VOT
visual object tracking

The Visual Object Tracking VOT2015: Challenge and results

Matej Kristan, Aleš Leonardis, Jiri Matas, Michael Felsberg
Luka Čehovin, Gustavo Fernandez, Tomaš Vojir, Georg Nebehay,
Roman Pflugfelder, et al.

University *of Ljubljana*
Faculty *of Computer and Information Science*

UNIVERSITY OF BIRMINGHAM

LINKÖPING UNIVERSITY

AUSTRIAN INSTITUTE OF TECHNOLOGY

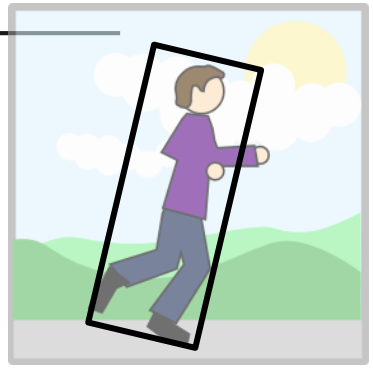
Outline

1. Scope of the VOT challenge
2. VOT2015 challenge overview
   - Evaluation system
   - Dataset
   - Performance evaluation measures
3. VOT2015 results overview
4. Summary and outlook

VOT2015

SCOPE OF THE VOT2015 CHALLENGE

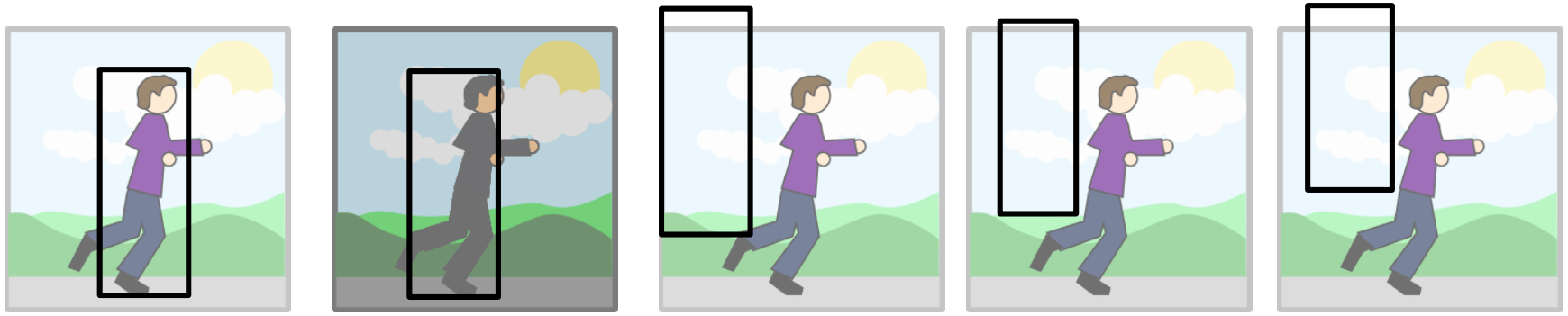
Selected class of trackers



- *Single-object, single-camera, model-free, short-term, causal trackers*
- Model-free:
  - Nothing but a single training example is provided by the BBox in the first frame
- Short-term:
  - Tracker does not perform re-detection
  - Once it drifts off the target we consider that a failure
- Causality:
  - Tracker does not use any future frames for pose estimation
- Object state defined as a rotated bounding box (rectangle)

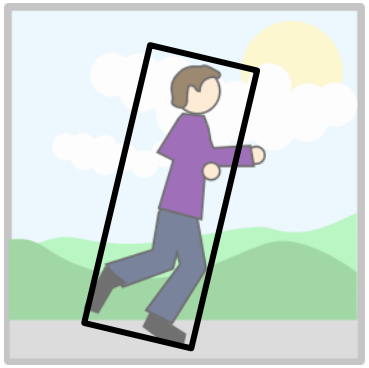
Requirements for tracker implementation

- VOT approach: Use the data fully



- Renitialize once the tracker drifts from the object

Requirements for tracker implementation



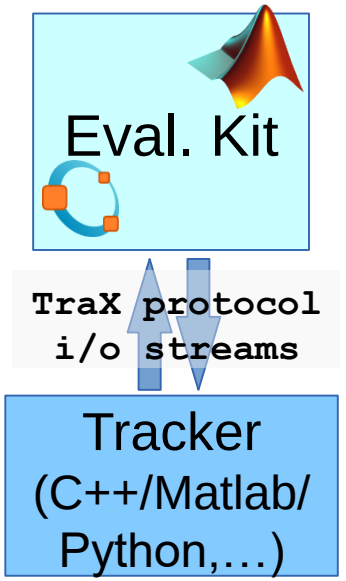
- Complete reset:
  - Tracker is not allowed to use any information obtained before reset, e.g., learnt dynamics, visual model.
- Trackers required to predict a single BB per frame
- Parameters may be set internally, but not by detecting a specific sequence
  - Verified for the top-performing trackers

VOT2015

VOT2015 EVALUATION SYSTEM

VOT2015 Challenge evaluation kit

- Matlab-based kit to automatically perform a battery of standard experiments
- Plug and play!
  - Supports multiple platforms and programming languages (C/C++/Matlab/Python, etc.)



- Easy to evaluate your tracker on all our benchmarks
- Backward compatibility with VOT2013/VOT2014

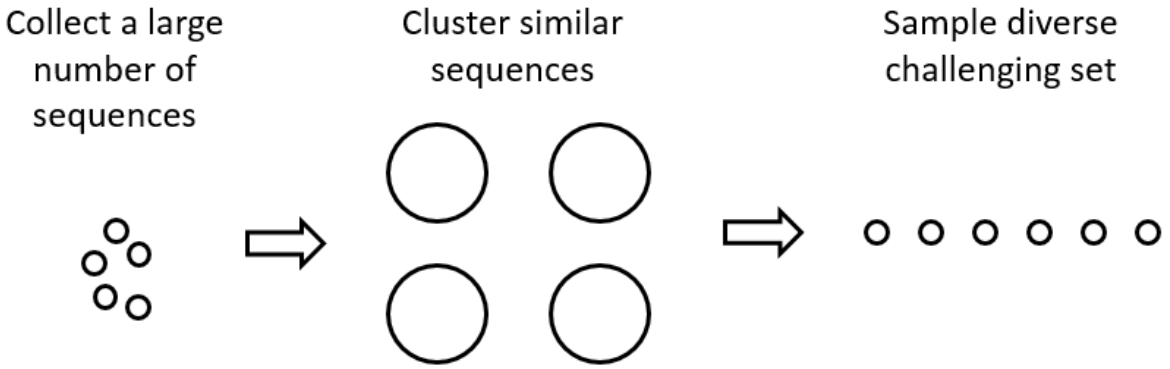
- Download from our homepage
  https://github.com/vicoslab/vot-toolkit

VOT2015

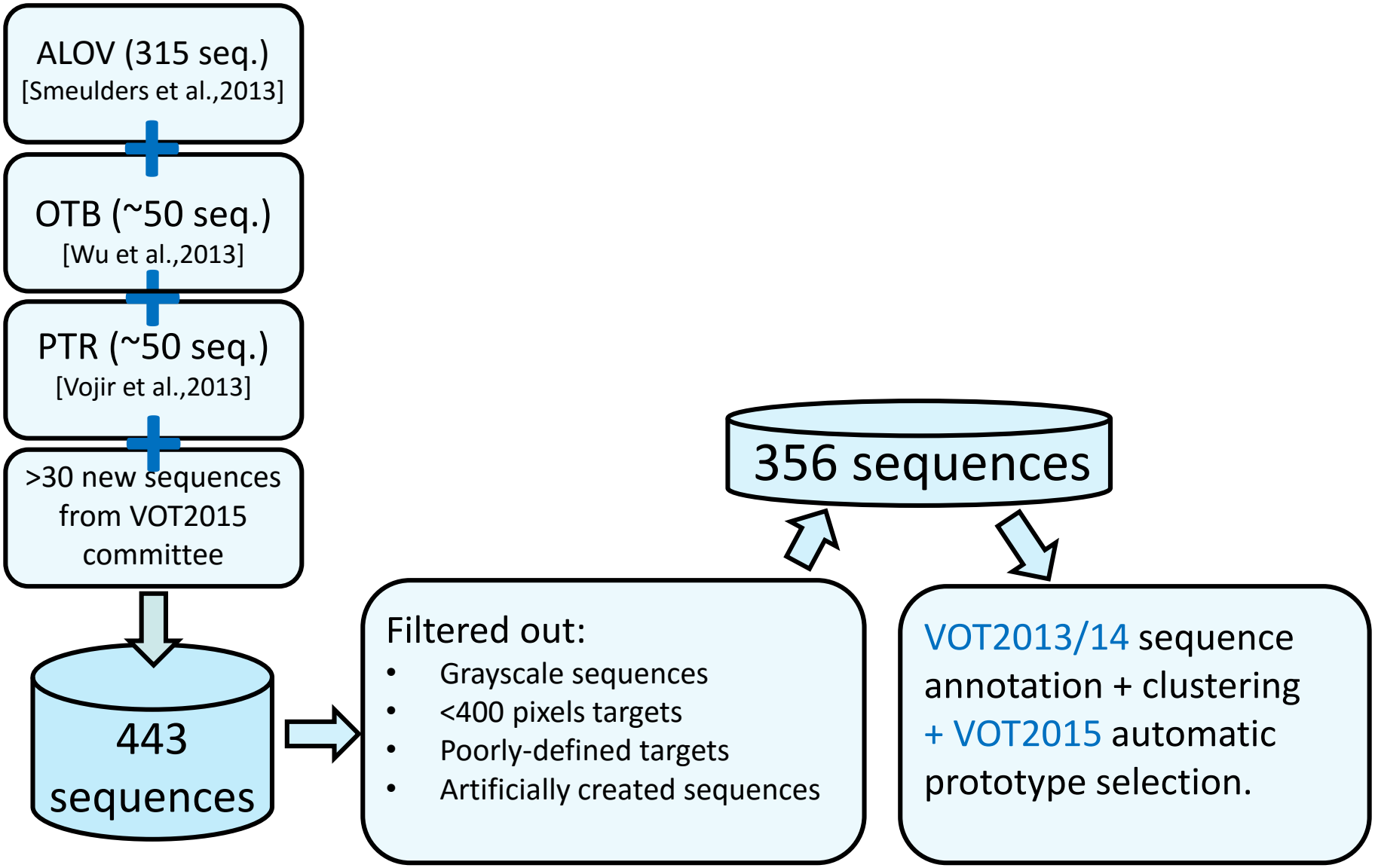
VOT2015 DATASET

Dataset construction approach

- Current trend [Wu et al. CVPR2013, Smeulders et al. PAMI2013, Wang et al. arXiv2015, Wu et al. PAMI2015]:
  - Large datasets by collecting many sequences from internet
  - Large dataset ≠ diverse or useful
- VOT2013/2014 approach:
  - Keep it sufficiently small, well annotated and diverse
  - Developed the VOT dataset construction methodology

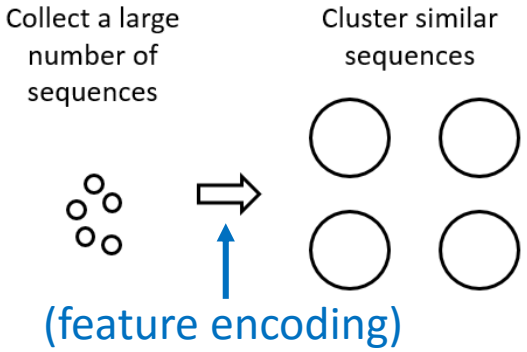


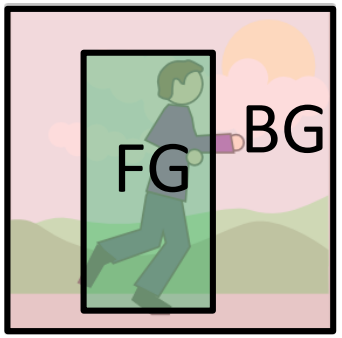
VOT2015 dataset: collection and filtering



VOT2015 dataset: Clustering 1/2

- 11 global attributes estimated automatically for 356 sequences
  - Each sequence represented as 11dim feature vector.





Global attributes:

1. Illumination change
   (difference of min/max FG intensity)
2. Size change
   (average of sequential BB size difference)
3. Motion
   (average of sequential BB center difference)
4. Clutter
   (FG/BG color histogram difference)
5. Camera motion
   (patch features motion between frames)
6. Blur
   (Camera focus measure [Kristan et al., 2006])
7. Aspect-ratio change
   (relative to initial BB aspect ratio change)
8. Object color change
   (average hue change inside BB w.r.t initial)
9. Deformation
   (mean intensity change in BB subregions)
10. Scene complexity
    (entropy of grayscale image)
11. Absolute motion
    (median difference between first and current BB center)

VOT2015 dataset: Clustering 2/2

- Sequences clustered by Affinity Propagation [Frey and Dueck 2007]
  - Automatic selection of the number of clusters (K=28)

Collect a large number of sequences

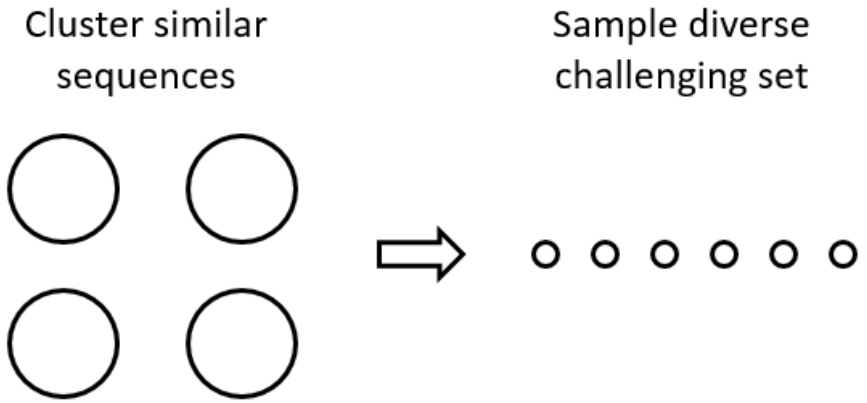
Feature encoding

11 dim

Cluster similar sequences by Affinity Propagation

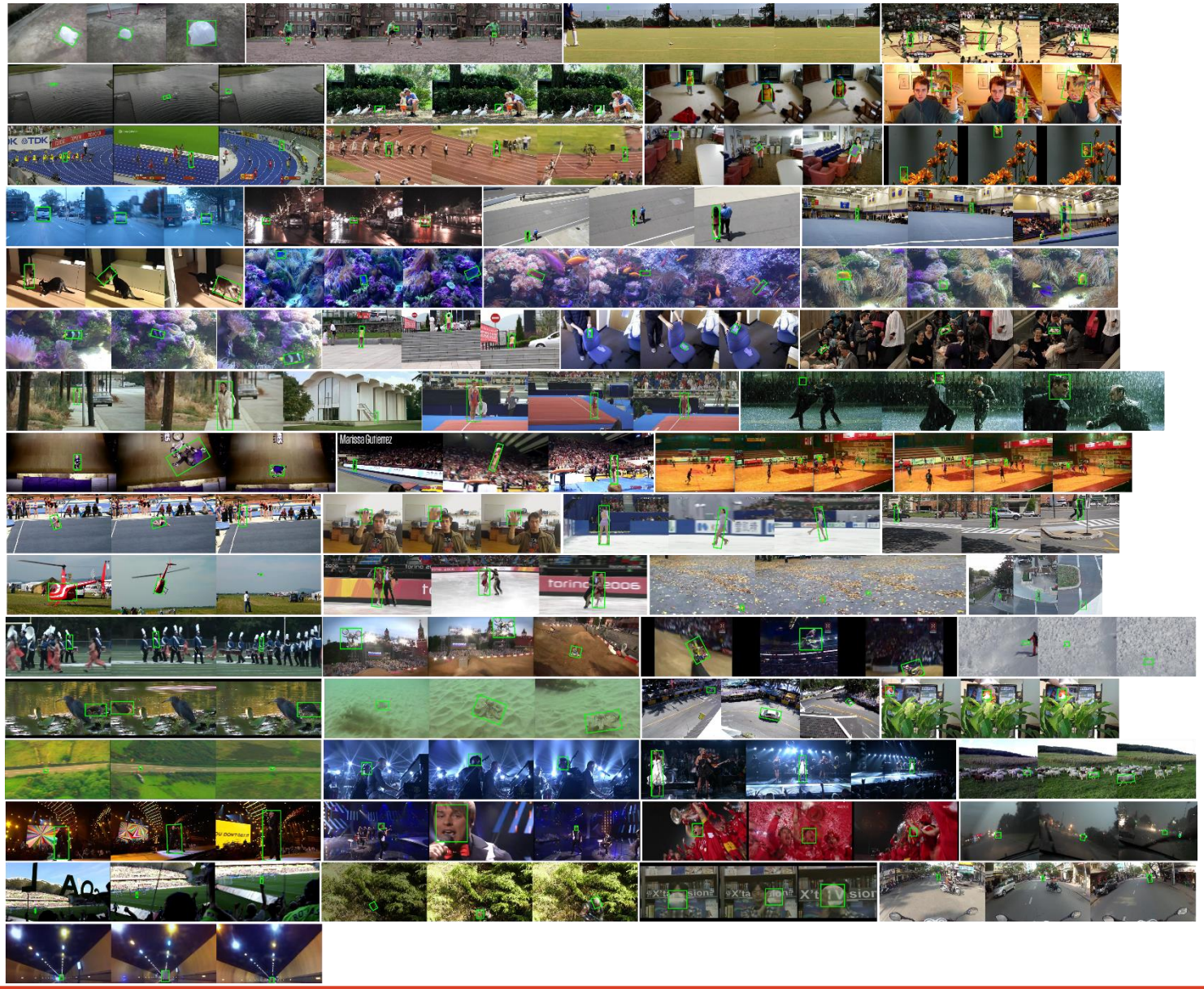
VOT2015 dataset: Cluster sampling

- Requirement:
  - Diverse visual attributes
  - Challenging subset



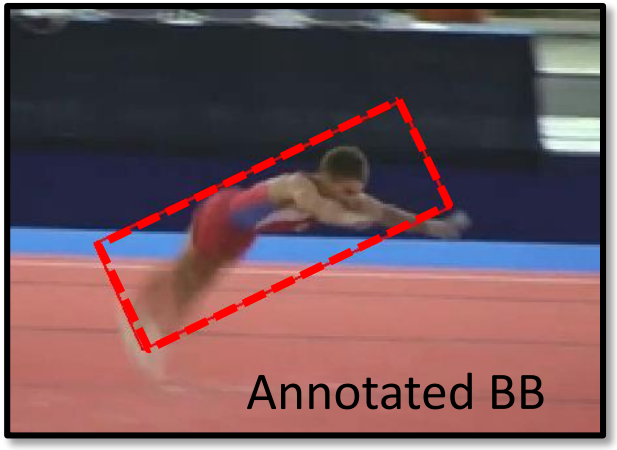
- Global visual attributes: computed
- Tracking difficulty attribute: Applied FoT, ASMS, KCF trackers
- Developed a sampling strategy that sampled challenging sequences while keeping the global attributes diverse.

VOT2015 dataset: 60 sequences



VOT2015 dataset – object annotation

- All sequences re-annotated by (rotated) bounding boxes



- Annotation guidelines distributed among annotators
- Each annotation cross-checked by two annotators
- Approximately square rotated BBs changed to axis-aligned.

VOT2015 dataset – frame annotation

- Manually and automatically labeled each frame with VOT2013 visual attributes:
  - i. Occlusion (M)
  - ii. Illumination change (M)
  - iii. Object motion (A)
  - iv. Object size change (A)
  - v. Camera motion (M)
  - vi. Unassigned (A)

M ... manual annotation, A ... automatic annotation

(i)	0	1	1	0
(ii)	0	0	0	0
(iii)	0	0	0	0
(iv)	1	1	1	0
(v)	0	0	0	0
(vi)	0	0	0	1

VOT2015

EVALUATION METHODOLOGY

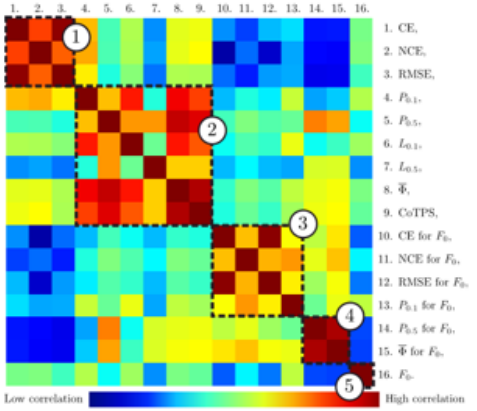
Performance measures

- Target localization properties measured using the VOT2013/VOT2014 methodology.

- Approach in VOT2013/VOT2014:
  - Interpretability of performance measures
  - Select as few as possible to provide clear comparison

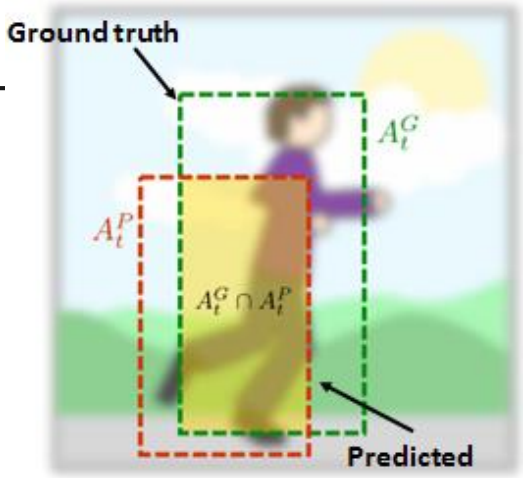
- Based on a recent study[1] two basic weakly-correlated measures are chosen:
  - Robustness
  - Accuracy

Correlation analysis of performance measures[1]



[1]Čehovin, Kristan, and Leonardis, "Is my new tracker really better than yours?",WACV2014

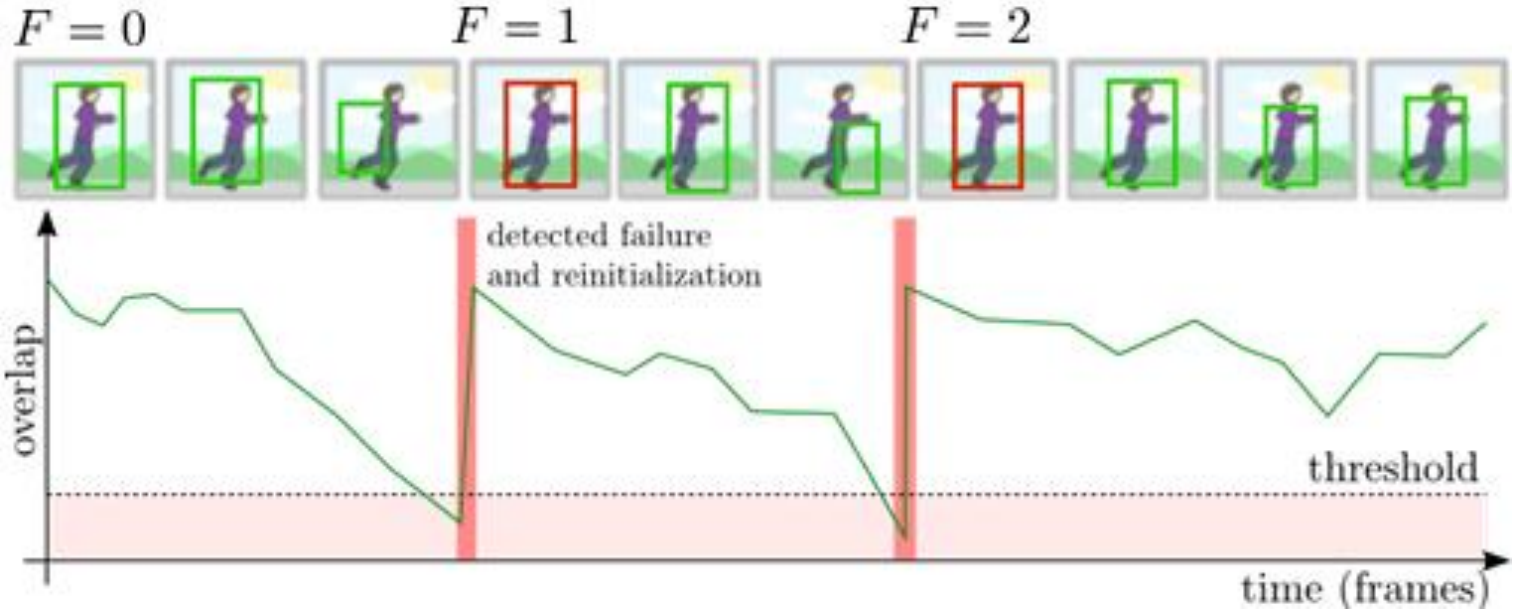
VOT performance measures



- Robustness:

  Number of times a tracker drifts off the target.

- Accuracy: *Average overlap during successful tracking.*



VOT Accuracy/Robustness ranking

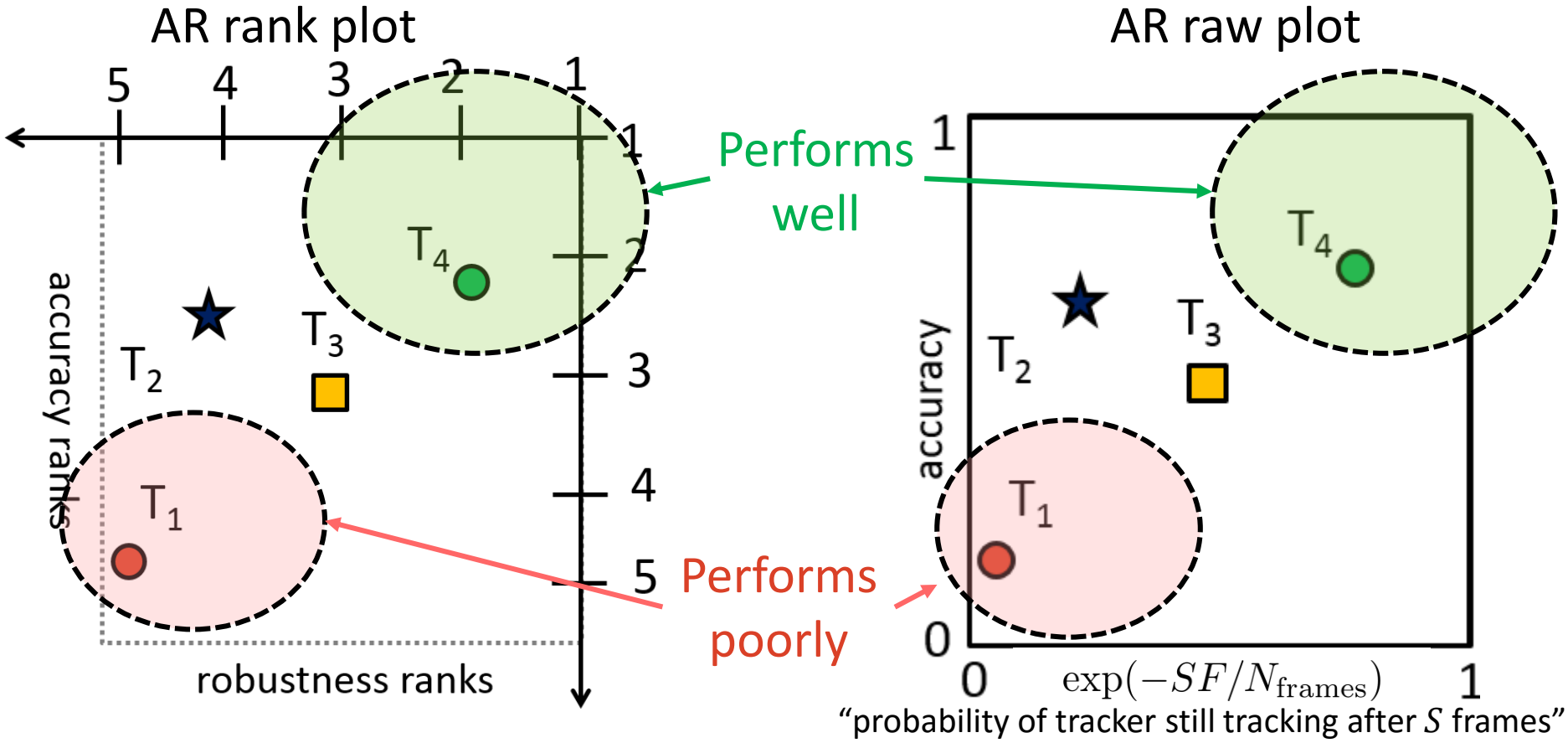
- VOT2014 ranking methodology used[1]
- Rank trackers for accuracy and robustness separately

[1]Kristan et al., A Novel Performance Evaluation Methodology for Single-Target Trackers, ArXiv, 2015

- Two types of ranking
  - Pooled ranking: Concatenate the results from all sequences and rank trackers.
  - Per-attribute ranking: Rank trackers on each attribute subset separately and average the ranks.
- Rank equivalency
  - Several trackers may perform equally well and should be assigned an equal rank.

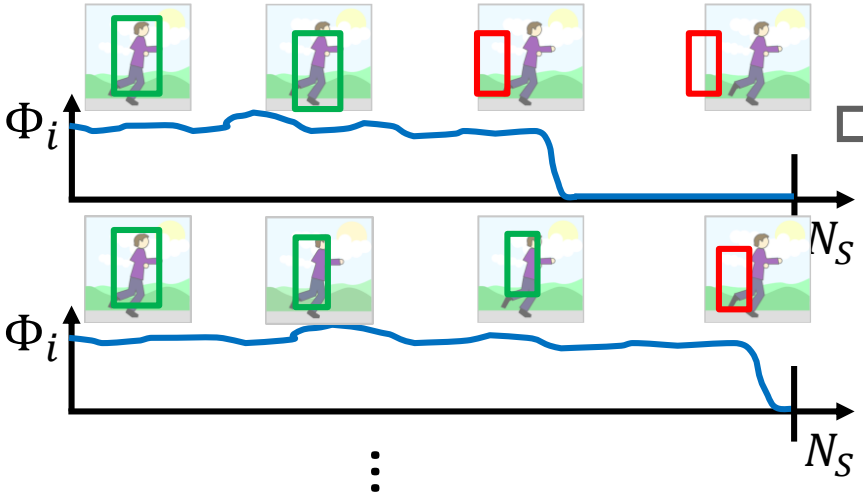
Visualizing the accuracy/robustness

- AR rank plots as proposed in VOT2013
- AR raw plots as proposed by [Čehovin et al. 2014]



New primary performance measure

- A new single score for challenge ranking
- Principled combination of accuracy and robustness
- Roots in application and clear interpretation
- Based on the *"Expected average overlap on N_S frames long sequence."*

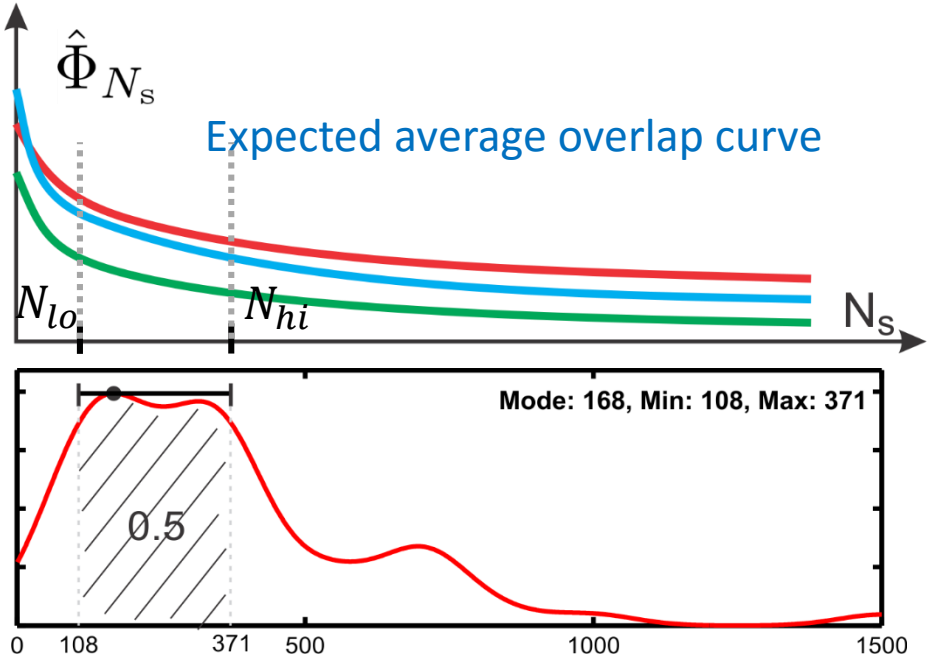


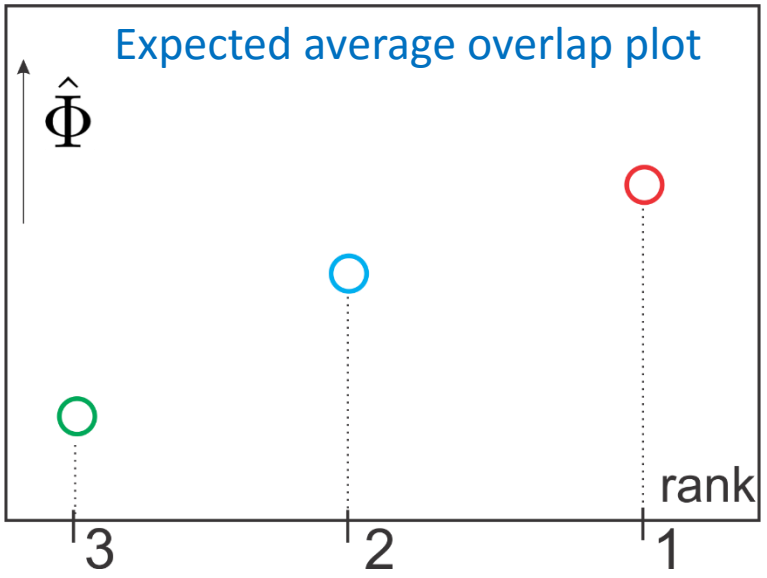
$$\Phi_{N_S} = \frac{1}{N_S}\sum \Phi_i$$

$$\hat{\Phi}_{N_S} = \langle \Phi_{N_S} \rangle$$

New primary performance measure

- Expected average overlap curve: $\hat{\Phi}_{N_S}$ for different values of N_S

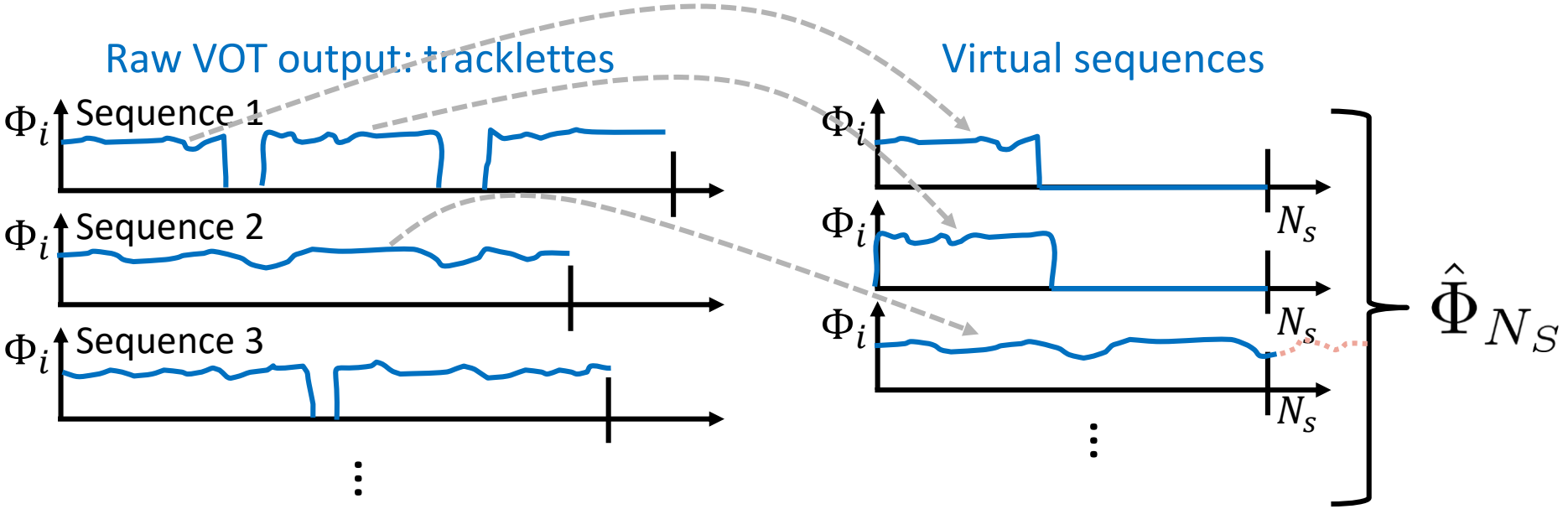




- "VOT2015 expected average overlap measure" $\hat{\Phi}$:
  - Φ_{N_S} averaged over typical short-term sequence lengths interval $[N_{lo}, N_{hi}]$.
  - Pdf of sequence lengths estimated from VOT2015 using KDE [Kristan2009].
  - Interval set to capture 50% of density at the mode.

Implementation of the measure

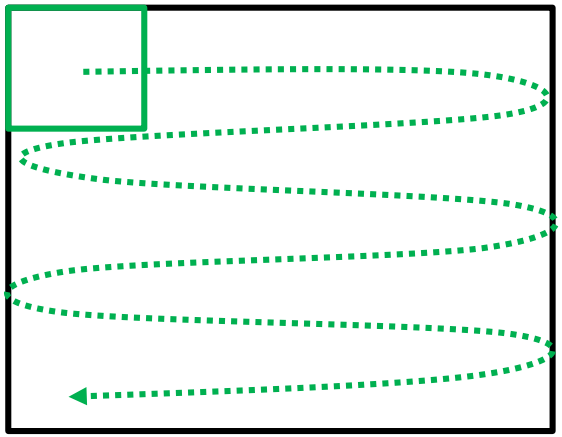
- Require a large dataset of equal length sequences to reduce the variance of the estimator Φ_{N_S} [1].
- Approximate from VOT raw results



[1]Kristan et al., A Novel Performance Evaluation Methodology for Single-Target Trackers, ArXiv, 2015

VOT2015 Speed measurement

- Reduce the hardware bias in reporting tracking speed.
- Approach: The VOT2014 speed benchmark



600x600 image
Max operation in 30x30 window
Apply this filter to all pixels
Measure the time for filtering

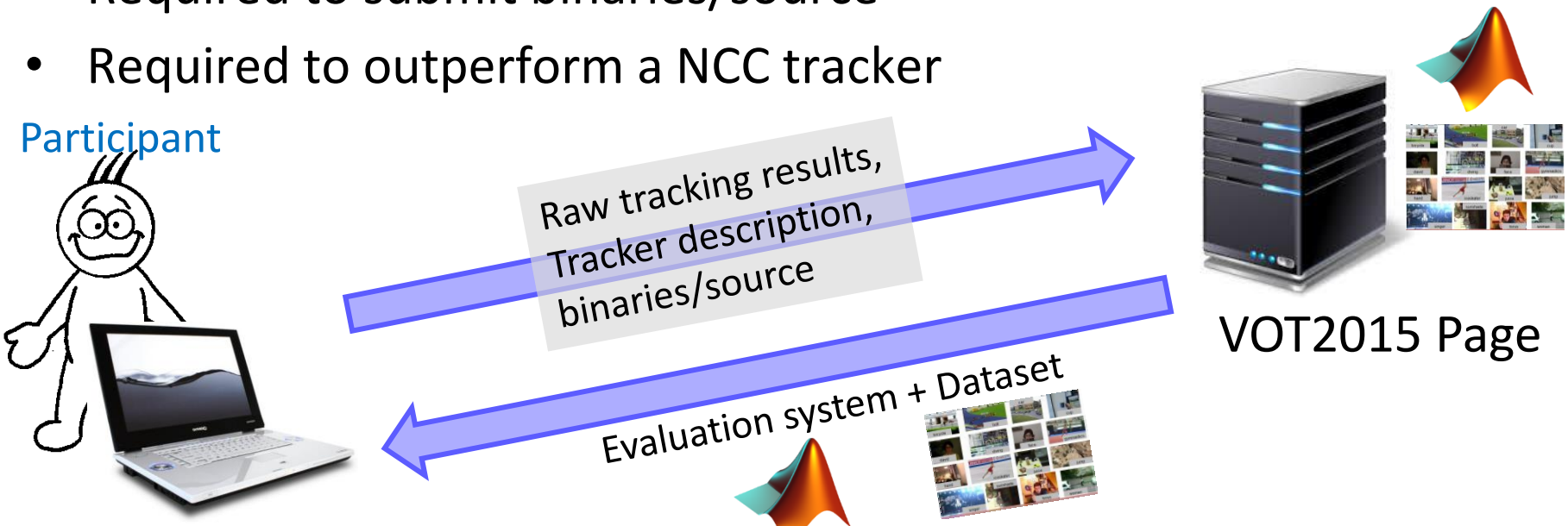
- Divide tracking time with time required to perform the filtering operation
- Equivalent Filter Operations (EFO)

VOT2015

CHALLENGE PARTICIPATION AND SUBMITTED TRACKERS

VOT2015 Challenge: participation

- Participants would download the evaluation kit:
  - Evaluation system + Dataset
- Integrate their tracker into the evaluation system
- Predefined set of experiments automatically performed – submit the results back
- Required to submit binaries/source
- Required to outperform a NCC tracker



62 trackers tested!

Diverse set of entries: 62 = 41 submissions + 21 baselines

- Deep convolutionan neural networks
  (MDNet, DeepSRDCF, SO-DLT)
- Object proposals based
  (EBT, KCFDP,SPST)
- General part-based
  (LDP, TRIC-track, G2T, AOG-track, LGT, HoughTrack,
  MatFlow, CMT, LT-FLO, THANG, FoT, BDF, FCT, FragTrack)
- Global generative-model-based
  (ASMS, SumShift, S3Tracker, PKLTF, DFT, IVT, CT, L1APG, DAT)
- Discriminative models – single part
  (OAB, MIL, MCT, CMIL)
- Discriminative regression-based techniques
  (Struck, RobStruck, SRAT, TGPR, HRP, ACT, KCFv2, DSST, SAMF, SRDCF, PTZ-MOSSE,
  NSAMF, RAJSSC, OACF, sKCF, LOFT-lite, STC, MKCF+, MTSA-KCF, MvCFT)
- Combinations of multiple trackers
  (HMM-TxD, MEEM, SCEBT, MUSTer, SME)

VOT2015

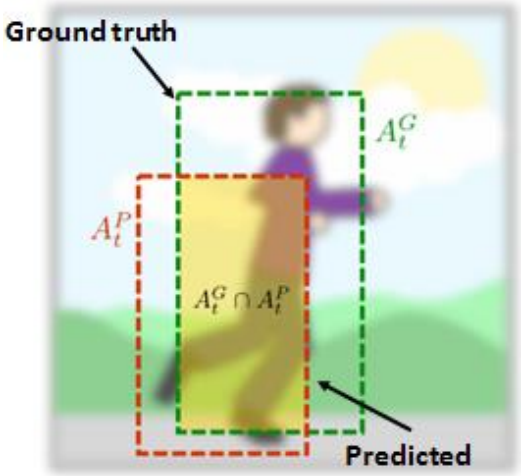
EXPERIMENTS AND RESULTS

VOT2015 Experiment

- Experiment 1– Baseline:
  - All sequences, initialization on ground truth BBs

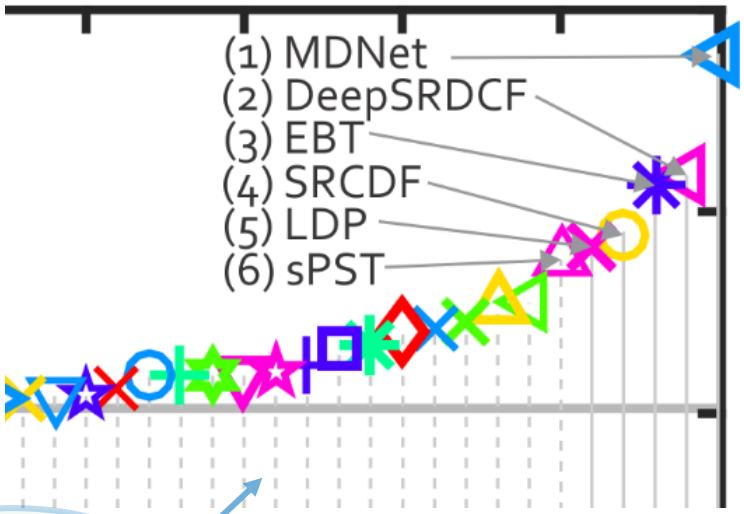
- Each tracker run 15 times on each sequence to obtain a better statistic on its performance.

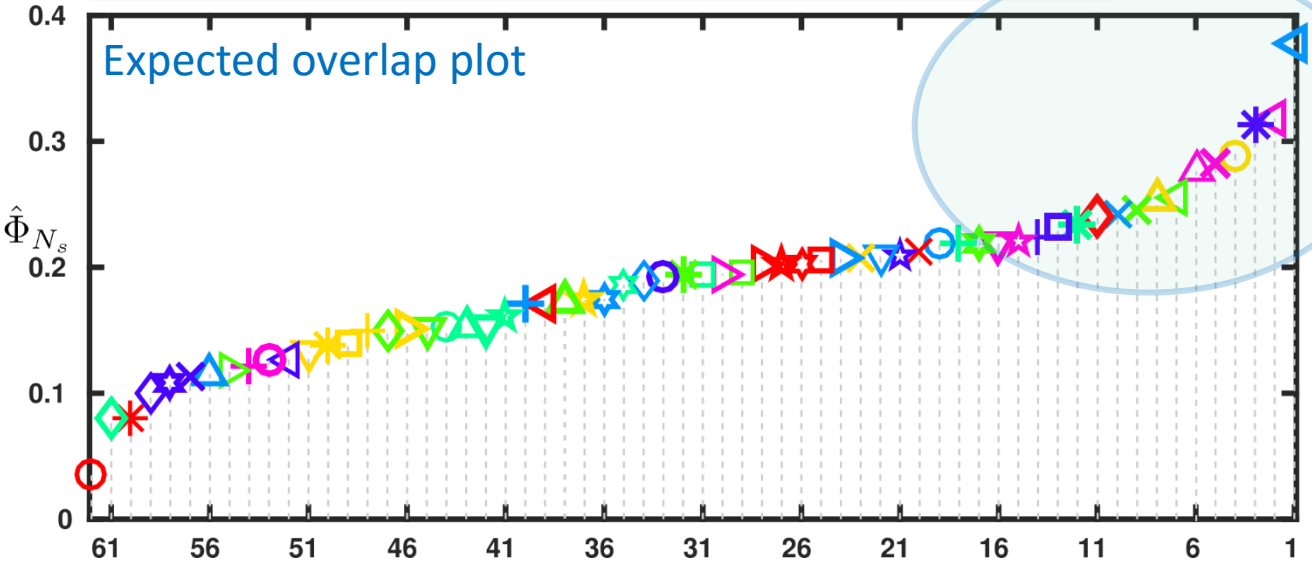
- Reinitialization at overlap 0.



Expected average overlap

Tracker	Type
MDNet*	CNN learned on video sequences
DeepSRDCF	Corr. Filter + CNN feats
EBT	Edgebox features+SSVM+color hist.
SRDCF	Corr. Filter + color names + HoG
LDP	Part-based corr. filter
sPST	Flow + Edgebox feats + SVM





Diverse trackers!

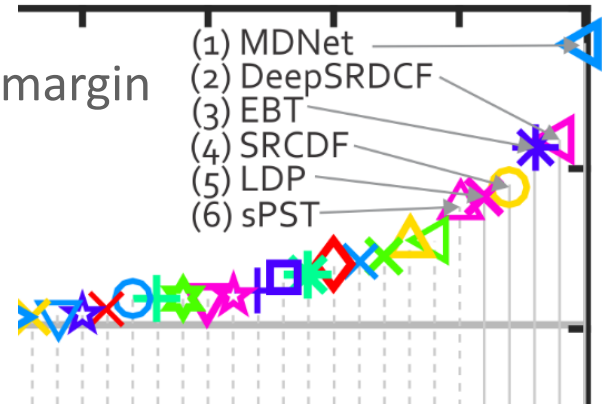
* N., Hyeonseob and H., Bohyung, *Multi-Domain Convolutional Neural Network Tracker*, (Talk today at 13:50)

Detailed analysis

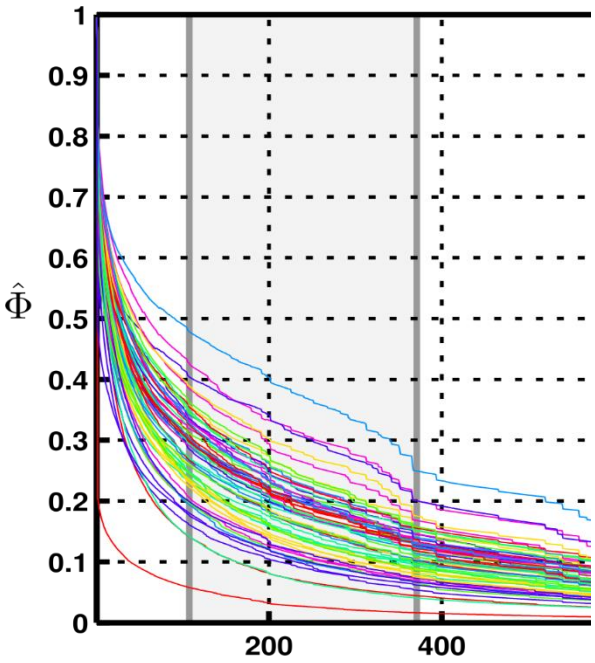
- MDNet ◁ consistently ahead
- DeepSRDCF ◁ similar to EBT ✱
  - EBT fails a bit less, but less accurate.

Obvious margin

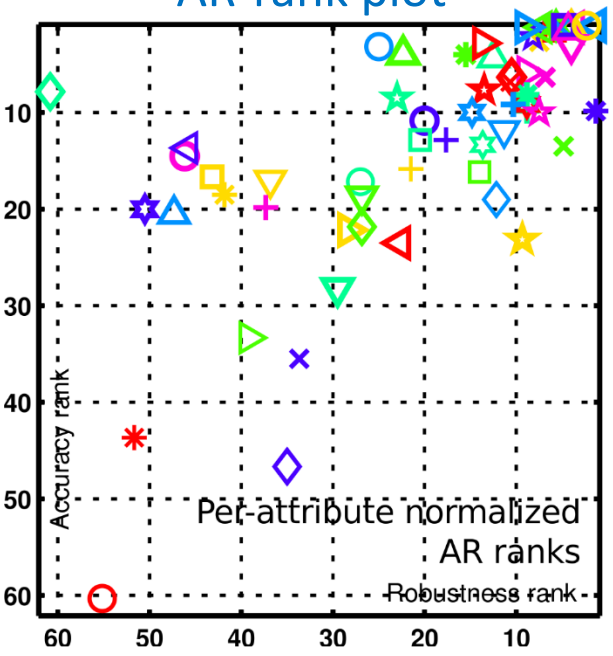
Expected overlap plots



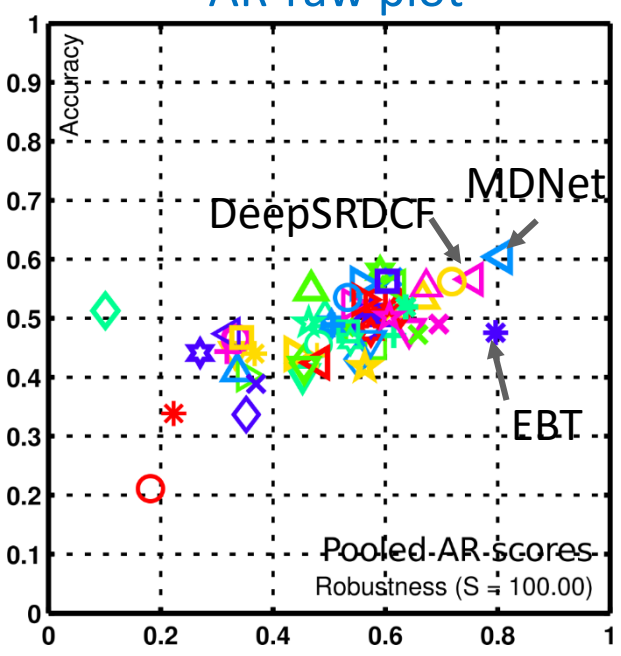
Overlap curves



AR-rank plot



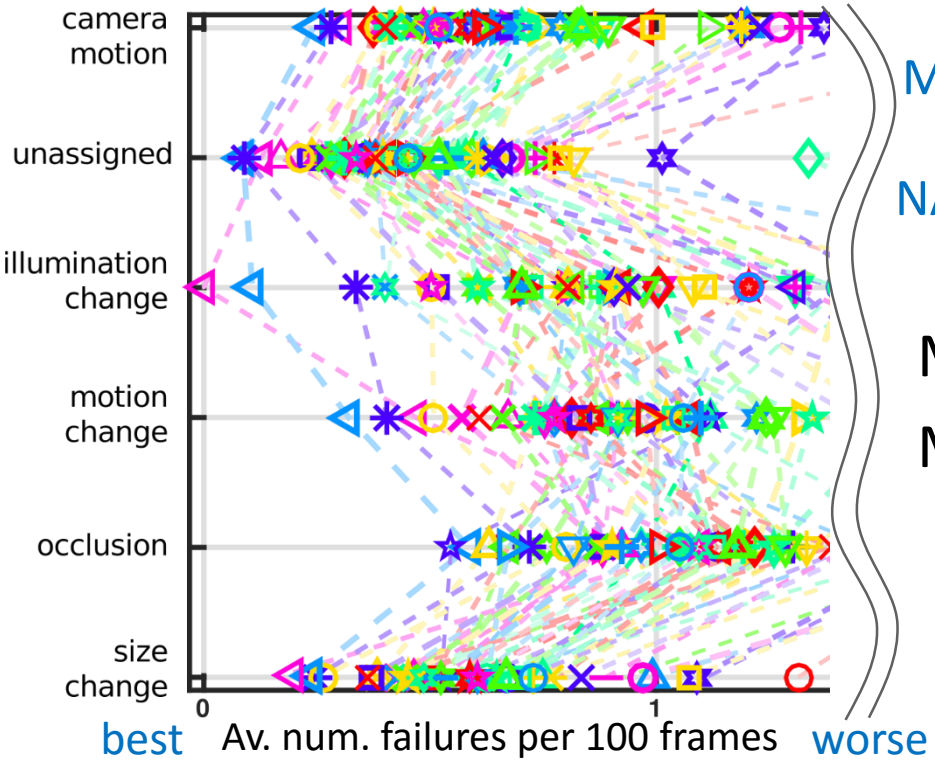
AR-raw plot



Detailed analysis: attributes

- Mostly at the top: MDNet, DeepSRDCF, EBT
- But in occlusion: MKCF+, MDNet, NSAMF

	cam. mot.	unass.	illum. ch.	motion ch.	occlusion	size ch.
A	0.49	0.54	0.49	0.45	*0.41*	**0.39**
R	0.66	0.41	0.89	*0.98*	**1.13**	0.61



MKCF+: Target loss explicitly addressed

NASMF: Multimodel SAMF (VOT2014 top perf.)

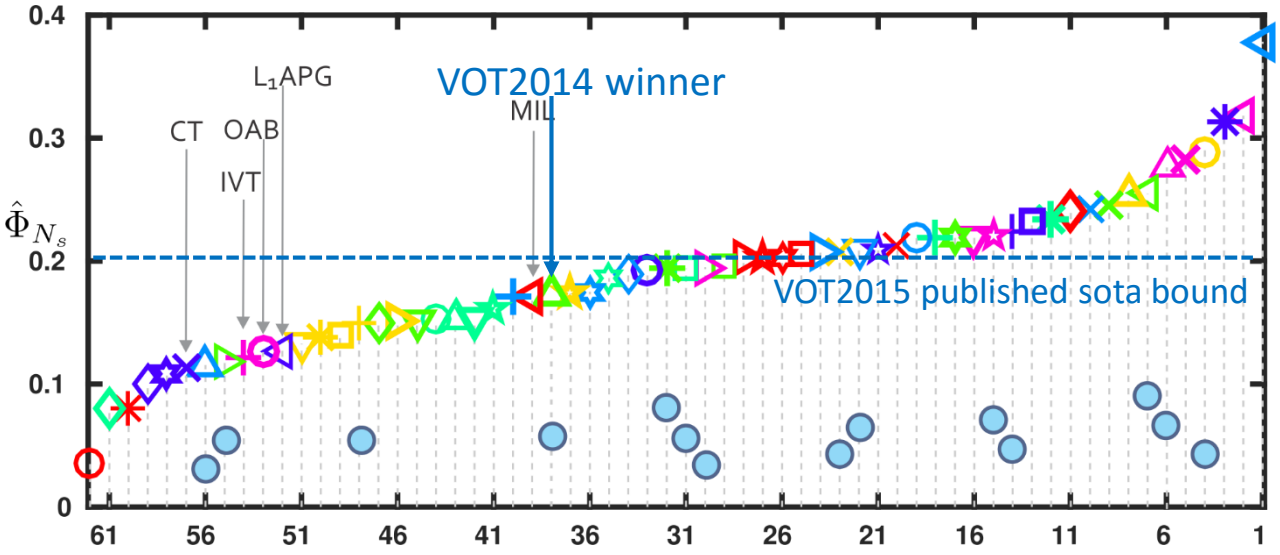
Most challenging (R): occlusion

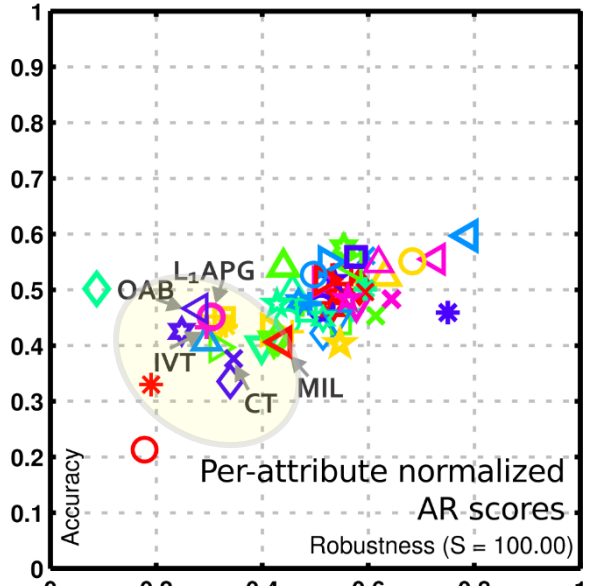
Most challenging (A): size change

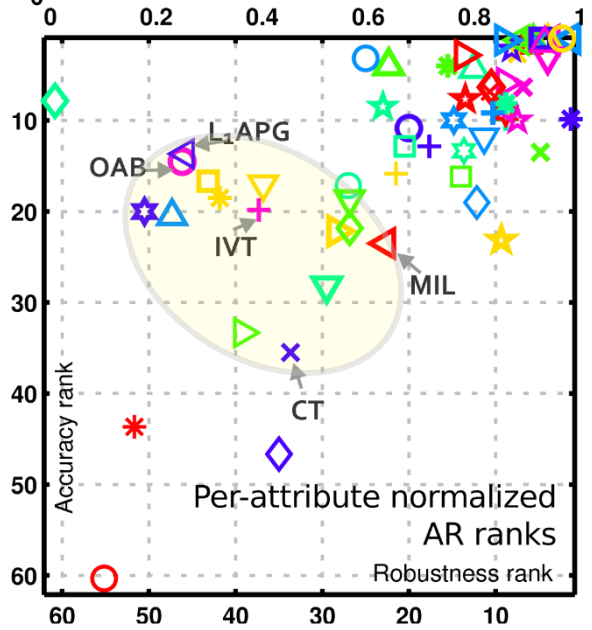
Detailed analysis: baselines + sota

- Baselines: OAB,IVT,CT,MIL,L_1APG
- 14 trackers: (2014-2015) ICCV,ECCV,CVPR,ICML,BMVC
  - Over 40% submissions exceed the VOT2015 published sota bound.
- The VOT2014 winner
  - Over 60% submissions outperform the VOT2014 winner.

Expected overlap plot

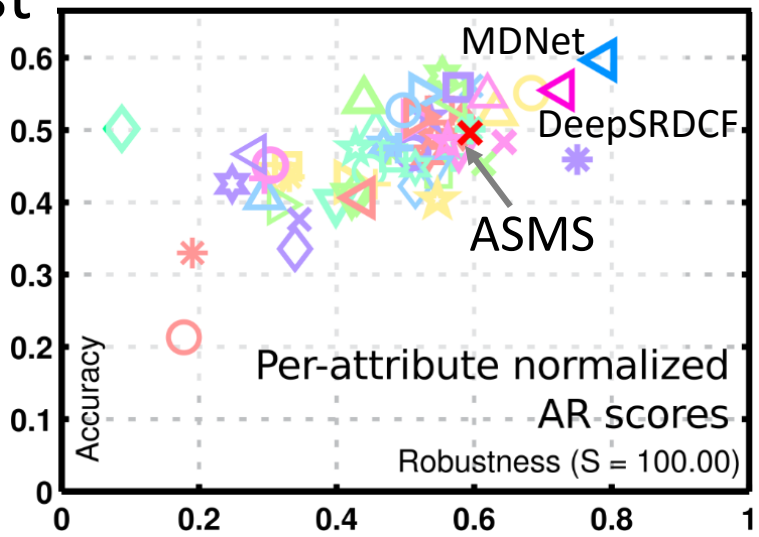






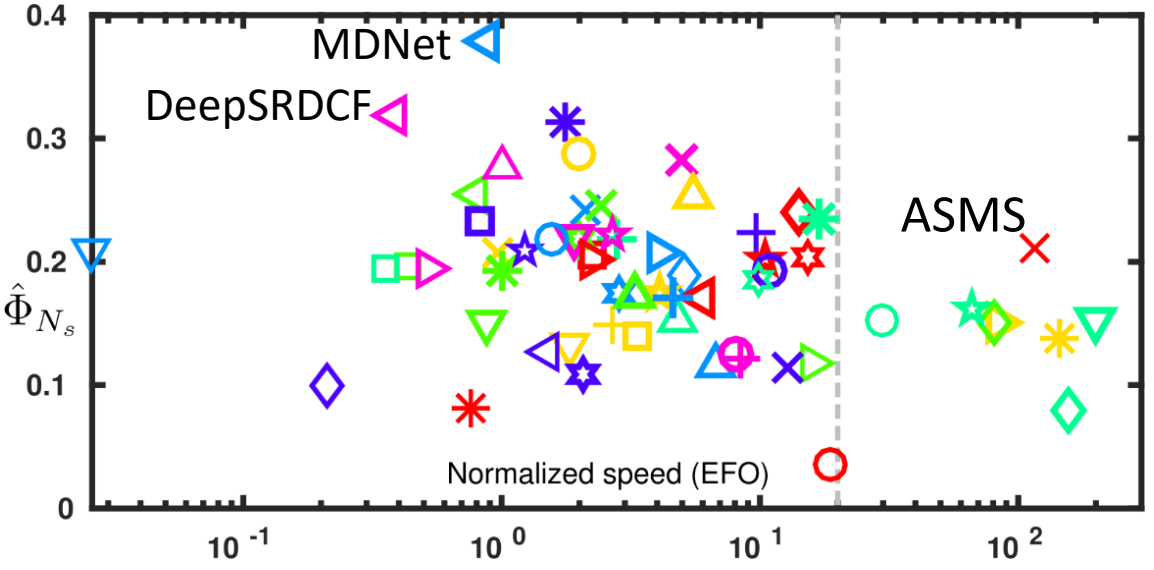
Tracking speed

- Top-performing trackers slowest
  - Plausible cause: CNN ◁ ◁
- Real-time bound: ASMS ✖
  - Decent accuracy,
  - Mid-level robustness
  - Tracks well on short run



For reference:

NCC implemented in C++
Intel Core i5 2557M processor
Measured speed: 140 fps
EFO units: ~160 EFO



Sequence ranking

- VOT2013 approach
  - Average number of trackers failed per frame (A_f)
  - Max. number of trackers failed at a single frame (M_f)

Sequence	Sequence	Sequence	Sequence
Ball1	Gymnastics1	Blanket	Shaking
Ball2	Gymnastics2	Bolt2	Singer2
Birds1	Handball1	Crossing	Sphere
Book	Handball2	Dinosaur	Traffic
Butterfly	Motocross1	Girl	Car2
Gymnastics3	Singer3	Iceskater1	Fish4
Hand	Soccer1	Iceskater2	Godfather
Leaves	Tiger	Nature	Helicopter
Matrix	Bolt1	Wiper	Pedestrian2
Pedestrian1	Car1	Bag	Tunnel
Rabbit	Fernando	Birds2	Fish3
Soccer2	Graduate	Bmx	Sheep
Fish1	Motocross2	Gymnastics4	Octopus
Fish2	Soldier	Marching	Racing
Glove	Basketball	Road	Singer1

Challenging:

$0.1 \leq A_f \leq 0.41$
$31 \leq M_f \leq 60$

$0.04 \leq A_f \leq 0.15$
$15 \leq M_f \leq 56$

Intermediate:

$0.02 \leq A_f \leq 0.09$
$8 \leq M_f \leq 27$

Easiest:

$0.01 \leq A_f \leq 0.02$
$3 \leq M_f \leq 12$

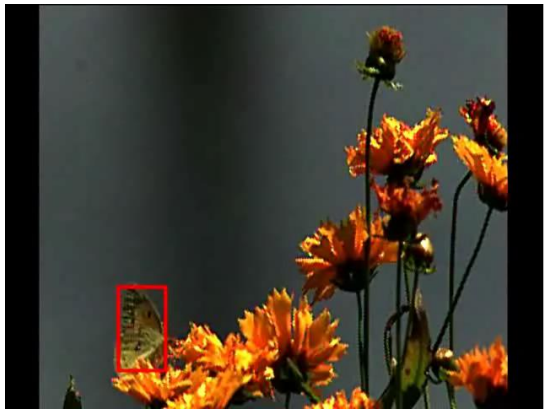
Sequence ranking

- Among the most challenging sequences

Matrix ($A_f = 0.36, M_f = 54$)

Rabbit ($A_f = 0.31, M_f = 39$)

Butterfly ($A_f = 0.22, M_f = 44$)



- Among the easiest sequences

Singer1 ($A_f = 0.01, M_f = 3$)

Octopus ($A_f = 0.01, M_f = 11$)

Sheep ($A_f = 0.02, M_f = 12$)

VOT Summary

- New VOT measure + highly challenging dataset
- Top-performing tracker MDNet (in expected average overlap)
  - AR analysis indicates high accuracy and rare failures
  - Computationally quite complex (EFO)
- Both top-performing trackers applied "learned" features by CNN but different localization strategy
- Most submitted trackers outperform standard baselines
- 40% of submitted trackers outperform the published sota bound as defined in VOT2015.

The VOT2015 online resources

Available at: http://www.votchallenge.net/vot2015

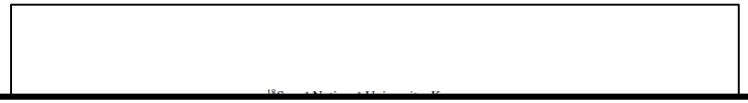
- This presentation + papers + Dataset + Evaluation kit
- Guidelines on how to evaluate your trackers on VOT2015 and produce graphs for your papers (directly comparable to >60 trackers!)
- Two VOT methodology cornerstone papers:
  - Kristan et al., *A Novel Performance Evaluation Methodology for Single-Target Trackers* , ArXiv, 2015 (under review)
  - Čehovin et al., *Visual object tracking performance measures revisited*, ArXiv, 2015 (under review)

Plan to release all versions along with the reviews and our responses

- VOT is open source !
  - Čehovin, ”Ask not what the VOT challenge can do for you …”

VOT2015 summary

- Results published in a 23 pages joint paper ~128 coauthors!

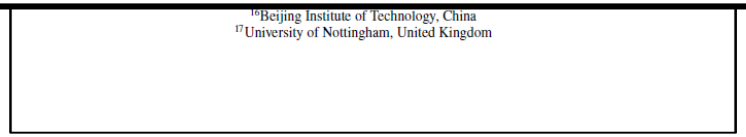


Winners of the VOT2015 challenge:

MDNet by Hyeonseob Nam and Bohyung Han

Multi-Domain Convolutional Neural Network Tracker

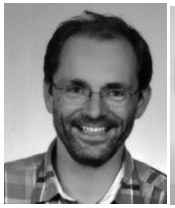
Presentation at VOT2015 today at 13:50



Thanks

- The VOT2015 committee


M. Kristan


J. Matas


A. Leonardis


M. Felsberg


L. Čehovin


T. Vojir


G. Fernandez


G. Häger


G. Nebehay


R. Pflugfelder

- Everyone who participated or contributed

Abhinav Gupta (Carnegie Mellon University), Adel Bibi (King Abdullah University of Science and Technology), Alan Lukežič (Ljubljana University), Alvaro Garcia-Martin (Universidad Autónoma de Madrid), Alfredo Petrosino (Parthenope University of Naples), Amir Saffari (Affectv Limited), Andrés Solís Montero (University of Ottawa), Anton Varfolomieiev (National Technical University of Ukraine), Atilla Baskurt (Universitè de Lyon), Baojun Zhao (Beijing Institute of Technology), Bernard Ghanem (King Abdullah University of Science and Technology), Brais Martinez (University of Nottingham), ByeongJu Lee (Seoul National University), Bohyung Han (POSTECH), Chaohui Wang (Universitè Paris-Est), Christophe Garcia (LIRIS), Chunyuan Zhang (National University of Defense Technology and National Key Laboratory of Parallel and Distributed Processing Changsha), Cordelia Schmid (INRIA Grenoble Rhône-Alpes), Dacheng Tao (University of Technology), Daijin Kim (POSTECH), Dafei Huang (National University of Defense Technology and National Key Laboratory of Parallel and Distributed Processing Changsha), Danil Prokhorov (Toyota Research Institute), Dawei Du (University at Albany and SCCE Chinese Academy of Sciences), Dit-Yan Yeung (Hong Kong University of Science and Technology), Eraldo Ribeiro (Florida Institute of Technology), Fahad Shahbaz Khan (Linköping University), Fatih Porikli (Australian National University and NICTA), Filiz Bunyak (University of Missouri), Gao Zhu (Australian National University), Guna Seetharaman (Naval Research Lab), Hilke Kieritz (Fraunhofer IOSB), Hing Tuen Yau (Chinese University of Hong Kong), Hongdong Li (Chinese University of Hong Kong and ARC Centre of Excellence for Robotic Vision), Honggang Qi (University at Albany and SCCE Chinese Academy of Sciences), Horst Bischof (Graz University of Technology), Horst Possegger (Graz University of Technology), Hyemin Lee (POSTECH), Hyeonseob Nam (POSTECH), Ivan Bogun (Florida Institute of Technology), Jae-chan Jeong (Electronics and Telecommunications Research Institute), Jae-il Cho (Electronics and Telecommunications Research Institute), Jae-Yeong Lee (Electronics and Telecommunications Research Institute), Jianke Zhu (Zhejiang University), Jianping Shi (CUHK), Jiatong Li (Beijing Institute of Technology and University of Technology), Jiaya Jia (CUHK), Jiayi Feng (Institute of Automation Chinese Academy of Sciences), Jin Gao (Institute of Automation Chinese Academy of Sciences), Jin Young Choi (Seoul National University), Ji-Wan Kim (Electronics and Telecommunications Research Institute), Jochen Lang (University of Ottawa), Jose M. Martinez (Universidad Autónoma de Madrid), Jongwon Choi (Seoul National University), Junliang Xing (Institute of Automation Chinese Academy of Sciences), Kai Xue (Harbin Engineering University), Kannappan Palaniappan (University of Missouri), Karel Lebeda (University of Surrey), Karteek Alahari (INRIA Grenoble Rhône-Alpes), Ke Gao (University of Missouri), Kimin Yun (Seoul National University), Kin Hong Wong (Chinese University of Hong Kong), Lei Luo (National University of Defense Technology), Liang Ma (Harbin Engineering University), Lipeng Ke (University at Albany and SCCE Chinese Academy of Sciences), Longyin Wen (University at Albany), Luca Bertinetto (Oxford University), Mahdieh Pootschi (University of Missouri), Mario Maresca (Parthenope University of Naples), Martin Danelljan (Linköping University), Mei Wen (National University of Defense Technology and National Key Laboratory of Parallel and Distributed Processing Changsha), Mengdan Zhang (Institute of Automation Chinese Academy of Sciences), Michael Arens (Fraunhofer IOSB), Michel Valstar (University of Nottingham), Ming Tang (Institute of Automation Chinese Academy of Sciences), Ming-Ching Chang (University at Albany), Muhammad Haris Khan (University of Nottingham), Nana Fan (Harbin Institute of Technology), Naiyan Wang (Hong Kong University of Science and Technology and TuSimple LLC), Ondrej Miksik (Oxford University), Philip Torr (Oxford University), Qiang Wang (Institute of Automation Chinese Academy of Sciences), Rafael Martin-Nieto (Universidad Autónoma de Madrid), Rengarajan Pelapur (University of Missouri), Richard Bowden (University of Surrey), Robert Laganière (University of Ottawa), Salma Moujtahid (Universitè de Lyon), Sam Hare (Obvious Engineering Limited), Simon Hadfield (University of Surrey), Siwei Lyu (University at Albany), Siyi Li (Hong Kong University of Science and Technology), Song-Chun Zhu (University of California), Stefan Becker (Fraunhofer IOSB), Stefan Duffner (Universitè de Lyon and LIRIS), Stephen L Hicks (Oxford University), Stuart Golodetz (Oxford University), Sunglok Choi (Electronics and Telecommunications Research Institute), Tianfu Wu (University of California), Thomas Mauthner (Graz University of Technology), Tony Pridmore (University of Nottingham), Weiming Hu (Institute of Automation Chinese Academy of Sciences), Wolfgang Hübner (Fraunhofer IOSB), Xiaomeng Wang (University of Nottingham), Xin Li (Harbin Institute of Technology), Xinchu Shi (Institute of Automation Chinese Academy of Sciences), Xu Zhao (Institute of Automation Chinese Academy of Sciences), Xue Mei (Toyota Research Institute), Yao Shizeng (University of Missouri), Yang Hua (INRIA Grenoble Rhône-Alpes), Yang Li (Zhejiang University), Yang Lu (University of California), Yuezun Li (University at Albany), Zhaoyun Chen (National University of Defense Technology and National Key Laboratory of Parallel and Distributed Processing Changsha), Zehua Huang (Carnegie Mellon University), Zhe Chen (University of Technology), Zhe Zhang (Baidu Corporation), Zhenyu He (Harbin Institute of Technology), and Zhibin Hong (University of Technology).

VOT2015 main sponsor:

University *of Ljubljana*
Faculty *of Computer and Information Science*

VOT2015 sponsors:

SICK
Sensor Intelligence.

Autoliv