Visual Object Tracking: An overview

Pan He, Ph.D student @ UF MALT Lab
https://bestsonny.github.io/

Tracking of single, arbitrary objects

Problem. Track an arbitrary object with the sole supervision of a single bounding box in the first frame of the video.

Challenges.

- We need to be class-agnostic.
- Stability-Plasticity dilemma[Grossberg87]

"How can a learning system remain plastic in response to significant new events, yet also remain stable in response to irrelevant events?"

Initialization in the 1st frame

.....

Estimated states in the N-th frame

What?

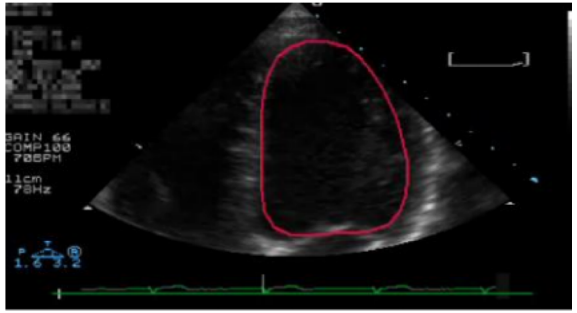
All sorts of “targets”

- Interest points
- Manually selected objects
- Specific known objects
- Cars, faces, people, etc.
- Moving cars, walking people, talking heads

Appearance/dynamical models and inference machineries

- Depend on task and setting
- Heavily influenced by CV/ML trends

With 2D (dynamic) shape prior

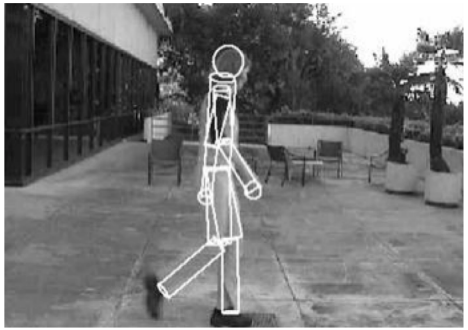


http://www2.imm.dtu.dk/~aam/tracking/

http://vision.ucsd.edu/~kbranson/research/cvpr2005.html

With 3D (cinematic) shape prior

http://cvlab.epfl.ch/research/completed/realtime_tracking/



http://www.cs.brown.edu/~black/3Dtracking.html

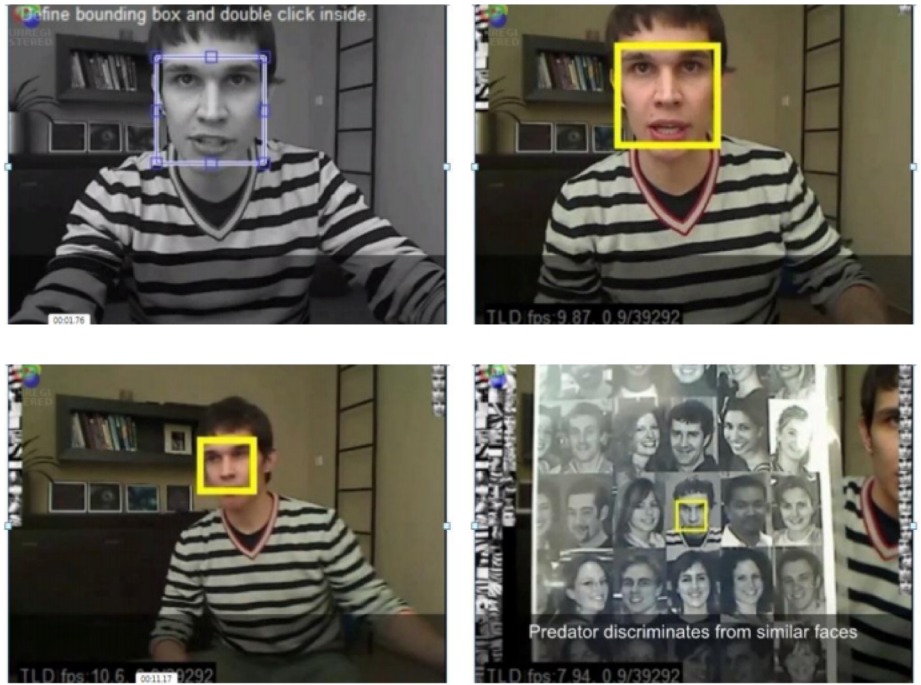
With appearance prior

Detect-before-tracking

http://www.cs.washington.edu/homes/xren/research/cvpr2008_casablanca/

With no appearance prior

Tracking bounding box from user selection



http://info.ee.surrey.ac.uk/Personal/Z.Kalal/

With no appearance prior

Tracking bounding box from user selection (query expansion)

http://www.robots.ox.ac.uk/~vgg/research/vgoogle/

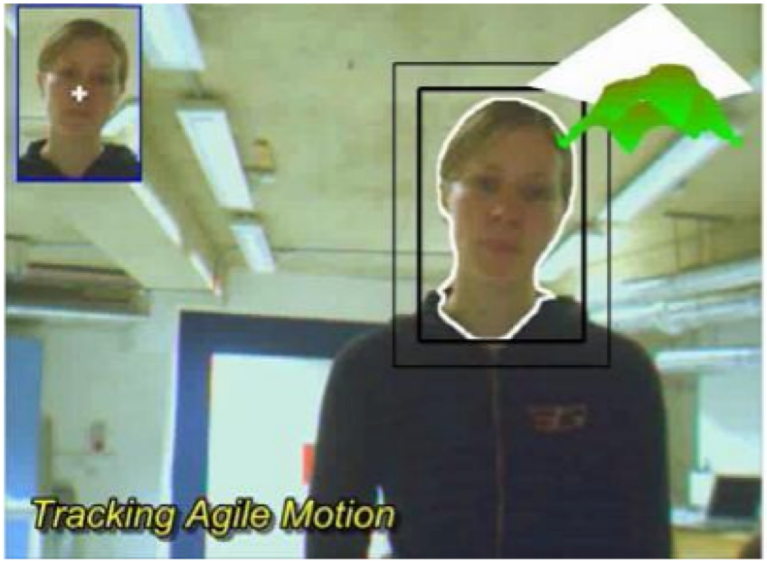
With no appearance prior

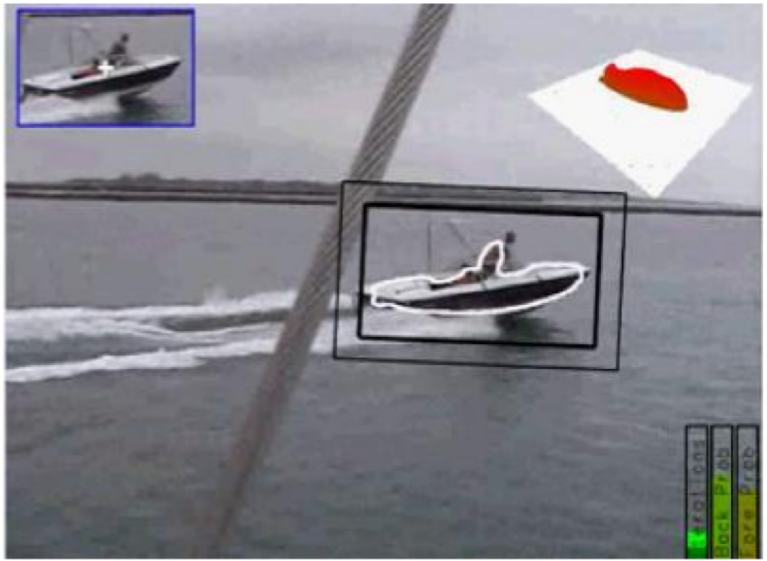
Tracking bounding box from user selection, and using context

http://server.cs.ucf.edu/~vision/projects/sali/CrowdTracking/index.html

With no appearance prior

Tracking bounding box and segmentation from user selection





http://www.robots.ox.ac.uk/~cbibby/index.shtml

Why?

Elementary or principal tool for multiple CV systems

- Other sciences (neuroscience, ethology, biomechanics, sport, medicine, biology, fluid mechanics, meteorology, oceanography)
- Defense, **surveillance**, safety, monitoring, control, assistance
- **Robotics**, Human-Computer Interfaces
- Video content production and post-production (compositing, **augmented reality**, editing, re-purposing, stereo3D authoring, motion capture for animation, clickable hyper videos, etc.
- Video content management (indexing, annotation, search, browsing)

Difficulties In Reliable Object Tracking

UF UNIVERSITY of FLORIDA

More than yet another search/matching/detection problem

- Specific issues
  - Drastic appearance variability through time
  - Non planar, deformable or articulated objects
  - More image quality problems: low resolution, motion blur
  - Speed/memory/causality constraints
- But
  - Sequential image ordering is key
  - Temporal continuity of appearance
  - Temporal continuity of object state

Formalizing tracking

Elementary or principal tool for multiple CV systems

- Other sciences (neuroscience, ethology, biomechanics, sport, medicine, biology, fluid mechanics, meteorology, oceanography)
- Defense, **surveillance**, safety, monitoring, control, assistance
- **Robotics**, Human-Computer Interfaces
- Video content production and post-production (compositing, **augmented reality**, editing, re-purposing, stereo3D authoring, motion capture for animation, clickable hyper videos, etc.
- Video content management (indexing, annotation, search, browsing)

Formalizing tracking

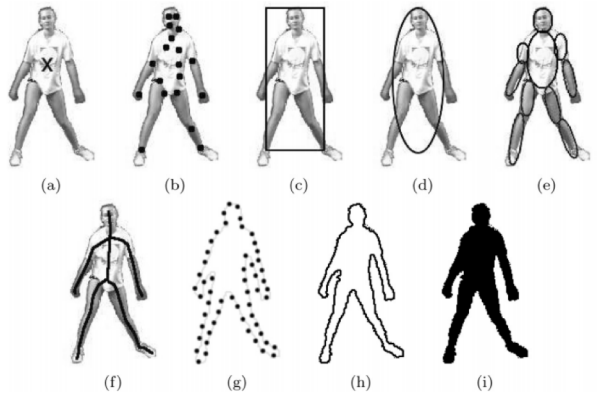
Tracking: Given past and current measurements → Output an estimate of current hidden state

Image-based “measurements”:

- Raw or filtered images (intensities, colors, texture)
- Low-level features (edges, corners, blobs, optical flow)
- High-level features (e.g., deep learning features)

Single target “state”

- Bounding box parameters (up to 6 DoF)
- 3D rigid pose (6 DoF)
- 2D/3D articulated pose (up to 30 DoF)
- 2D/3D principal deformations
- Discrete pixel-wise labels (segmentation)
- Discrete indices (activity, visibility, expression)



(a) Centroid, (b) multiple points, (c) rectangular patch, (d) elliptical patch, (e) part-based multiple patches, (f) object skeleton, (g) complete object contour, (h) control points on object contour, (i) object silhouette.

Tracking as Ridge Regression

The goal of training is to find a function

$$f(z) = w^T z$$

That minimizes the squared error over samples x_i and their regression targets y_i

$$\min_{\mathbf{w}} \sum_i \left(f(\mathbf{x}_i) - y_i\right)^2 + \lambda \left\|\mathbf{w}\right\|^2 .$$

According to [1], the solution is:

$$\mathbf{w} = \left(X^T X + \lambda I\right)^{-1} X^T \mathbf{y}.$$

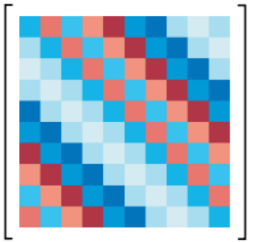
In general, a large system of linear equations must be solved to compute the solution, which can become prohibitive in a real-time setting

[1] R. Rifkin, G. Yeo, and T. Poggio, "Regularized least-squares classification," Nato Science Series Sub Series III Computer and Systems Sciences, vol. 190, pp. 131–154, 200

Cyclic shifts

+30 +15 Base sample -15 -30

$C\left(\quad\right) =$



Base sample
Shifted by 1 element
Shifted by 2 elements
⋮
Shifted by $n-1$ elements

cyclic shift operator

$$P = \begin{bmatrix} 0 & 0 & 0 & \cdots & 1 \\ 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}.$$

Due to the cyclic property, we get the same signal x periodically every n shifts. This means that the full set of shifted signals is obtained with

$$\{P^u \mathbf{x} \,|\, u = 0, \ldots n-1\}.$$

[1] R. Rifkin, G. Yeo, and T. Poggio, "Regularized least-squares classification," Nato Science Series Sub Series III Computer and Systems Sciences, vol. 190, pp. 131–154, 200

Cyclic shifts

To compute a regression with shifted samples, we can use them as the rows of a data matrix X:

$$X = C(\mathbf{x}) = \begin{bmatrix} x_1 & x_2 & x_3 & \cdots & x_n \\ x_n & x_1 & x_2 & \cdots & x_{n-1} \\ x_{n-1} & x_n & x_1 & \cdots & x_{n-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_2 & x_3 & x_4 & \cdots & x_1 \end{bmatrix}.$$

Correlation Filter

Given the template path $\varphi(x) \in \mathbb{R}^{M\times N\times D}$ and the idea response $y \in \mathbb{R}^{M\times N}$, the desired filter w can be obtained by minimizing the output ridge loss:

$$\epsilon = \left\| \sum_{l=1}^{D} \mathbf{w}^l \star \varphi^l(\mathbf{x}) - \mathbf{y} \right\|^2 + \lambda \sum_{l=1}^{D} \left\| \mathbf{w}^l \right\|^2$$

The solution can be gained as:

$$\hat{\mathbf{w}}^l = \frac{\hat{\varphi}^l(\mathbf{x}) \odot \hat{\mathbf{y}}^*}{\sum_{k=1}^{D} \hat{\varphi}^l(\mathbf{x}) \odot (\hat{\varphi}^l(\mathbf{x}))^* + \lambda}$$

Correlation Filter

For the detection process, we crop a search patch and obtain the features ϕ(z) in the new frame, the translation can be estimated by searching the maximum value of correlation response map g

$$\mathbf{g} = \mathcal{F}^{-1}\left(\sum_{l=1}^{D} \hat{\mathbf{w}}^{l*} \odot \hat{\varphi}^{l}(\mathbf{z})\right)$$

Correlation Filter

During the online tracking, we just update the filters w over time. The optimization problem can be formulated in a incremental mode:

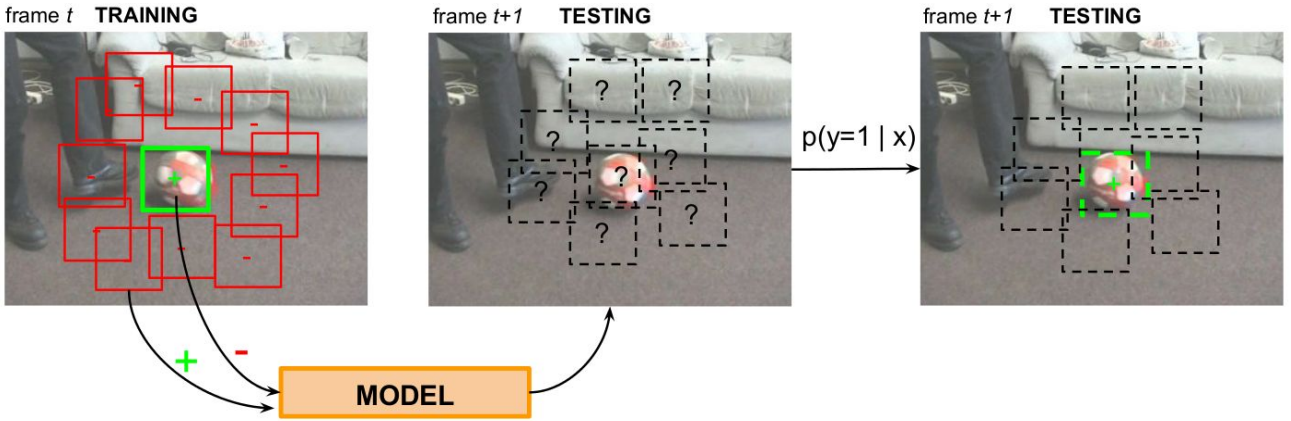
$$\epsilon = \sum_{t=1}^{p} \beta_t \left(\left\| \sum_{l=1}^{D} \mathbf{w}_p^l \star \varphi^l(\mathbf{x}_t) - \mathbf{y} \right\|^2 + \lambda \sum_{l=1}^{D} \left\| \mathbf{w}_p^l \right\|^2 \right)$$

The solution now can be extend to time series:

$$\hat{\mathbf{w}}_p^l = \frac{\sum_{t=1}^{p} \beta_t \hat{\mathbf{y}}^* \odot \hat{\varphi}^l(\mathbf{x}_t)}{\sum_{t=1}^{p} \beta_t \left(\sum_{k=1}^{D} \hat{\varphi}^k(\mathbf{x}_t) \odot (\hat{\varphi}^k(\mathbf{x}_t))^* + \lambda \right)}$$

Recent history of object tracking [2010 - today]

Tracking-by-detection paradigm



- Learn online a binary classifier (+ is object, - is background).
- Re-detect the object at every frame + update the classifier.

Slides adapted from Luca et. al. @Valse 2016

Recent history of object tracking [2010 - today]

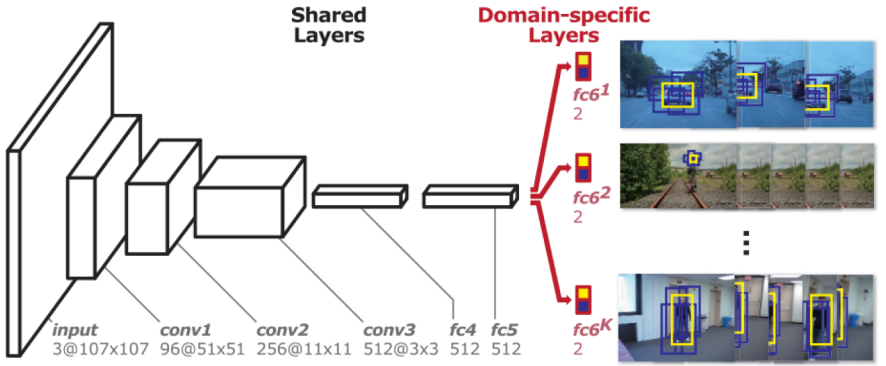
Correlation filters become the most popular choice

- Sampling space is loosely a circulant matrix → diagonalized with Discrete Fourier Transform.
- Fast training and evaluation of linear classifier in the Fourier Domain.
- Mostly used with HOG features.

Slides adapted from Luca et. al. @Valse 2016

MDNet [CVPR16, winner of VOT15]

- Rationale: separate domain-independent (e.g. the concept of “objectness”) to domain-dependent (video-specific) information.
- Training. fixed common part (3conv+2fc) and several “one-hot” fc branches.
- Tracking. fine-tuning of several layers, hard-negative mining, bbox regression.

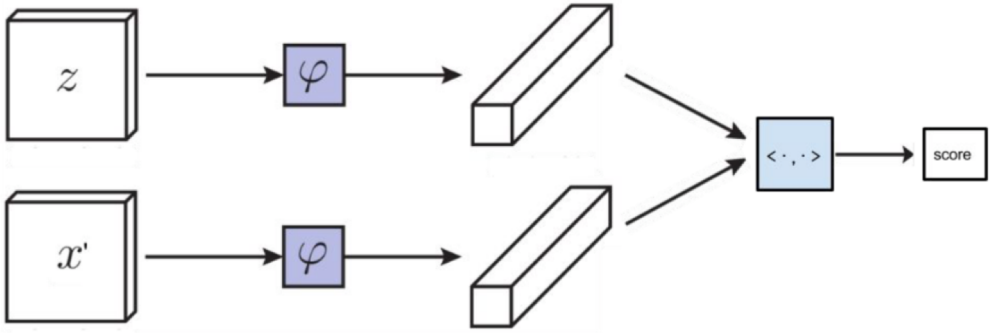


1 fps

Slides adapted from Luca et. al. @Valse 2016

Vanilla siamese conv-net for similarity learning

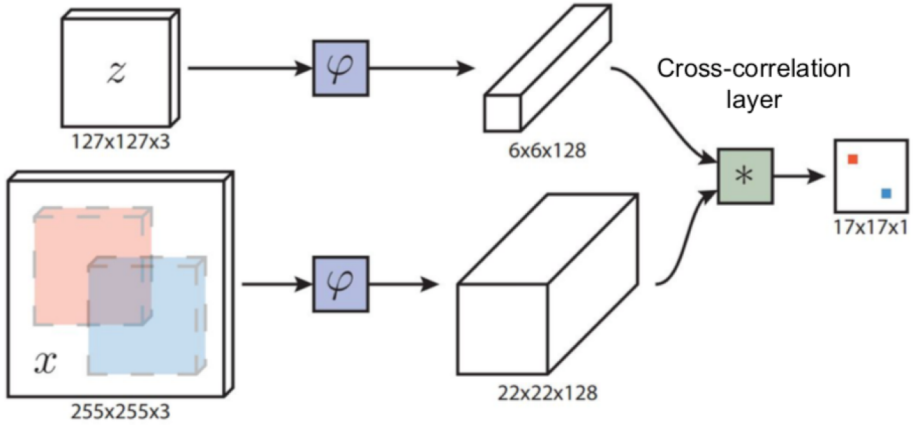
- Siamese conv-net trained to address a similarity learning problem in an offline phase.
- The conv-net learns a function that compares an exemplar z to a candidate of the same size x'.
- Score tell us how similar are the two image patches.



Slides adapted from Luca et. al. @Valse 2016

Fully-Convolutional Siamese Networks for Object Tracking (SiamFC CVPR17)

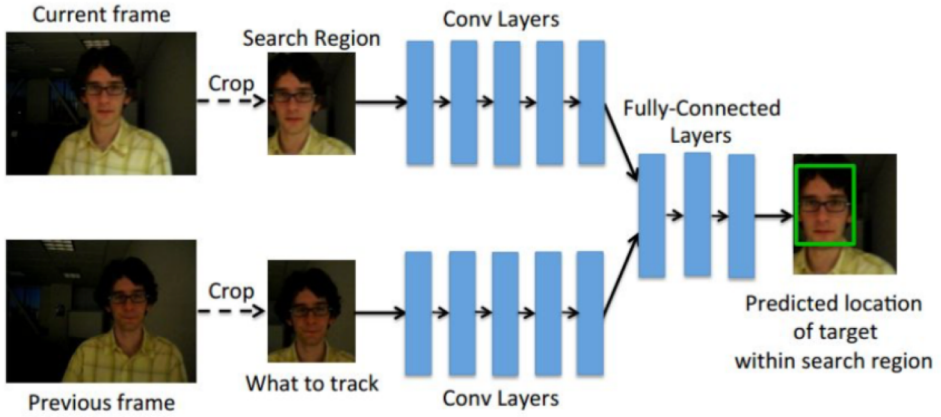
- One fully convolutional network (no padding, no fc).
- Two inputs of different sizes: smaller is the exemplar (target object during tracking), bigger is the search area.
- Output of embedding function has spatial support.
- Cross-correlation layer: computes the similarity at all translated sub-windows on a dense grid in a single evaluation.
- ● Output is a score map.



Forward pass: **>100Hz**

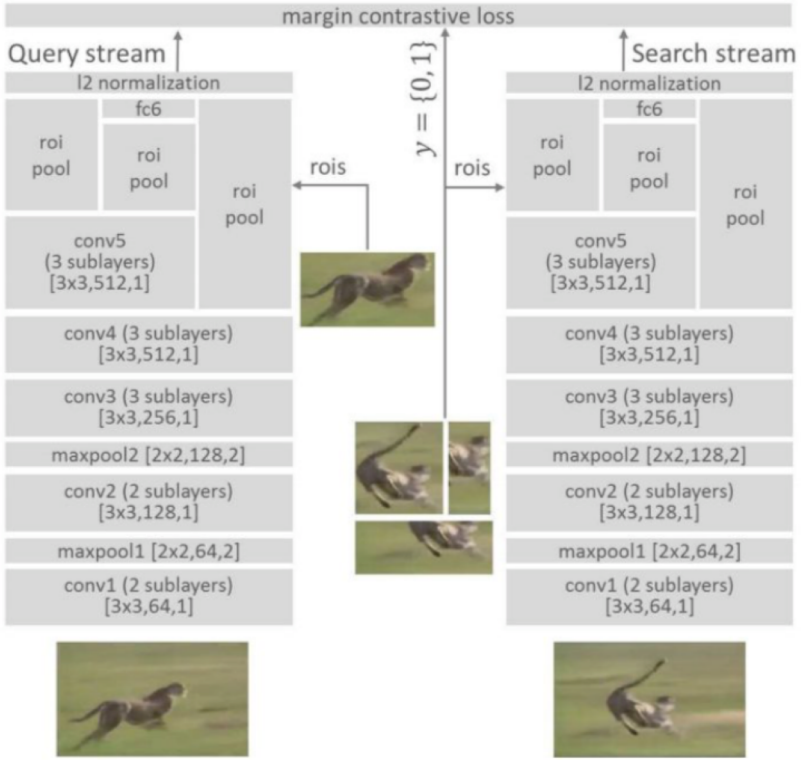
GOTURN [ECCV16]

- Siamese architecture trained to solve Bounding Box regression problems.
- Network is not fully convolutional.



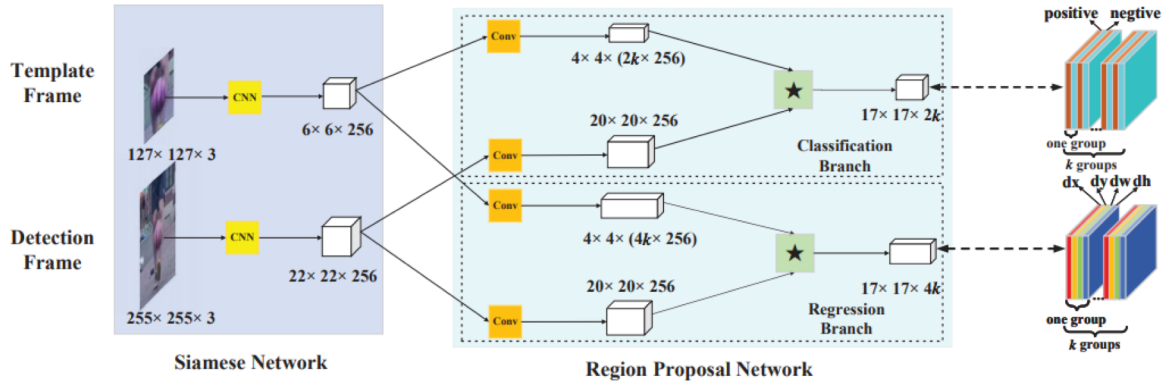
SINT [CVPR16]

- Siamese architecture trained to learn a generic similarity function.
- ROI pooling to sample candidates.
- BBox regression to improve tracking performance.



SiamRPN [CVPR18]

- Siamese subnetwork for feature extraction
- Region proposal subnetwork including the classification branch and regression branch.
- State-of-the-art method



Current trends

Leverage cutting-edge ML/DL tools

- Sparse appearance modeling
- Discriminative learning
- Adversarial learning

Exploitation of context

- Sparse appearance modeling
- Leveraging scene understanding
  - Geometry
  - Pixel-wise semantics
  - Interaction between scene elements

OpenSource Framework

https://github.com/huanglianghua/open-vot

UF UNIVERSITY of FLORIDA

Evaluation Methodology

We use the **precision and success rate** for quantitative analysis. In addition, we evaluate the **robustness** of tracking algorithms in two aspects:

- Precision plot
  - Center location error
- Success plot
  - Bounding box overlap
- Robustness Evaluation
  - One-pass evaluation (OPE)
  - Temporal robustness evaluation (TRE)
  - Spatial robustness evaluation (SRE)

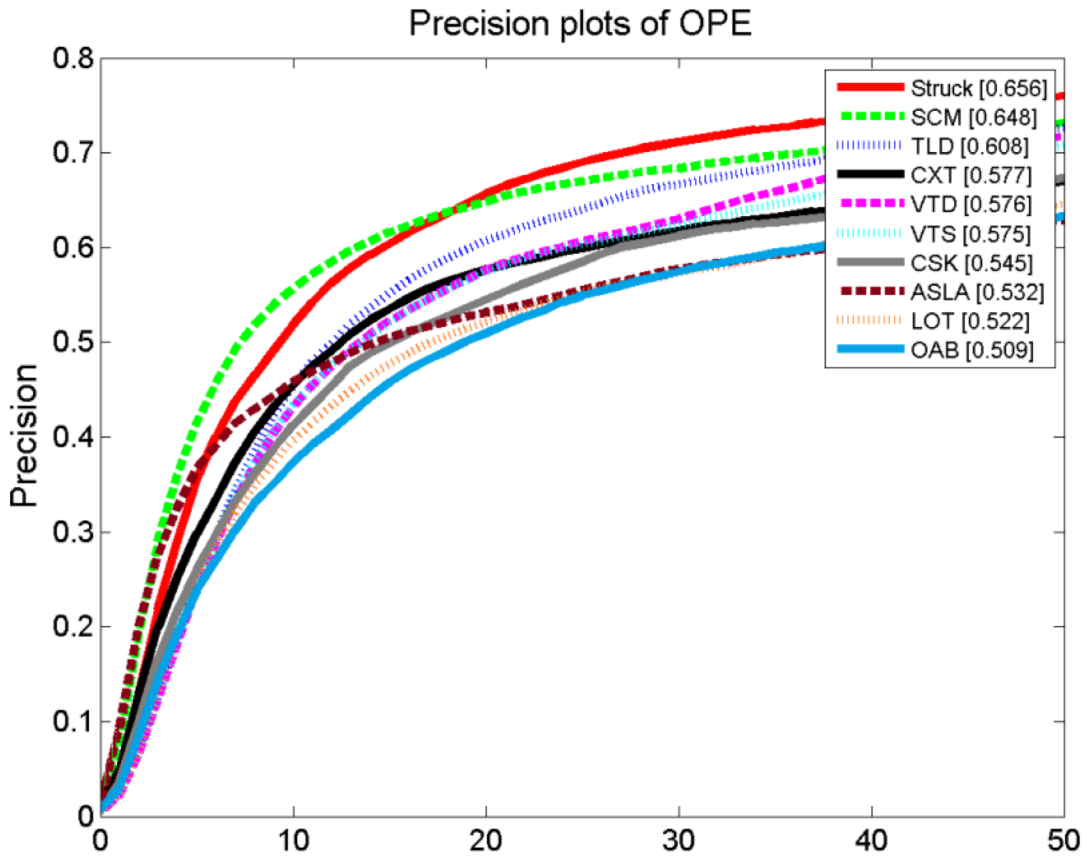
Evaluation Methodology

Center location error is defined as the average Euclidean distance between the center locations of the tracked targets and the manually labeled ground truths

The average center location error over all the frames of one sequence is used to summarize the overall performance for that sequence.

The **precision plot** has been adopted to measure the overall tracking performance. It shows the **percentage** of frames whose estimated location is within the given threshold distance of the ground truth.

Evaluation Methodology



Evaluation Methodology

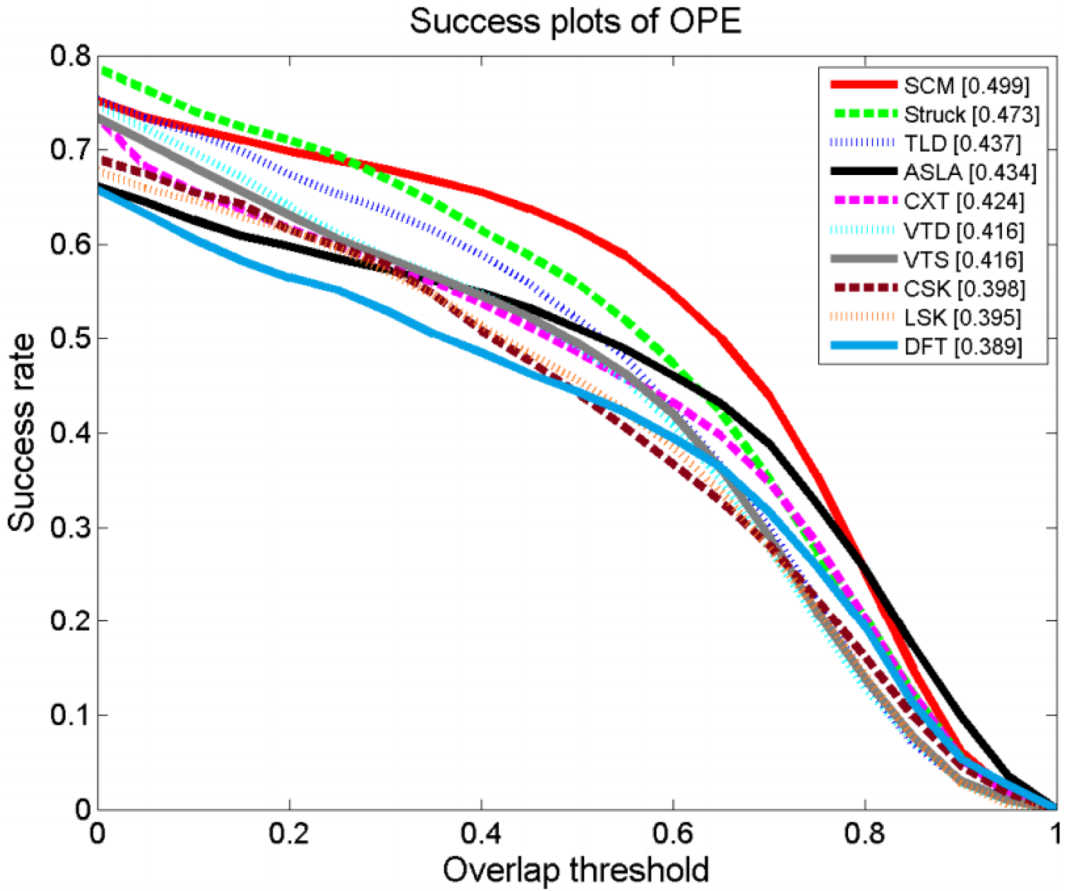
Bounding box overlap. Given the tracked bounding box r_t and the ground truth bounding box r_a, the overlap score is defined as

$$S = \frac{|\, r_t \bigcap r_a \,|}{|\, r_t \bigcup r_a \,|}$$

where ∩ and ∪ represent the **intersection** and **union** of two regions, respectively, and | · | denotes the number of pixels in the region. To measure the performance on a sequence of frames, we count the number of successful frames whose overlap S is larger than the given threshold t_o

The **success plot** shows the ratios of successful frames at the thresholds varied from 0 to 1. Use the **area under curve** (AUC) of each success plot to rank the tracking algorithms

Evaluation Methodology



Evaluation Methodology

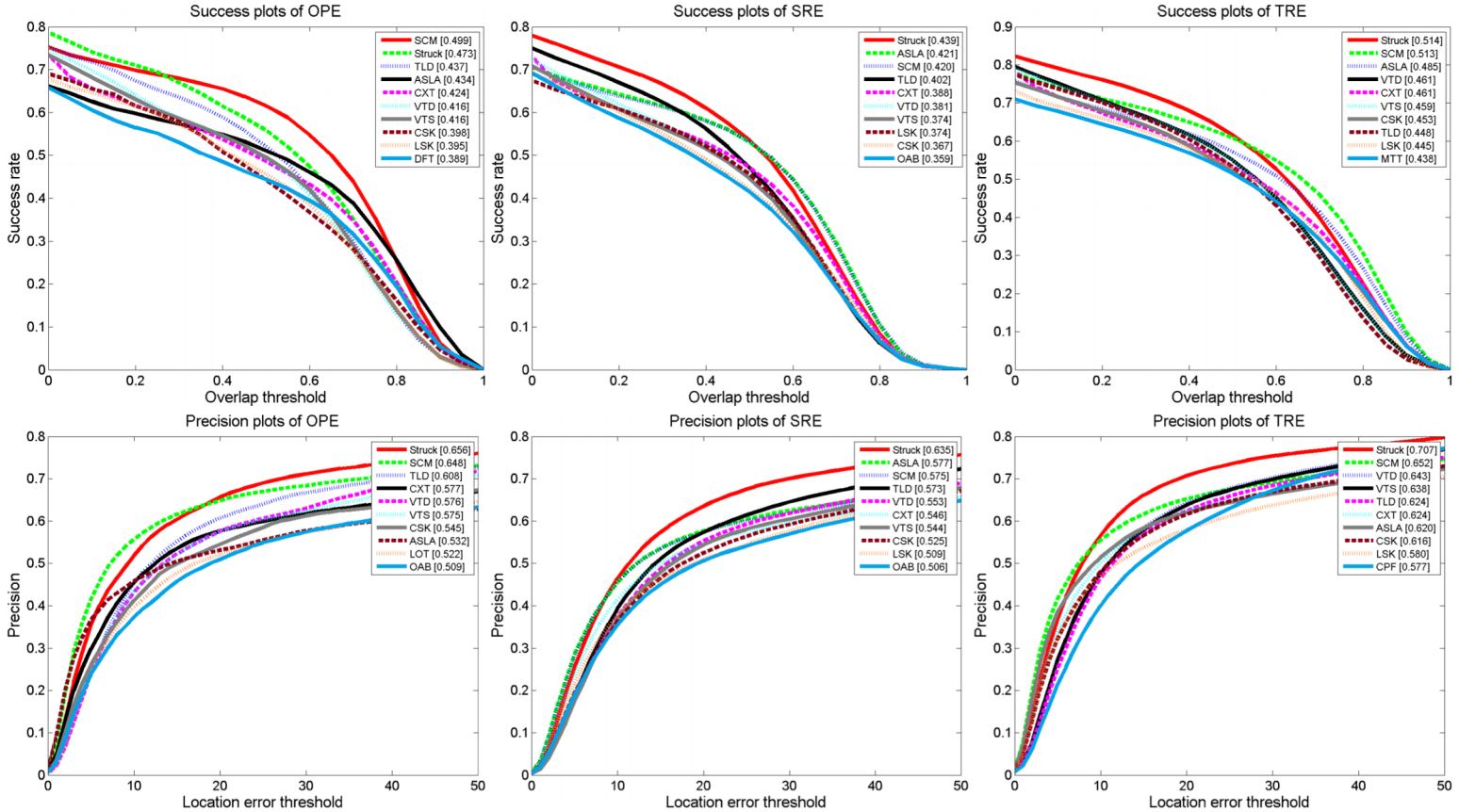
One-pass evaluation. To run them throughout a test sequence with initialization from the ground truth position in the first frame and report the average precision or success rate.

However a tracker may be **sensitive** to the **initialization**, and its performance with different initialization at a different start frame may become much worse or better

Evaluation Methodology

Two better ways to analyze a tracker's robustness to **initialization**, by perturbing the initialization **temporally** (i.e., start at different frames) and **spatially** (i.e., start by different bounding boxes), which are referred as **temporal robustness evaluation (TRE)** and **spatial robustness evaluation (SRE)** respectively

Evaluation Methodology



Visual Tracker Benchmarks

Several popular benchmarks

- Object Tracking Benchmark(OTB)
- Visual Object Tracking (VOT) challenge
- Need for Speed Dataset (NFS)

Reviews, tutorials

Computer vision: a modern approach, Chapter 19, Forsyth and Ponce

Object tracking: a survey, Yilmaz et al. 2006
http://vision.eecs.ucf.edu/papers/Object%20Tracking.pdf

A review of visual tracking, Cannons, 2008
http://www.cse.yorku.ca/techreports/2008/CSE-2008-07.pdf

Recent advances and trends in visual tracking: A review, Yang et al., 2011
http://210.75.252.83/bitstream/344010/6218/1/110201.pdf

Lucas-Kanade 20 years on: a unifying framework, Barker and Matthews, 2004
http://www.cs.cmu.edu/afs/cs/academic/class/15385-s12/www/lec_slides/Baker&Matthews.pdf

A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking, MS Arulampalam et al., 2002
http://www.dis.uniroma1.it/~visiope/Articoli/ParticleFilterTutorial.pdf

On sequential Monte Carlo sampling methods for Bayesian filtering, Doucet et al. 2000
http://www-sigproc.eng.cam.ac.uk/~sjg/papers/99/statcomp_final.ps

Thank you.