Multiple Object Tracking

Guest lecture (NYU, neuroinformatics class)

Amin Nejatbakhsh

Current: Flatiron Research Fellow in the Center for Computational Neuroscience and Visiting Scholar at NYU
Past: Ph.D. in the Center for Theoretical Neuroscience at Columbia University

Research Interests: Statistics, Machine Learning, Dynamical Systems, Computer Vision, Neuroscience

March 24, 2024

Research Projects

Computer Vision

Neural Atlas Construction

Cell 2021
MICCAI 2020
Development 2021
IPMI 2023

Neural Segmentation

MICCAI 2020

Neural Registration

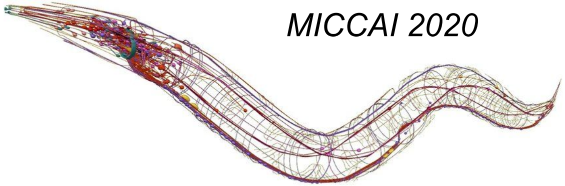
WACV 2021

Neural Signal Demixing

MICCAI 2020

Neural Tracking

Under Review



Stats & Neuroscience

Functional Causal Flow

Physical Review Research 2023

Causal Dynamical Systems

In Progress

Unique Information of Continuous Variables

NeurIPS 2021

Noise Correlation Estimation with Wishart Process

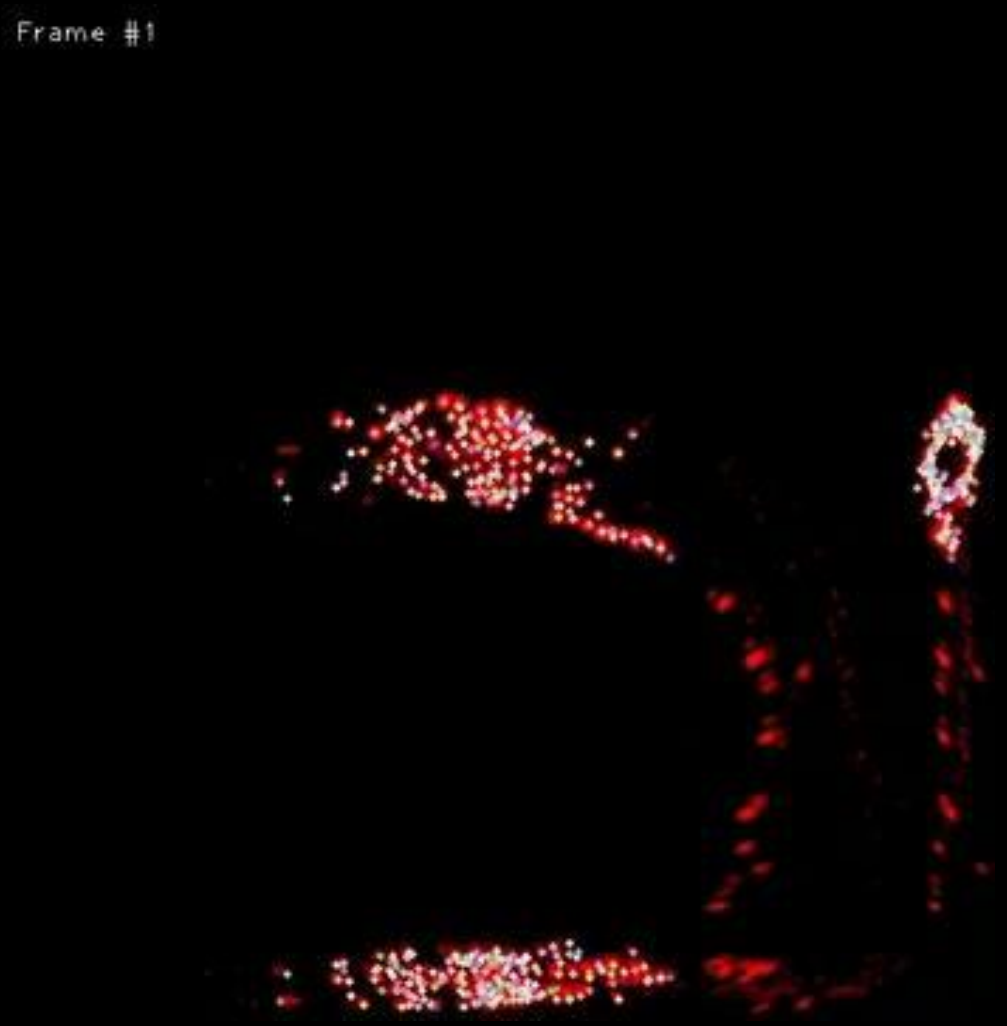
NeurIPS 2023

Dynamic Stochastic Shape Metrics

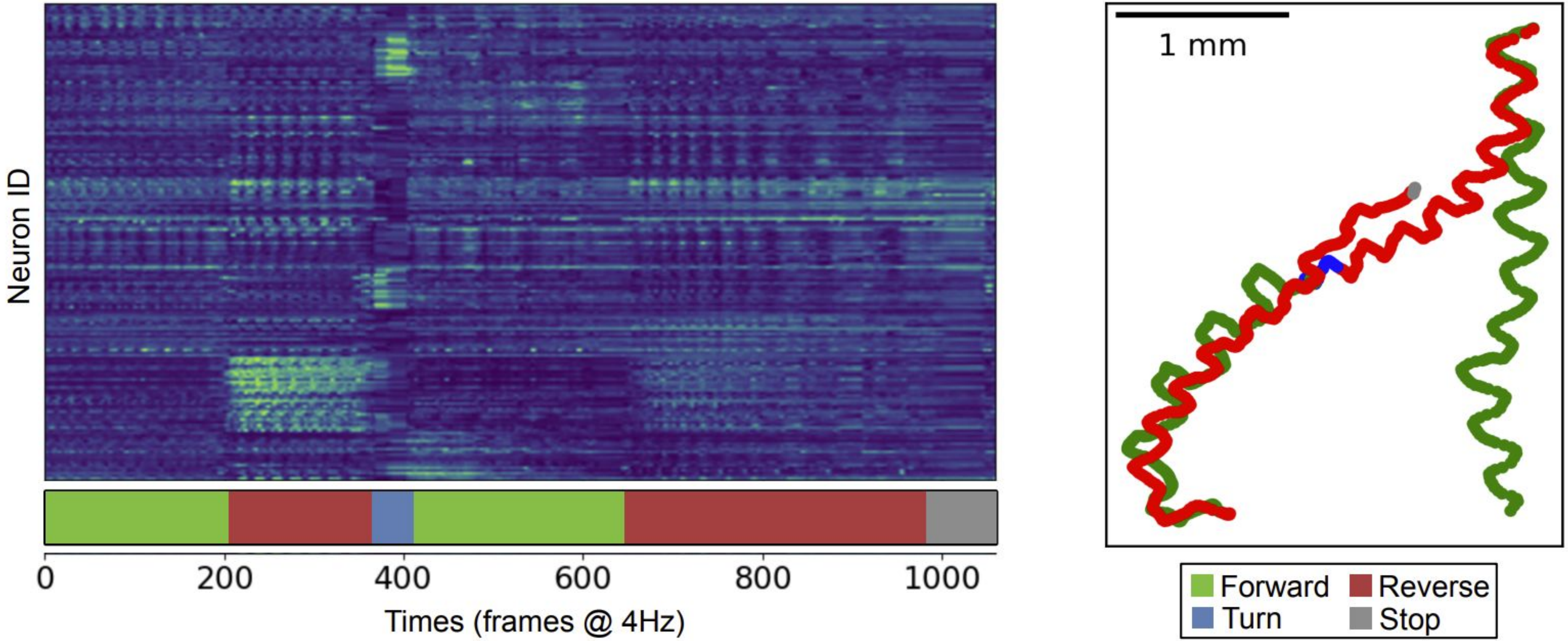
ICLR Workshops 2024

C. elegans Neural Tracking

James Yu, Amin Nejatbakhsh, … ***Under Review***



Tracking and Signal Extraction from Fully Moving *C. elegans*



James Yu, Amin Nejatbakhsh, … ***Under Review***

Introduction

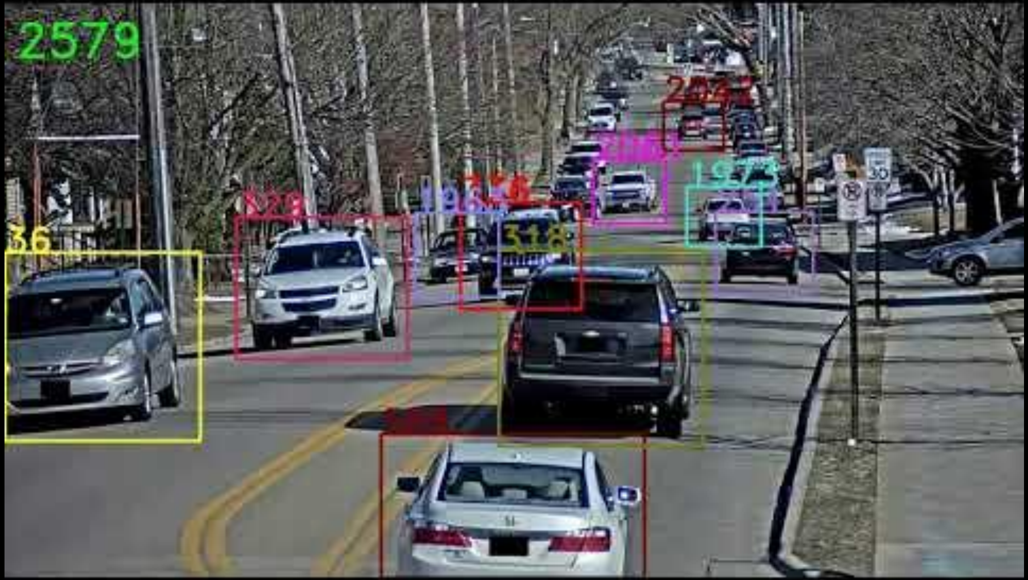
Multiple Object Tracking is not a single problem, it's a set of problems!

Let's look at a few examples to see why.

Vehicle Tracking

Properties

- Predictable motion patterns (linear models can be sufficient)
- Lack of unique appearance features
- Relative object size changes



Applications

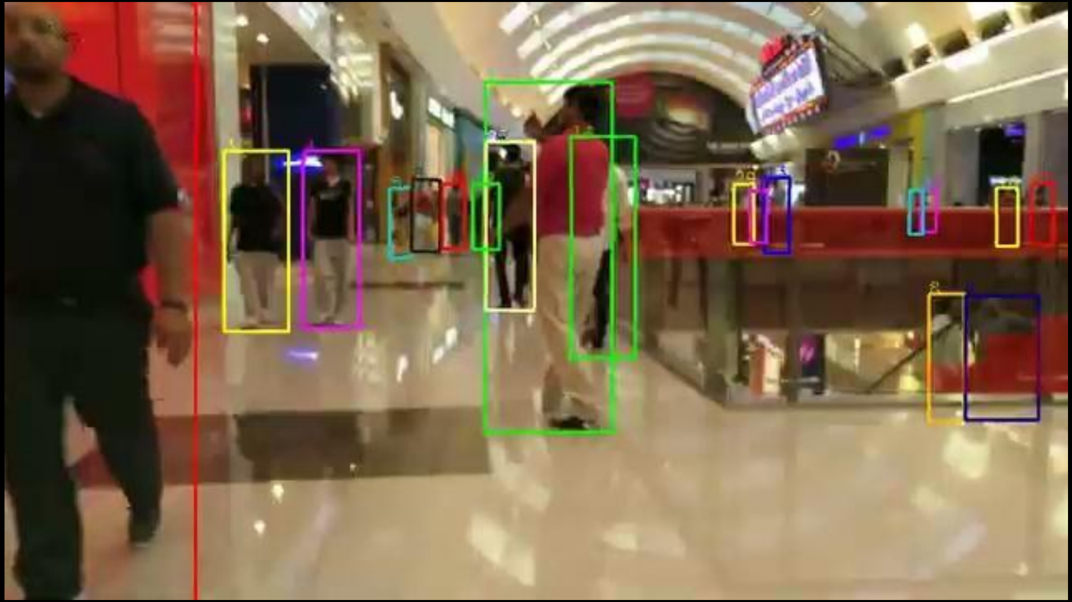
- Automated traffic monitoring

Lou et al, 2017

People in the Shopping Mall

Properties

- Egocentric view and camera angle changes
- Frequent birth and death events
- Missing data (occlusions)
- Background changes
- Noisy (Brownian) motion

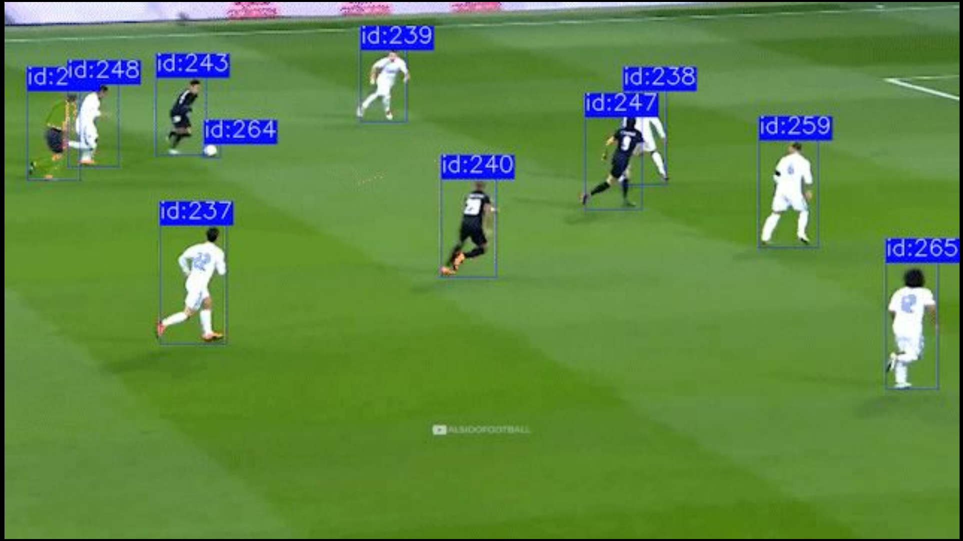


motchallenge

Tracking in Sports

Properties

- Sporadic sudden changes in the flow of the game (semi-noisy motion patterns)
- Large train/test distribution shift (in background, player jerseys, etc.)
- Lack of unique markers



Applications

- Collecting statistics
- Individual training

https://learnopencv.com/

Pedestrians on the Street

Properties

- Occlusions
- Birth and death events
- Out of plane rotations
- Unpredictable motion patterns
- Low spatial resolution
- Complex object transformations (humans walking or moving their arms)

Applications

- Automated monitoring

motchallenge

Zebrafish 3D Behavior Imaging

Properties

- Piecewise linear motion patterns
- Sparse spatial information
- Multi-camera recordings to avoid occlusions
- Accuracy is very important

Applications

- Understanding neural representations of behavior

motchallenge

Tracking Body Parts in Mouse

Properties

- Relative distances are fixed in 3 dimensions
- Multi-camera recordings

Applications

- Neural basis of motor control
- Understanding social behavior

International Brain Lab

Tracking Human Body Parts

Properties

- Relative distances are fixed in 3 dimensions

Applications

- Pose estimation
- Action recognition and classification

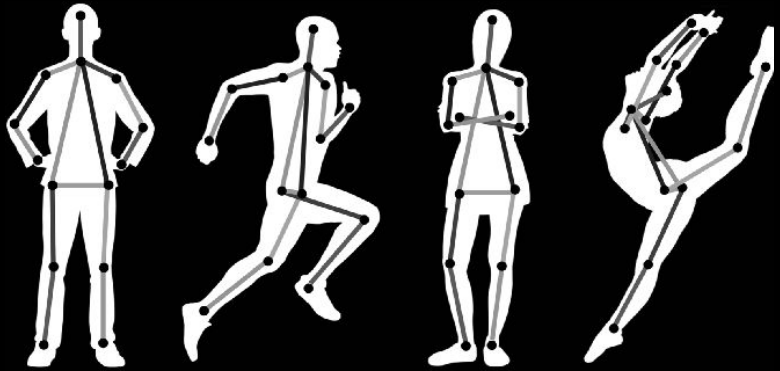


Photo credit: https://www.analyticsvidhya.com

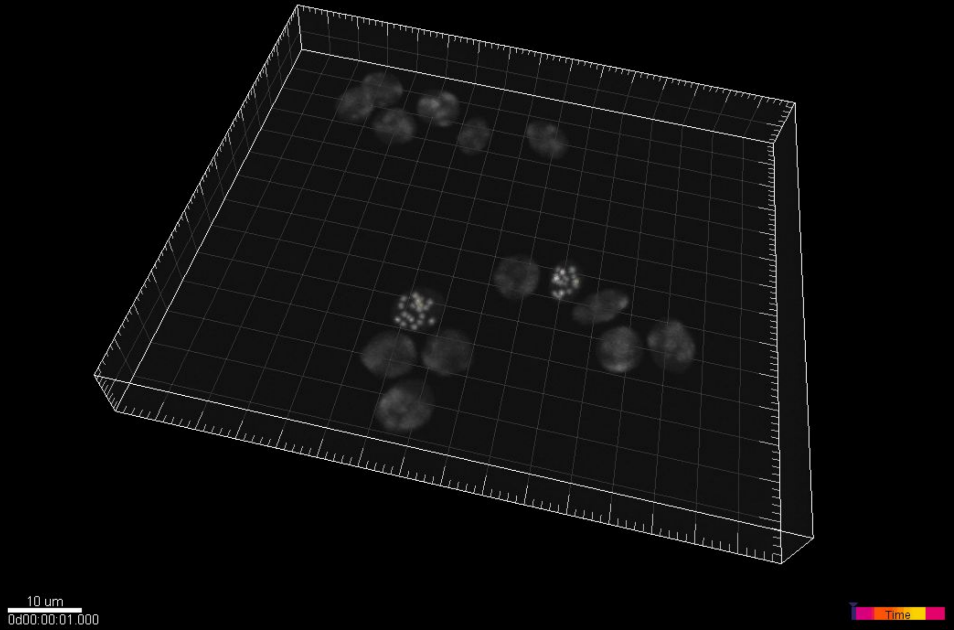
Cell Migration

Properties

- Cell division (frequent birth events)
- Lack of unique appearance and shape features
- Noisy motion

Applications

- Understanding cell development and migration



celltrackingchallenge

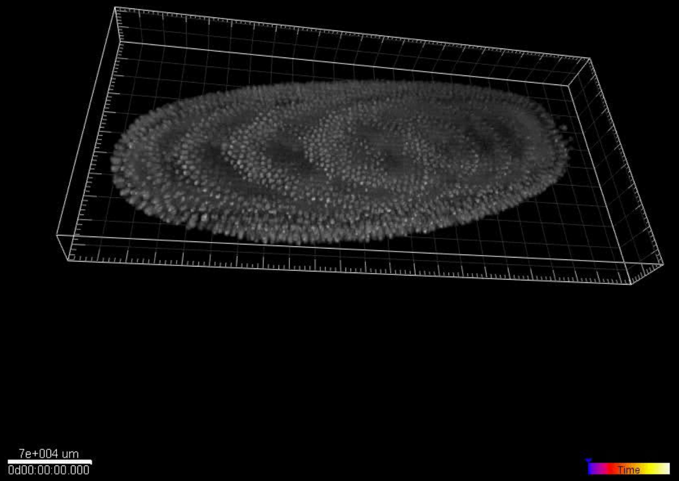
Developing Drosophila Melanogaster embryo

Properties

- Low spatial resolution

Application

- Extracting neural activities to understand neural basis of development

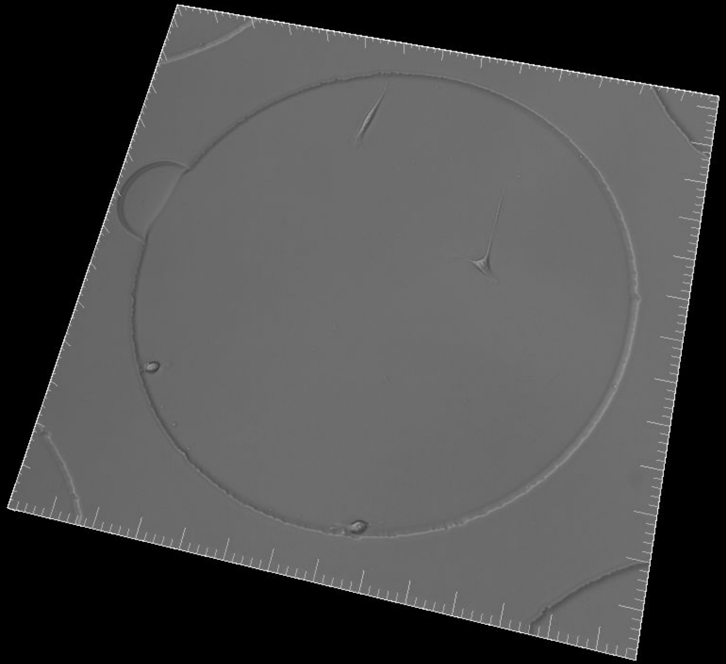


celltrackingchallenge

Mouse muscle stem cells in hydrogel microwells

Properties

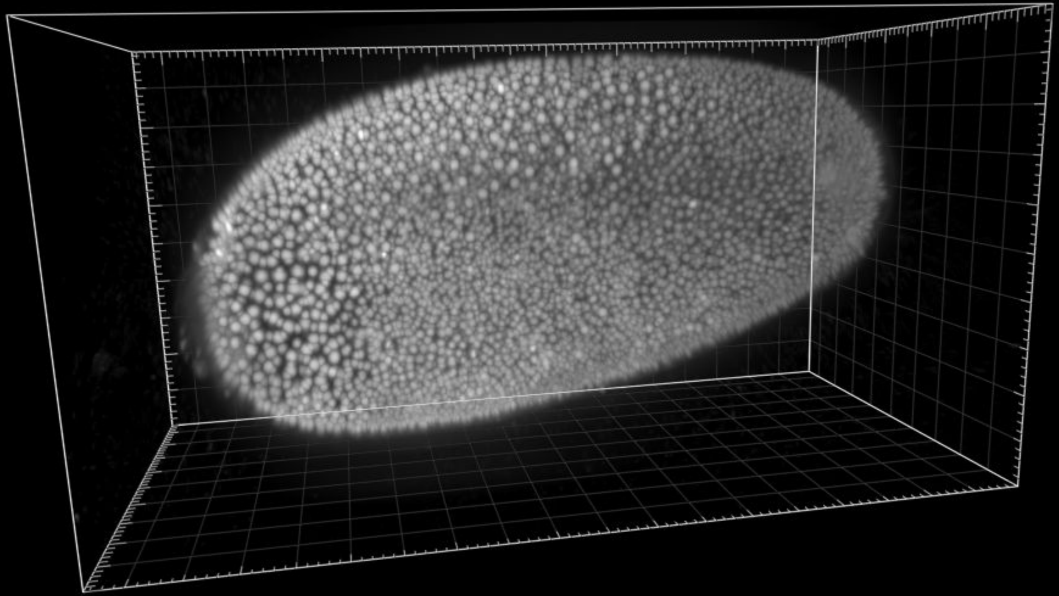
- Low temporal resolution



Developing Tribolium Castaneum embryo

Properties

- Complex motion and deformation

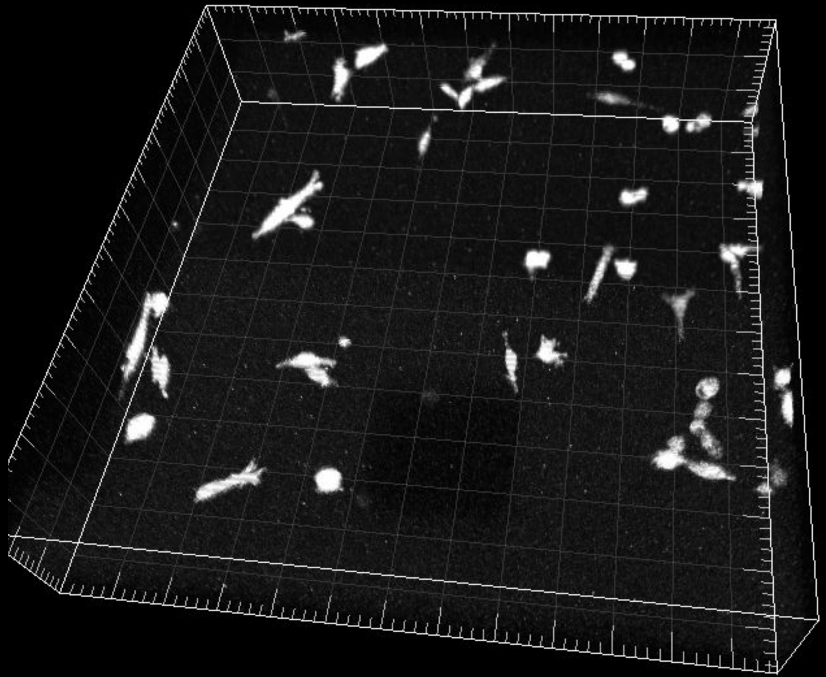


celltrackingchallenge

MDA231 human breast carcinoma cells

Properties

- Lack of unique shape and appearance markers
- No color information



celltrackingchallenge

Summary of challenges

- Low spatial or temporal resolution
- Diverse motion patterns (linear, nonlinear, piecewise linear, noisy)
- Lack of unique appearance or shape features
- Object transformations (relative size changes, out-of-plane rotations)
- Camera properties (egocentric view, multi-camera recordings)
- Frequent birth and death events
- Missing data and occlusions
- Train/test distribution shift (background changes)
- Spatial structure (relative distances fixed in 3 dimensions, complex motion and deformation)
- Online vs. offline tracking

<u>Important things to keep in mind</u>

Train vs. test distribution shift

- Lighting conditions
- Data coming from different labs/environments/cameras

Amount of training data (exploratory vs. deployed experiments)

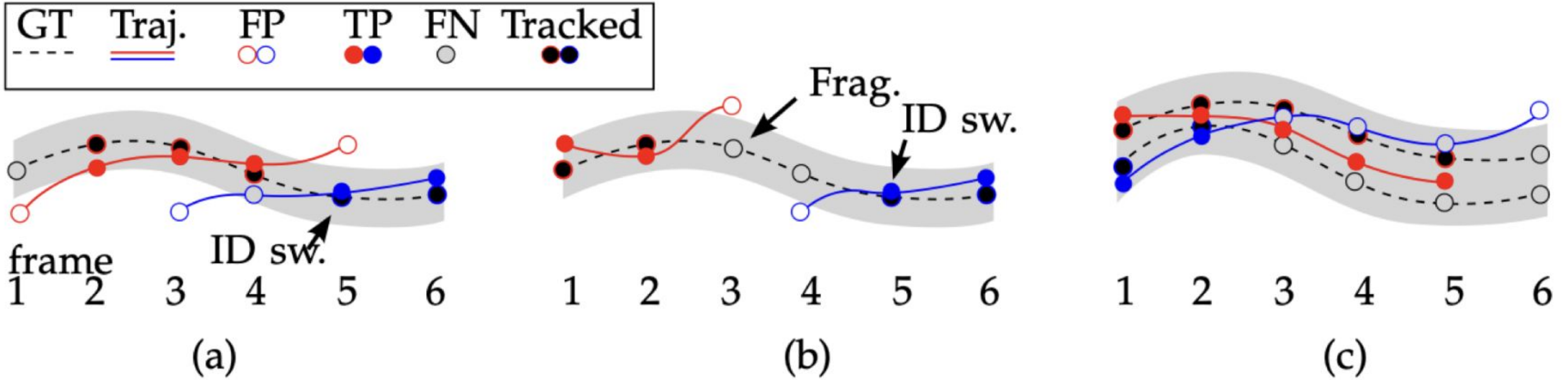
- Unsupervised (old school)
- Supervised (modern)
- Semi-supervised (SOTA)

Let's see how evaluation works before reviewing approaches

Recall ↑	Ratio of correctly matched detections to ground-truth detections
Precision ↑	Ratio of correctly matched detections to total result detections
MODP ↑	Average overlap between true positives and ground truth
MOTA ↑	**Combines false negatives, false positives and mismatch rate**
IDS ↓	**Number of times that a tracked trajectory changes its matched ground-truth identity (or vice versa)**
MOTP ↑	**Overlap between the estimated positions and the ground truth averaged over the matches**
TDE ↓	Distance between the ground-truth annotation and the tracking result
MT ↑	Percentage of ground-truth trajectories which are covered by the tracker output for more than 80% of their length
ML ↓	Percentage of ground-truth trajectories which are covered by the tracker output for less than 20% of their length

Luo et al, 2017

Let's see how evaluation works before reviewing approaches



ID switches are important, we want the tracked object to be stable across frames (different from accuracy evaluation in static images)

Leal (2021) - MOT Slides

Review of Existing Approaches

Unique markers (e.g. faces), perfect detection

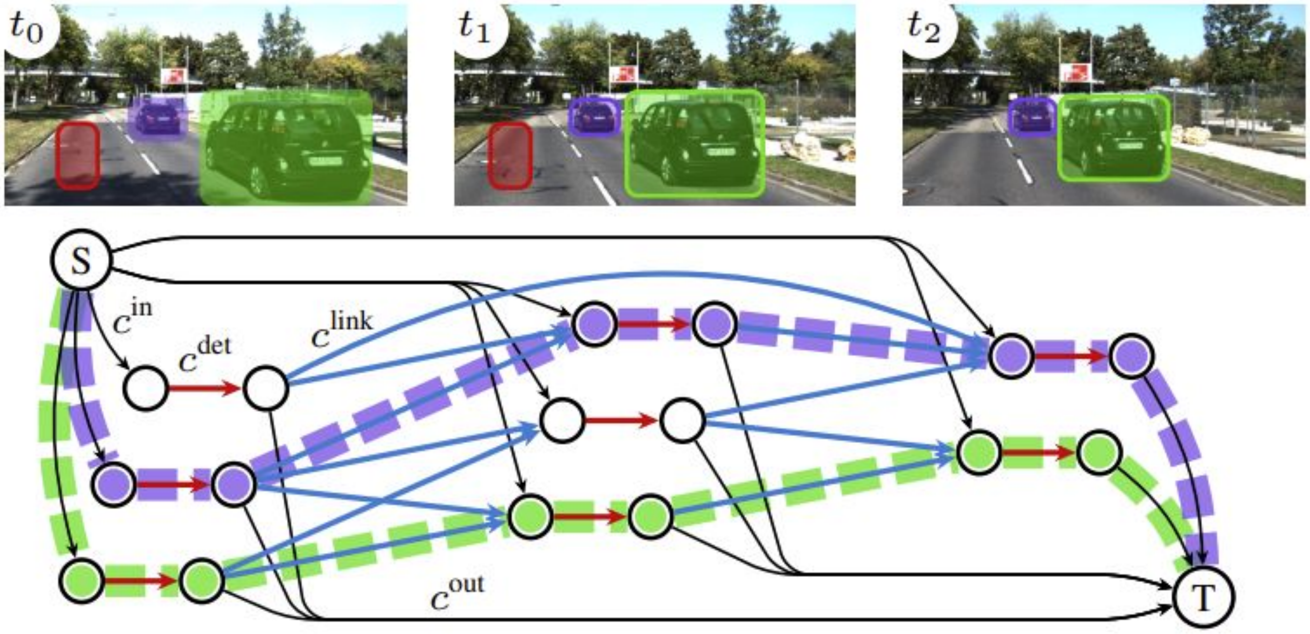
(1) Frame to frame matching

- Bipartite graph matching
- Hungarian algorithm

(2) Matching across all frames

- K-shortest paths
- Dynamic programming
- Max-flow network

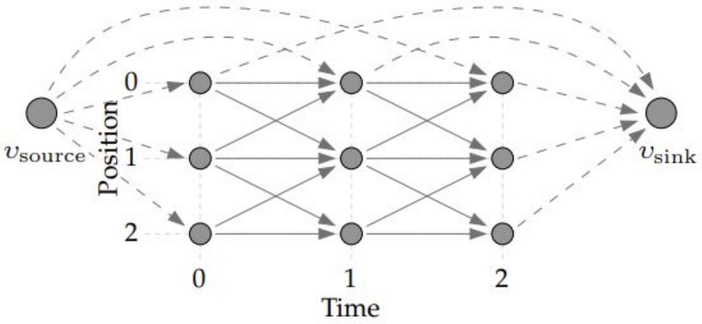
Graph construction



Schulter et al, 2017

Bipartite Graph Matching

Graph Construction



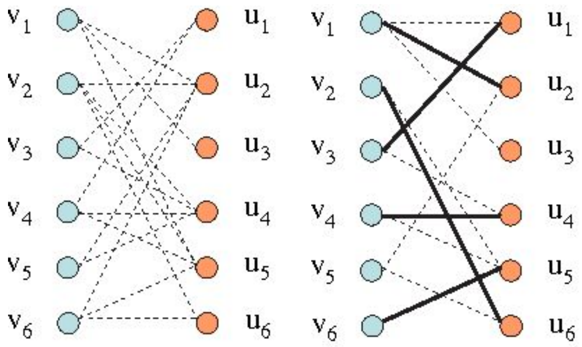
How to handle occlusions?

How to handle cell divisions

How to handle birth and death

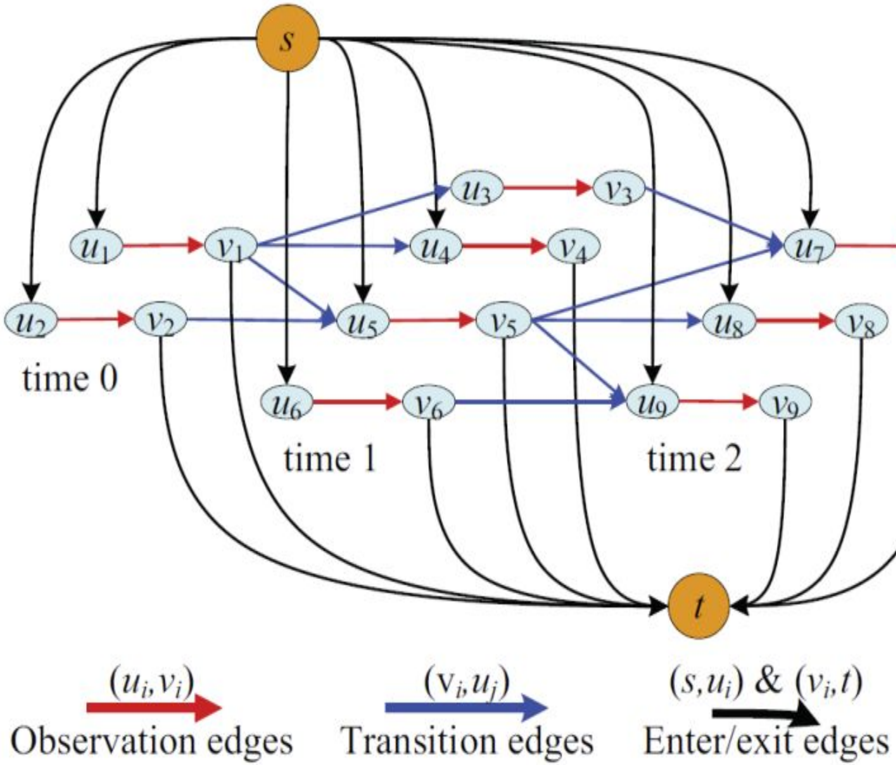
Hungarian Algorithm

w_{ij}	u_1	u_2	u_3	u_4	u_5	u_6
v_1	0	8	9	0	6	0
v_2	0	4	0	5	5	8
v_3	7	0	0	9	0	0
v_4	3	0	0	8	8	0
v_5	0	6	0	7	0	0
v_6	0	8	0	0	9	3



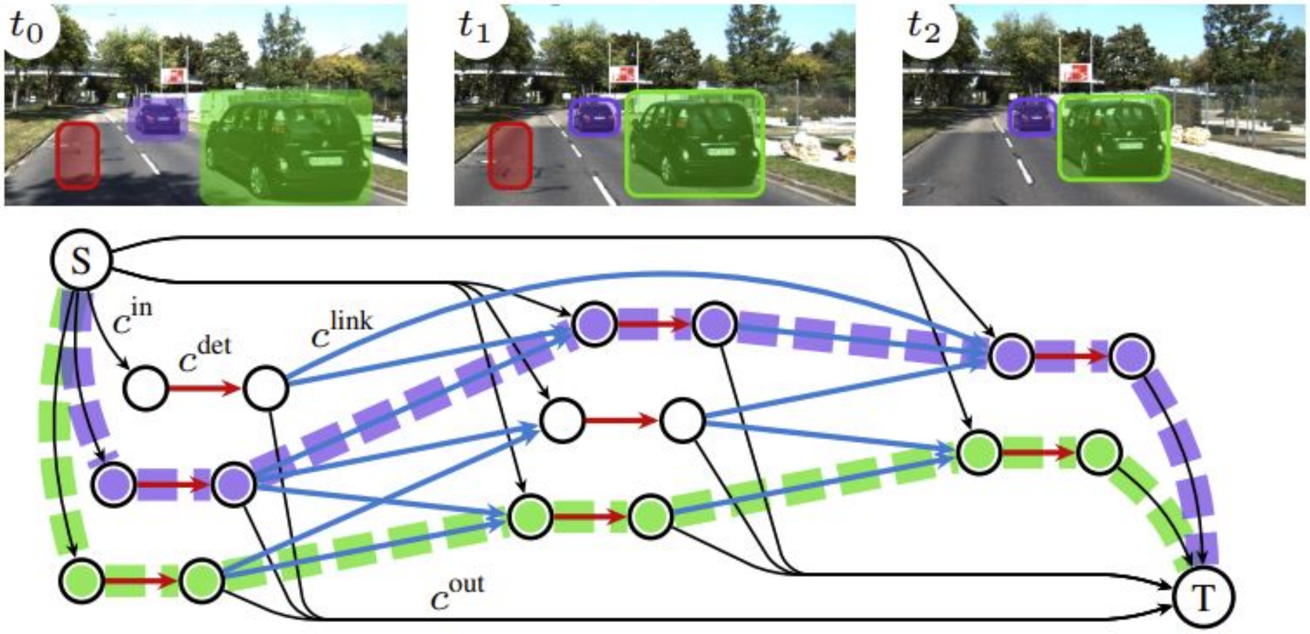
Mills-Tettey et al, 2007

Graph Construction: A More Complete View



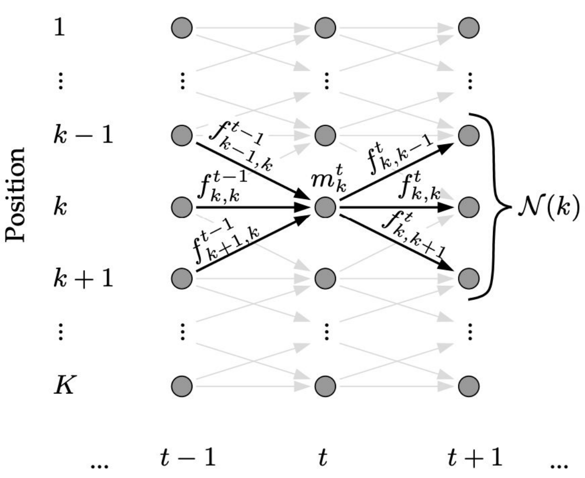
Zhang et al, 2008

K-Shortest Paths



Schulter et al, 2017

Linear/integer programming



Linear/integer program

$$\begin{aligned}
\text{Maximize} \quad & \sum_{t,i} \log\left(\frac{\rho_i^t}{1-\rho_i^t}\right) \sum_{j\in\mathcal{N}(i)} f_{i,j}^t \\
\text{subject to} \quad & \forall t,i,j, \quad f_{i,j}^t \geq 0 \\
& \forall t,i, \quad \sum_{j\in\mathcal{N}(i)} f_{i,j}^t \leq 1 \\
& \forall t,i, \quad \sum_{j\in\mathcal{N}(i)} f_{i,j}^t - \sum_{k:i\in\mathcal{N}(k)} f_{k,i}^{t-1} \leq 0 \\
& \sum_{j\in\mathcal{N}(\upsilon_{\text{source}})} f_{\upsilon_{\text{source}},j} - \sum_{k:\upsilon_{\text{sink}}\in\mathcal{N}(k)} f_{k,\upsilon_{\text{sink}}} \leq 0\,.
\end{aligned}$$

KSP formulation

$$\text{cost}(P_l) = \sum_{i=1}^{l} \text{cost}(p_i^*)\,.$$

$$\text{cost}(p_l^*) = \sum_{e_{i,j}^t \in p_l^*} c(e_{i,j}^t)\,.$$

IP: NP complete
KSP: O(k(m + n log n))
Min-Cut: O(kn^{2}m log n)
LP: polynomial-time

Berclaz et al, 2010

What if detections are not perfect? Crowded scenes or fast videos

(1) Incorporate more into the distance/cost

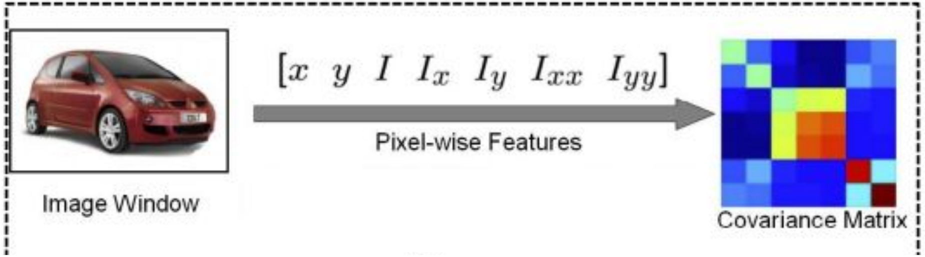
- Position
- Color or color-derived features
- Gradient/flow features
- Representational distance

(2) Learn the cost

- Linear program with training data

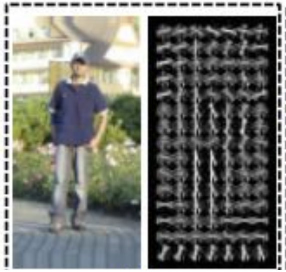
More complex distances and appearance models

Optical flow



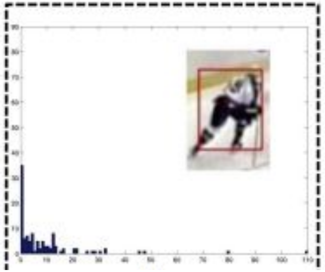
Covariance matrix

Point features



Gradient based features (HOG

Depth features



Color features

Luo et al, 2017

K-Shortest Paths Cost Learning

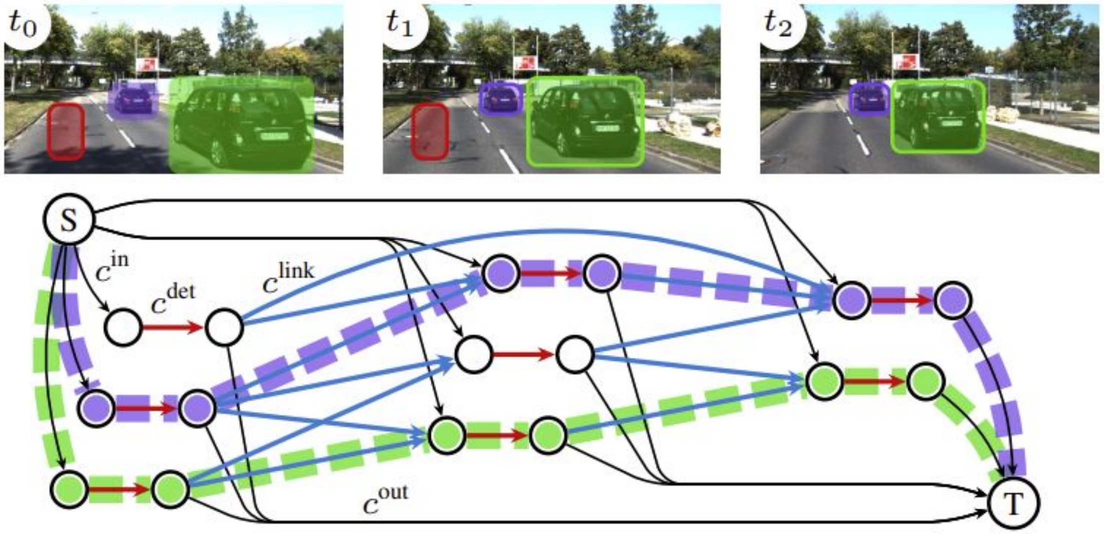
Optimization

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} \mathbf{c}^\top \mathbf{x}$$
$$\text{s.t.} \quad \mathbf{A}\mathbf{x} \leq \mathbf{b}, \; \mathbf{C}\mathbf{x} = \mathbf{0},$$

Cost Learning

$$\arg\min_{\Theta} \mathcal{L}\left(\mathbf{x}^{gt}, \mathbf{x}^*\right)$$
$$\text{s.t.} \quad \mathbf{x}^* = \arg\min_{\mathbf{x}} \mathbf{c}(\mathbf{f}, \Theta)^\top \mathbf{x}$$
$$\mathbf{A}\mathbf{x} \leq \mathbf{b}, \; \mathbf{C}\mathbf{x} = \mathbf{0},$$

Graph Construction



Schulter et al, 2017

What if there are no unique markers, detection is really bad, and images are noisy?

(1) Use probabilistic formulation

- State space models and Kalman filter
- Particle filtering

(2) Incorporate spatial and temporal structure

- Conditional random fields
- Quadratic programming
- Temporal smoothness
- Impose motion model (linear, piecewise linear, etc.)

Probabilistic Formulation

States

$$\mathbf{S}_t = (\mathbf{s}_t^1, \mathbf{s}_t^2, ..., \mathbf{s}_t^{M_t})$$

Observations

$$\mathbf{O}_t = (\mathbf{o}_t^1, \mathbf{o}_t^2, ..., \mathbf{o}_t^{M_t})$$

Inference

$$\widehat{\mathbf{S}}_{1:t} = \underset{\mathbf{S}_{1:t}}{\arg\max}\, P\left(\mathbf{S}_{1:t} | \mathbf{O}_{1:t}\right).$$

Important benefit: cost function is automatically given

$$\textbf{Predict: } P(\mathbf{S}_t|\mathbf{O}_{1:t-1}) = \int P(\mathbf{S}_t|\mathbf{S}_{t-1}) P(\mathbf{S}_{t-1}|\mathbf{O}_{1:t-1}) d\mathbf{S}_{t-1},$$

$$\textbf{Update: } P(\mathbf{S}_t|\mathbf{O}_{1:t}) \propto P(\mathbf{O}_t|\mathbf{S}_t) P(\mathbf{S}_t|\mathbf{O}_{1:t-1}).$$

Emissions (a.k.a. Appearance Model)

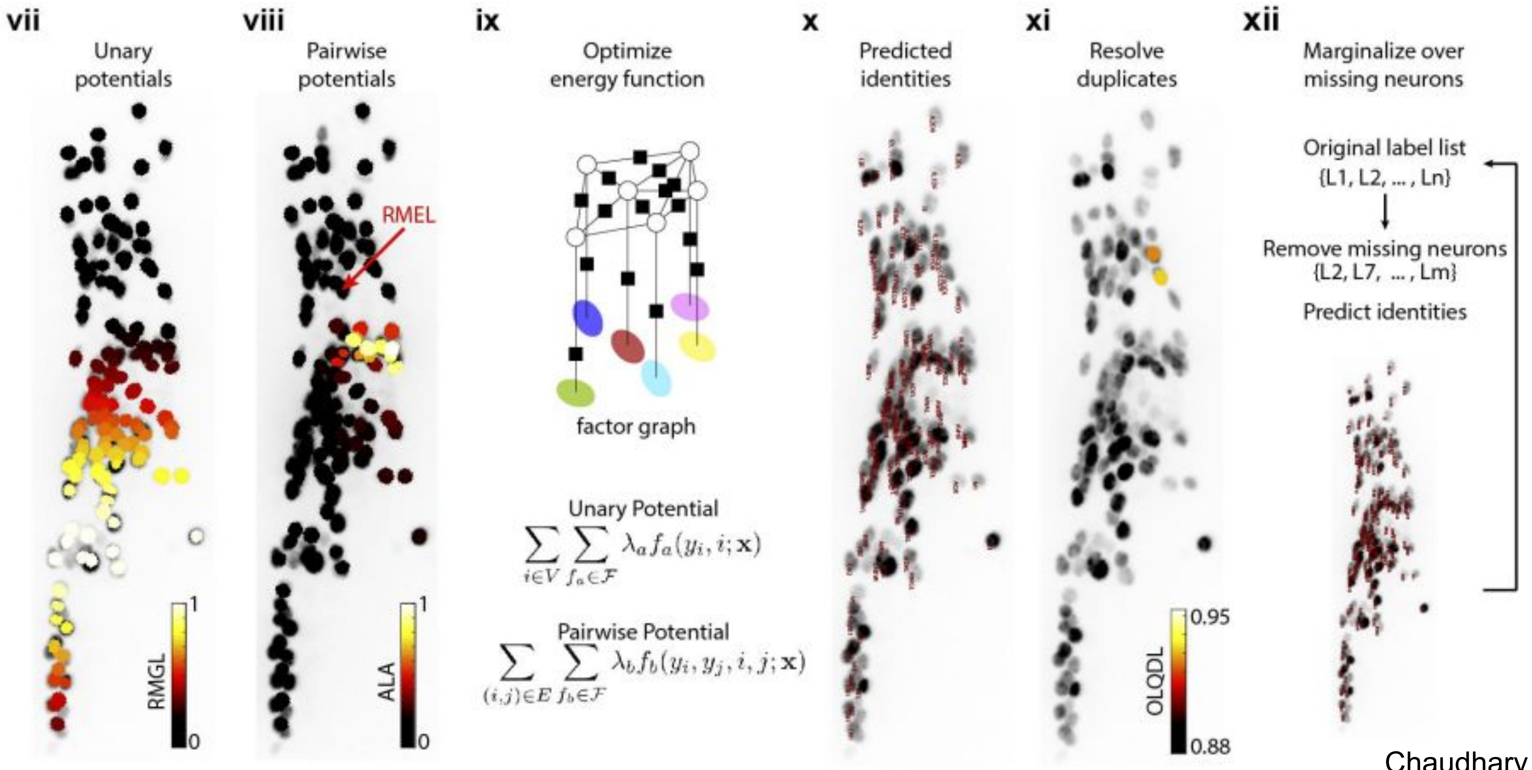
$$P\left(\mathbf{O}_t|\mathbf{S}_t\right)$$

State Dynamics (a.k.a. Motion Model)

$$P\left(\mathbf{S}_t|\mathbf{S}_{t-1}\right)$$

Luo et al, 2017

Conditional Random Fields



Chaudhary et al, 2021

Better use of training data, combine with recent advances in AI

(1) Deep learning based appearance models

- Deep lab cut

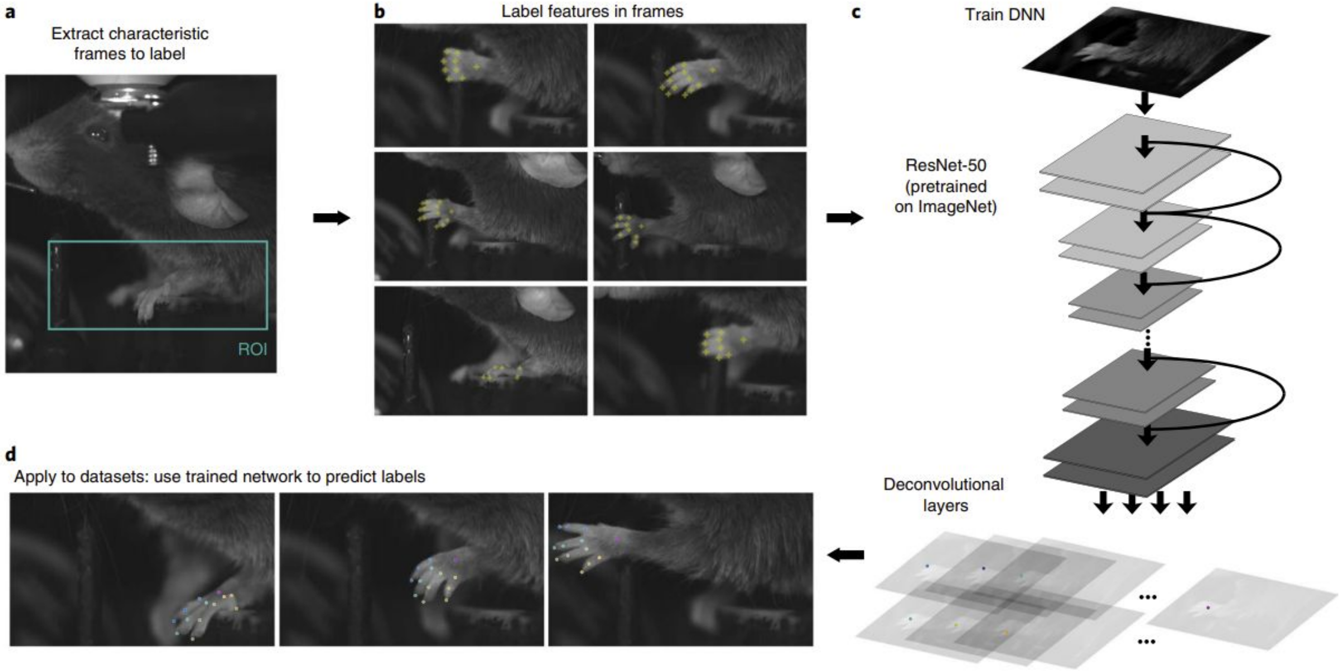
(2) Extensions to probabilistic formulation

- Deep graph pose

(3) Incorporating spatial and temporal structure

- Lightning pose

Deep Lab Cut (DLC)



Benefits

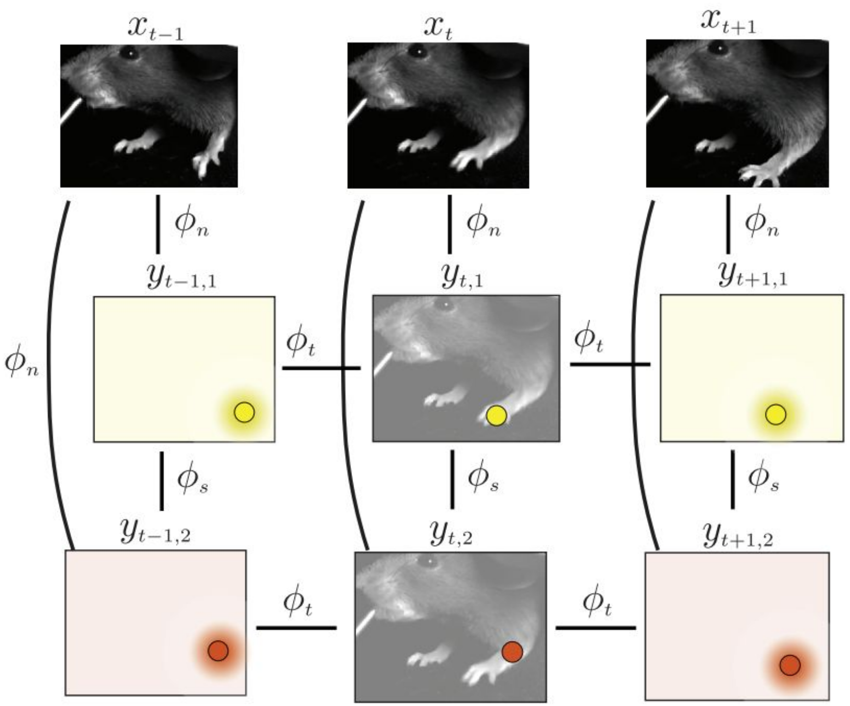
- Fast and scalable

Drawbacks

- Requires labeled data
- Requires fine-tuning or retraining for new datasets
- Does not have an underlying temporal/spatial/motion model

Mathis et al, 2019

From DLC to Deep Graph Pose (DGP)



$$p(y|x,\beta) = \frac{1}{Z(x,\beta)} \exp\Bigg(- \sum_{t=1}^{T} \sum_{j=1}^{J} \phi_n^j(y_{t,j}, x_t) - \sum_{t=1}^{T-1} \sum_{j=1}^{J} \phi_t^j(y_{t,j}, y_{t+1,j}) - \sum_{t=1}^{T} \sum_{i,j \in \mathcal{E}} \phi_s^{ij}(y_{t,i}, y_{t,j}) \Bigg),$$

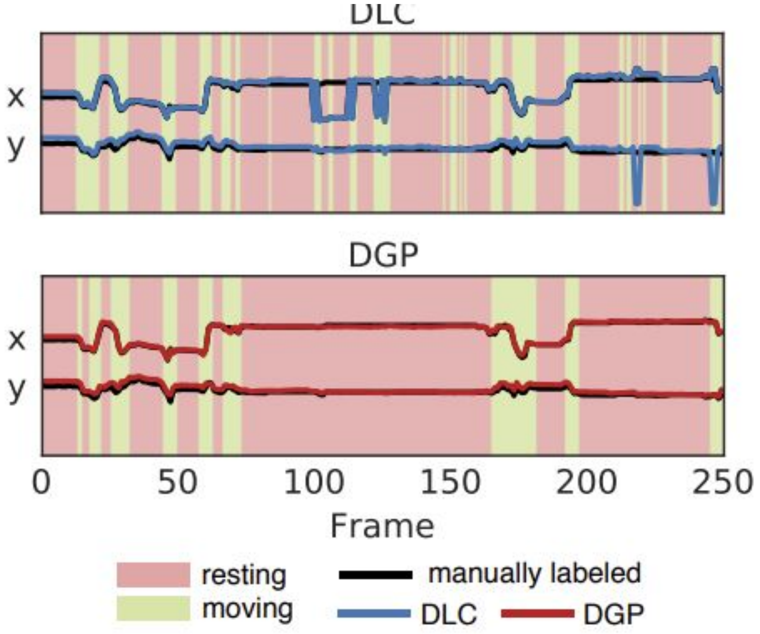
$$\boxed{\begin{aligned} \phi_s^{ij}(y_{t,i}, y_{t,j}) &= \frac{1}{2} w_s^{ij} ||y_{t,i} - y_{t,j}||^2, \\ \phi_t^{j}(y_{t,j}, y_{t+1,j}) &= \frac{1}{2} w_t^{j} ||y_{t,j} - y_{t+1,j}||^2, \end{aligned}}$$

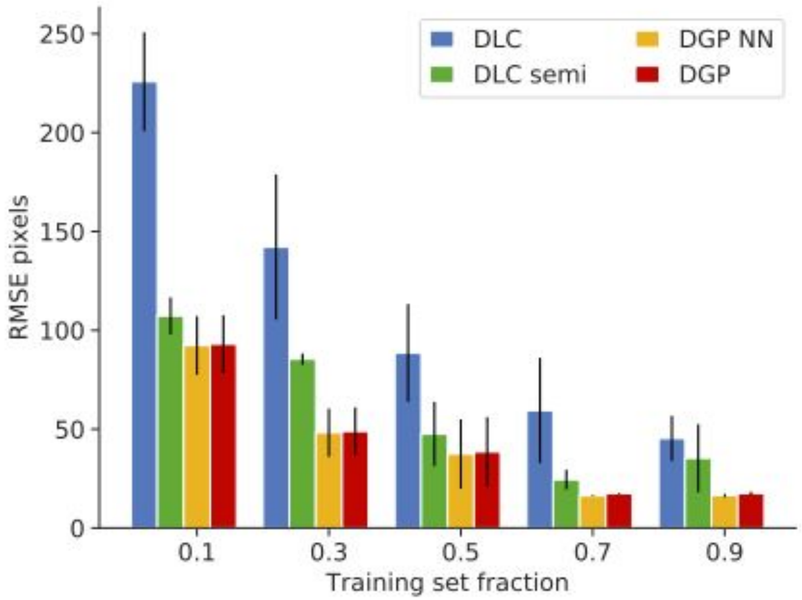
DGP solves major DLC issues

- Uses unlabeled data
- Incorporates temporal smoothness
- Incorporates spatial structure
- Uses probabilistic formulation

Wu et al, 2021

Deep Graph Pose

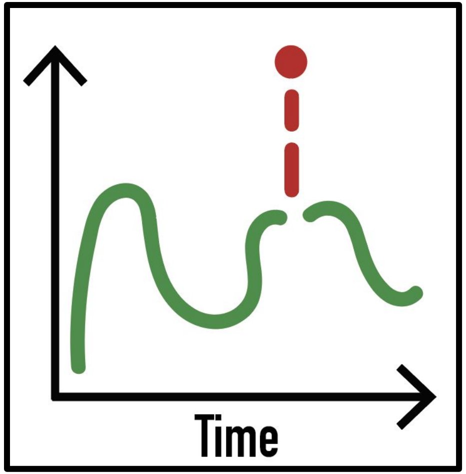




Wu et al, 2021

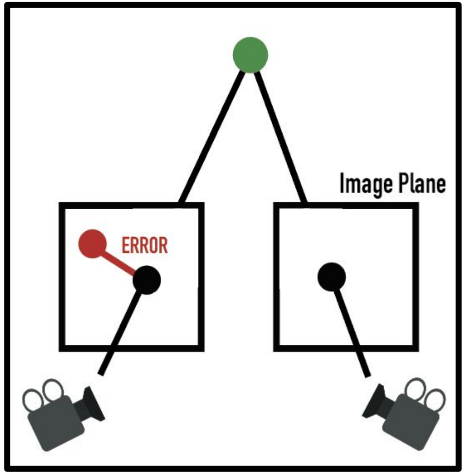
From DGP to Lightning Pose

Temporal smoothness

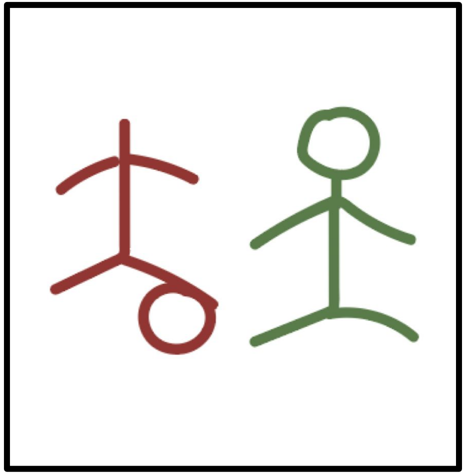


DGP does this too!

Multiview consistency

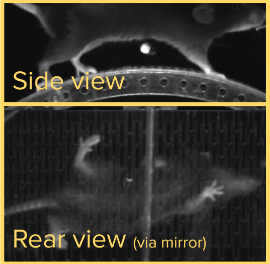


Low dimensionality



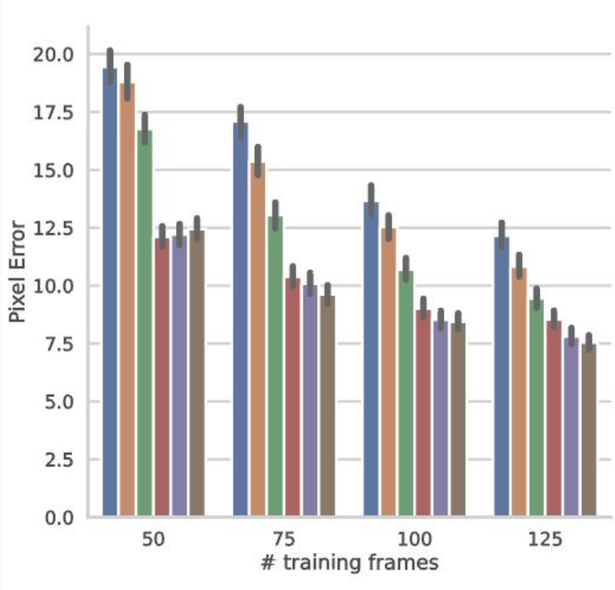
Biderman et al, 2023

Lightning Pose Results

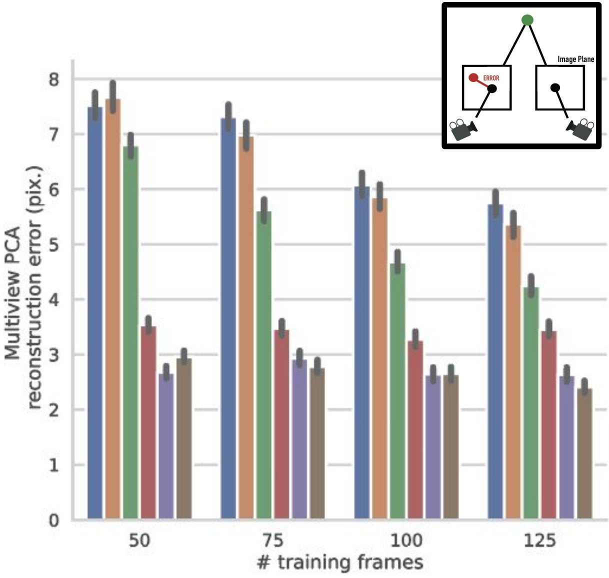


Supervised Unimodal Temporal Dimensionality Multiview Combined

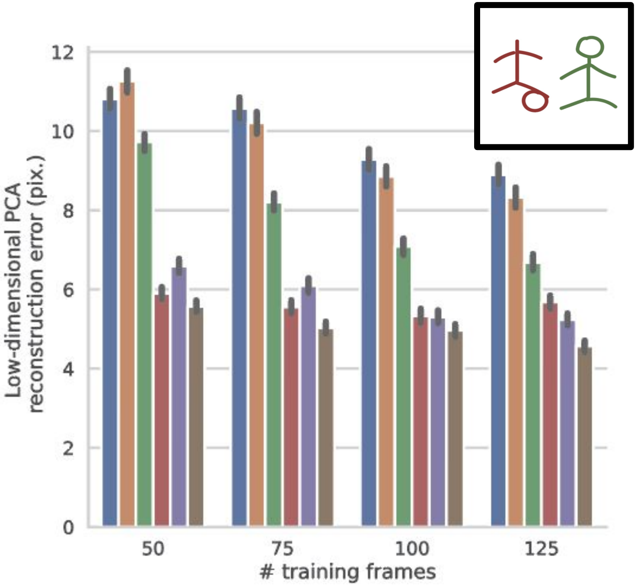
test-set error (preds. VS labels)



Less multiview inconsistency



Less implausible poses



References

1. Luo, Wenhan, et al. "Multiple object tracking: A literature review." Artificial intelligence 293 (2021): 103448.
2. Korsah, G. A., A. T. Stentz, and M. B. Dias. "The dynamic hungarian algorithm for the assignment problem with changing costs." Robotics Institute, Pittsburgh, PA, Tech. Rep. CMU-RI-TR-07-27 (2007).
3. Zhang, Li, Yuan Li, and Ramakant Nevatia. "Global data association for multi-object tracking using network flows." 2008 IEEE conference on computer vision and pattern recognition. IEEE, 2008.
4. Schulter, Samuel, et al. "Deep network flow for multi-object tracking." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.
5. Berclaz, Jerome, et al. "Multiple object tracking using k-shortest paths optimization." IEEE transactions on pattern analysis and machine intelligence 33.9 (2011): 1806-1819.
6. Chaudhary, Shivesh, et al. "Graphical-model framework for automated annotation of cell identities in dense cellular images." Elife 10 (2021): e60321.
7. Mathis, Alexander, et al. "DeepLabCut: markerless pose estimation of user-defined body parts with deep learning." Nature neuroscience 21.9 (2018): 1281-1289.
8. Wu, Anqi, et al. "Deep Graph Pose: a semi-supervised deep graphical model for improved animal pose tracking." Advances in Neural Information Processing Systems 33 (2020): 6040-6052.
9. Biderman, Dan, et al. "Lightning Pose: improved animal pose estimation via semi-supervised learning, Bayesian ensembling, and cloud-native open-source tools." bioRxiv (2023).

Public Datasets:

1. https://motchallenge.net/
2. https://celltrackingchallenge.net/
3. https://www.crcv.ucf.edu/data/