Unsupervised Online Multiple Animal Tracking

Dor Litvak, Or Dinari, and Oren Freifeld
The Deptartment of Computer Science, Ben-Gurion University of the Negev, Israel

CV4Animals Workshop CVPR JUNE 19-24 2022 NEW ORLEANS LOUISIANA

Introduction: Multiple Animal Tracking (MAT)

Goal: Track each individual animal present in the video.

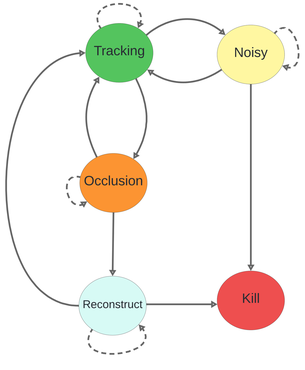
Key Challenges:
- data association;
- unknown # of animals;
- each animal may be present in any subset of the frames;
- the appearance, shape, and location of each trajectory can change across the video;
- occlusions;
- for animals, it is usually hard to obtain video-level supervision.

Contributions
- A new unsupervised online method for multiple animal tracking;
- A new dataset.

The MAT Dataset

- 15 movies with various animals;
- We provide ground-truth trajectories;
- We use it only for evaluation.

The Proposed State Machine



Unsupervised Multiple Animal Tracking (UMAT)

Our Unsupervised and Online Approach Combines:
- Tracking-by-detection
- Bayesian nonparametric (BNP) modeling:
  - Streaming-data BNP mixture inference for modeling appearance;
  - A velocity-informed Gaussian Processes Regression (GPR) for predicting bounding boxes;
  - A data-association mechanism related to DP means (a BNP extension of K-means).
- A new state machine for handling occlusions and detection noise.

Quantitative Results

Methods	MOTA ↑	HOTA ↑	IDF1 ↑	IDSW ↓	TP ↑	FN ↓	FP ↓
MaskTrack	39.681	44.817	48.138	265	14252	19769	**487**
UniTrack	84.333	69.464	76.781	354	29830	4191	785
GTR	80.336	70.628	75.273	334	29050	4971	1385
UMAT+Detectron2	**86.591**	**76.529**	79.356	**211**	30853	3168	1183
UMAT+HTC	85.447	75.708	**80.797**	215	**31085**	**2936**	1800
UMAT+SCNet	84.548	74.417	79.092	228	31039	2982	2047
UMAT+QueryInst	84.271	71.955	76.724	286	30195	3826	1239

Capturing Dynamics via a Velocity-informed GPR [1]

$\boldsymbol{x}_i = (x_i, y_i)$: center of the object i
$\boldsymbol{v}_i$: velocity.

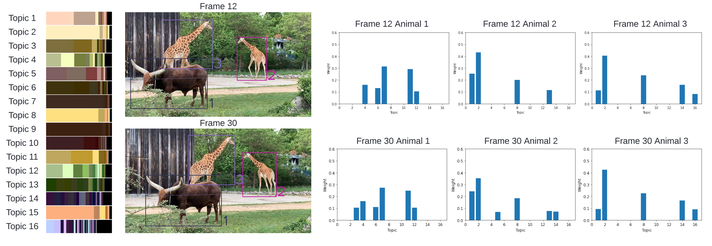
$$\boldsymbol{x}_i = f(t) \sim \mathcal{GP}\left(m(t), \kappa(t_i, t_j)\right) \quad (1)$$

$$\boldsymbol{v}_i = f'(t) \sim \mathcal{GP}'\left(m'(t), \kappa'(t_i, t_j)\right) \quad (2)$$

GPR predictive mean/variance ⇒ an ellipse the object center can be in.

Left: at first, the uncertainty is high (large ellipses). Right: as more measurements arrive, the uncertainty decreases (smaller ellipses).

Capturing Appearance via Streaming BNP Topic Modeling [2]



- N_t: # of visual topics at time t (in this example: $N_t = 16$).
- Visual words are assumed to be drawn *i.i.d.* from a visual topic model.

$$\sum_{k=1}^{K_t} \pi_{t,k}^{\text{global}} \, \text{topic}_{t,k} \,, \quad (3)$$

- The appearance model of a new Object $O_t + 1, i$ at time $t + 1$:

$$\sum_{k=1}^{K_t} \pi_{t,k}^{\text{O}_t+1,\text{i}} \, \text{topic}_{t,k} \,. \quad (4)$$

References

[1] C. Rasmussen and C Williams, "Gaussian processes for machine learning., (MIT Press: Cambridge, MA)," 2006.
[2] O. Dinari and O. Freifeld, "Sampling in Dirichlet process mixture models for clustering streaming data," in *AISTATS*, 2022.