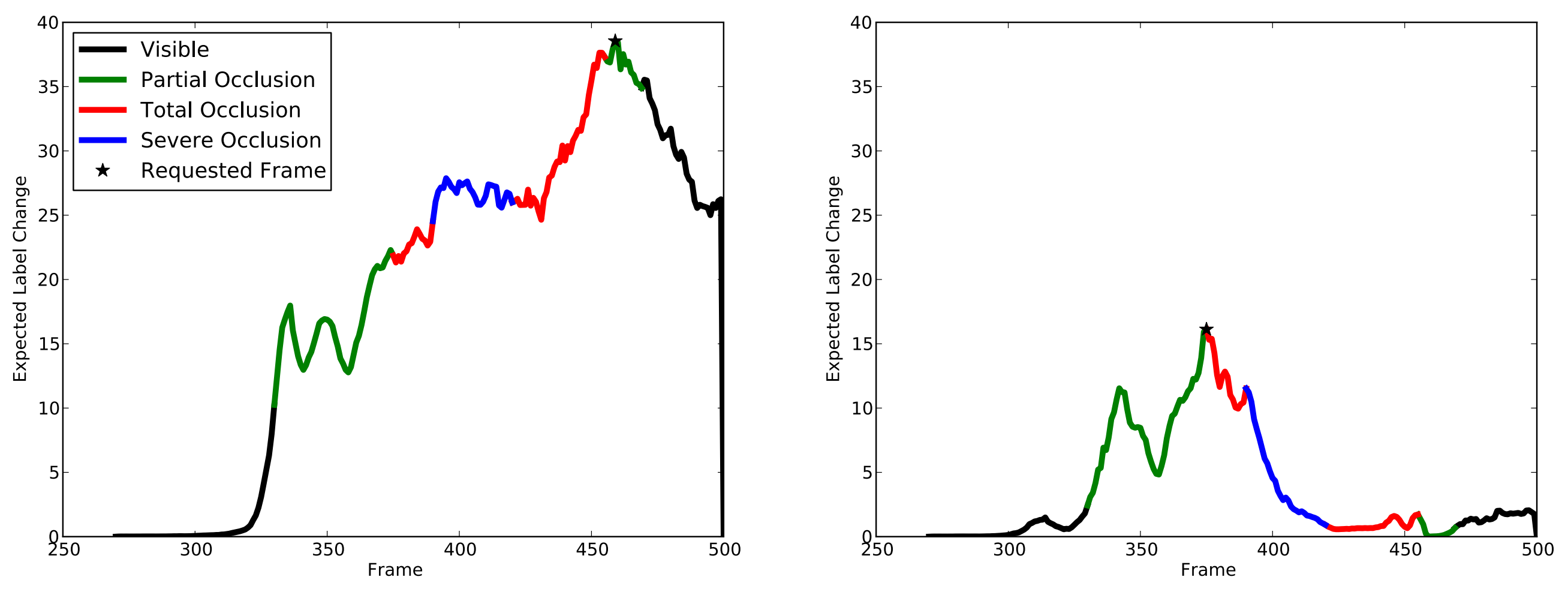




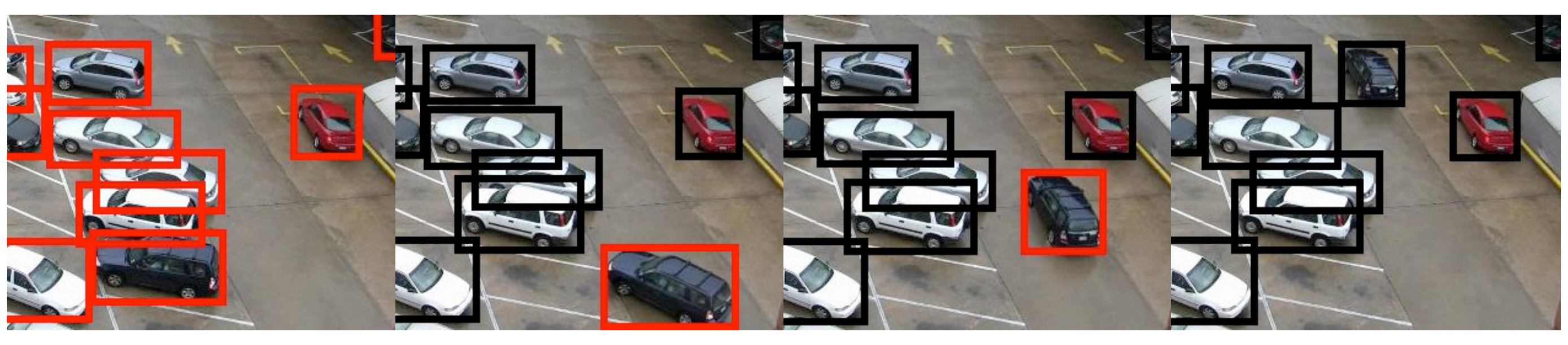
Tracking Under Deformation



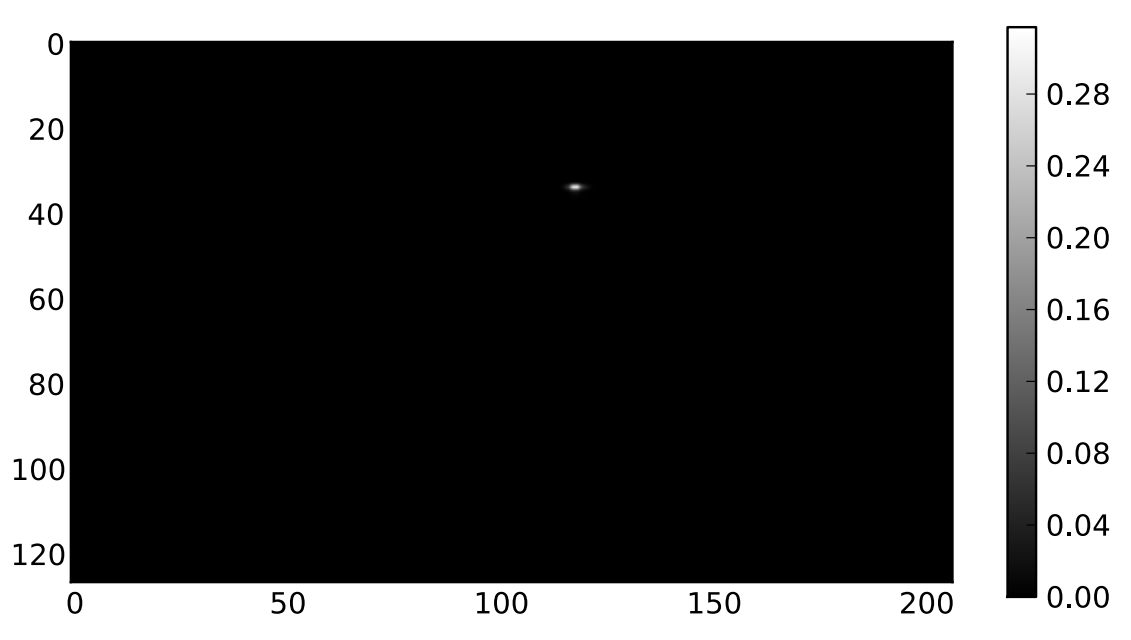
Tracking Under Occlusion



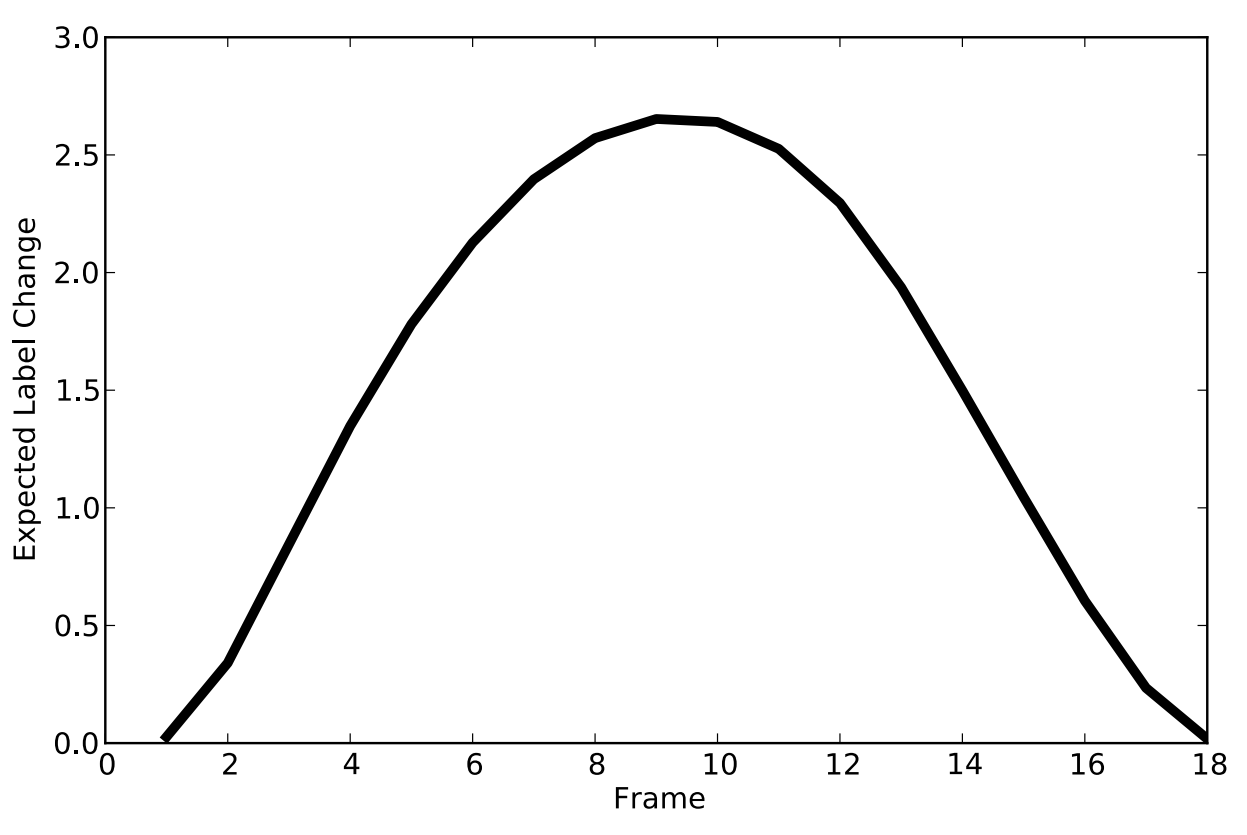
Transferring Wasted Clicks



Trivial Tracking



Cluttered Tracking



Annotation Tool: mit.edu/vondrick/vatic