# MBestStruck: M-Best Diverse Sampling for Structured Tracker

Ivan Bogun
http://my.fit.edu/~ibogun2010
Eraldo Ribeiro
http://cs.fit.edu/~eribeiro

Department of Computer Sciences and Cybersecurity
School of Computing
Florida Institute of Technology
Melbourne, U.S.A.

We approach the problem of *model-free* visual tracking of objects in videos. Model-free tracking has its state-of-the-art in a class of methods called *tracking-by-detection*, as shown in recent benchmarks. Some top-performing methods use deep neural networks (i.e., convnets) to solve the learning-based steps of the tracking algorithm (e.g., bounding box prediction and evaluation). Despite improving accuracy, *convnets* impose a high computational cost on trackers, limiting their real-time applications. In this paper, we propose to use deep features from a pre-learned deep-convolutional network in a computationally efficient way. Here, we use the *M-Best diverse-sampling* approach for sampling a small yet diverse set of bounding boxes that are likely to contain the object being tracked. These bounding boxes are then used by our method to perform detection using deep features. The resulting tracker, which we call *MBestStruck*, uses high-quality feature representation while being computationally efficient. Our tracking approach compares very well with the state-of-the-art, as we demonstrate by experiments done on popular benchmark datasets.

**M-Best-Diverse Labeling.** Let $E : \mathcal{Y} \to \mathbb{R}$ be an energy function that we define as a negative ObjStruck [2] discriminative function:

$$E(\mathbf{y}) = -\sum_{i,\bar{\mathbf{y}}} \beta_i^{\bar{\mathbf{y}}} \langle \phi(\mathbf{x_i}, \bar{\mathbf{y}}), \phi(\mathbf{x}, \mathbf{y}) \rangle - \\ - \lambda_s s(\mathbf{y}) - \lambda_e e(\mathbf{y}), \qquad (1)$$

where $\lambda_e, \lambda_s > 0$ are objectness parameters, and $s(\cdot), e(\cdot)$ are the straddling and the edge-density measures of objectness, respectively. Batra et al. [1] uses a greedy sequential procedure for finding $M$ diverse labelings, $\mathbf{y_1}, \ldots, \mathbf{y_M}$, according to the following criterion:

$$\mathbf{y}^m = \arg\min_{\mathbf{y} \in \mathcal{Y}} \left[ E(\mathbf{y}) - \lambda \sum_{i=1}^{m-1} \Delta(\mathbf{y}, \mathbf{y}^i) \right], \qquad (2)$$

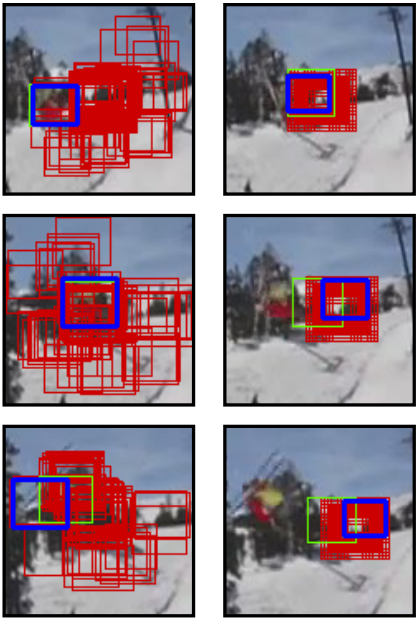

Figure 1: Left row: sampling using MBest procedure. Right row: sampling deterministically using linearly spaced bounding boxes.

for $i = 1, \ldots, M$, where parameter $\lambda > 0$ controls a trade-off between the diversity of the labelings and their quality. The function $\Delta(\cdot,\cdot) : \mathcal{Y} \times \mathcal{Y} \to \mathbf{R}$ is called a *dissimilarity kernel*.

We compared our method with the other trackers on benchmarks OTB50, OTB100, and VOT2015. Results show that our sampling strategy compares favorably to the state-of-the-art while using fewer bounding boxes for detection.

[1] Dhruv Batra, Payman Yadollahpour, Abner Guzman-Rivera, and Gregory Shakhnarovich. Diverse M-best solutions in Markov random fields. In *ECCV*, pages 1–16. Springer, 2012.

[2] Ivan Bogun and Eraldo Ribeiro. Object-aware tracking. *ICPR 2016 (to appear)*, 2016.