VISMAC 2016
SUMMER SCHOOL

june 2016, GRADO

Human Motion Understanding

RE-IDENTIFICATION and
MULTIPLE-TARGET TRACKING
in environmental and egocentric views

Rita Cucchiara, Simone Calderara, **Francesco Solera**

DIPARTIMENTO DI INGEGNERIA Enzo Ferrari

Università di Modena e Reggio Emilia, Italia

http://www.imagelab.ing.unimore.it

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

TOPICS OF TODAY

1) Introduction, motivation and applications
2) The pipeline: the tracking models
3) Tracking by detection.. Thus we need detection, first!
4) And Re-identification?

5) If Single target tracking is difficult,
6) **Multiple target tracking** is more difficult…

7) The problem of **performance**
8) New ideas: Cognitive Based and Deep Learning based MTT

9) Multi camera Multiple target tarcking
10) Conclusions, at the end.

1. INTRODUCTION,MOTIVATIONS AND APPLICATION

RESEARCH TOPICS

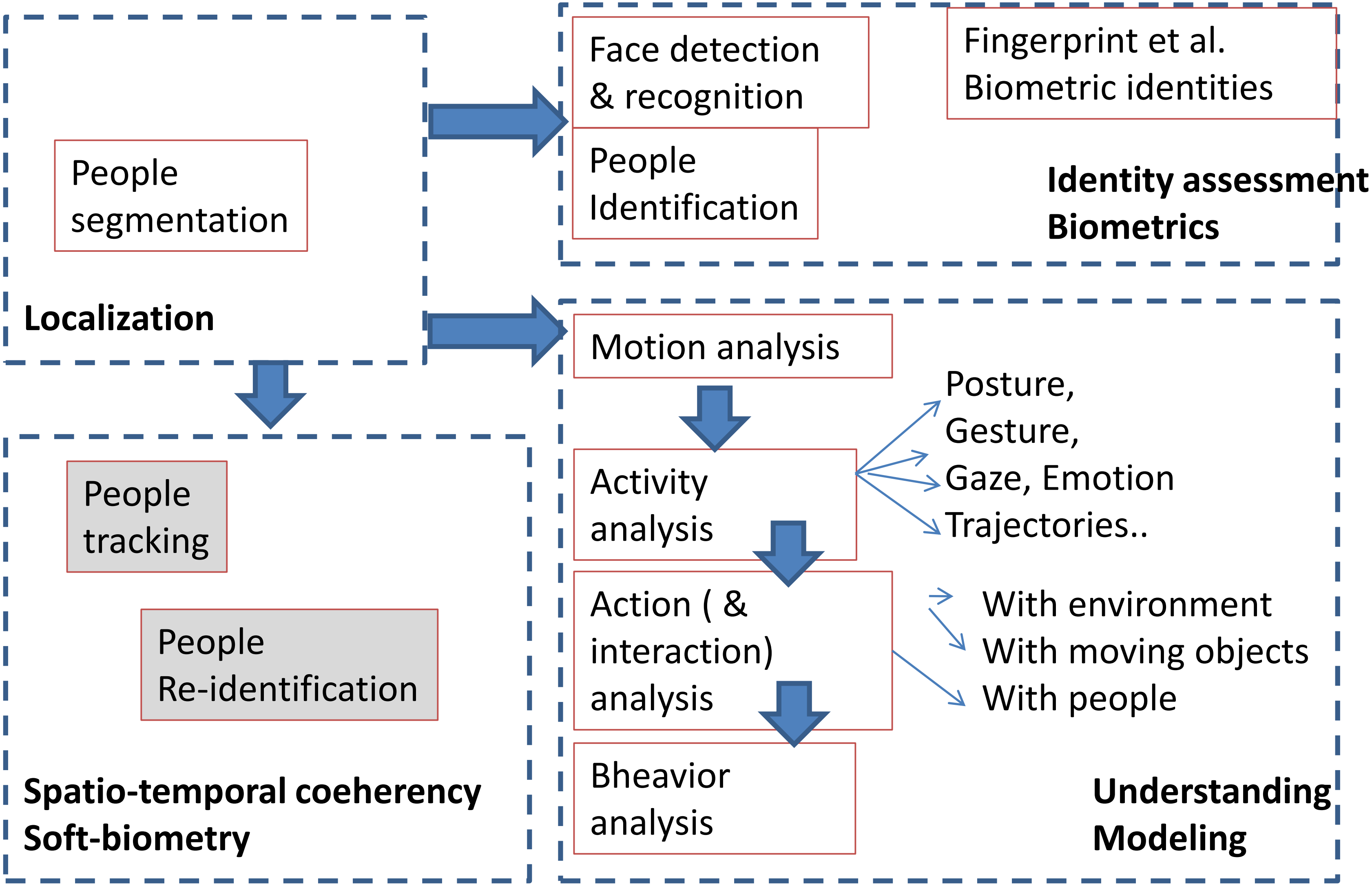
Googling at «google scholar», for research topics

(search june 2016)

VERY HOT RESEARCH TOPICS !!

	Since 2012	Since 2015	Average 2012-2014
deep learning	23000	16600	2133
video-surveillance	18300	8820	3160
tracking video	267000	81300	61900
background suppression	5190	1620	1190
multi- target tracking	3150	1620	510
people detection	3160	999	720
people tracking	3950	1120	943
re-identification	8430	3020	1803
egocentric vision	291	161	43

FROM DETECTING TO REASONING ON PEOPLE



Application trends

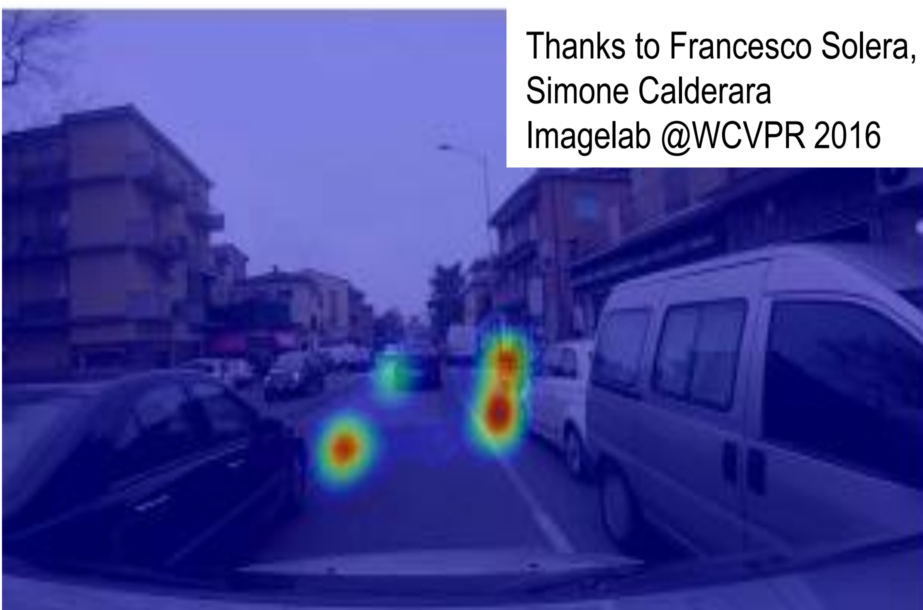
SURVEILLANCE

SPORT AND ENTERTAINMENT

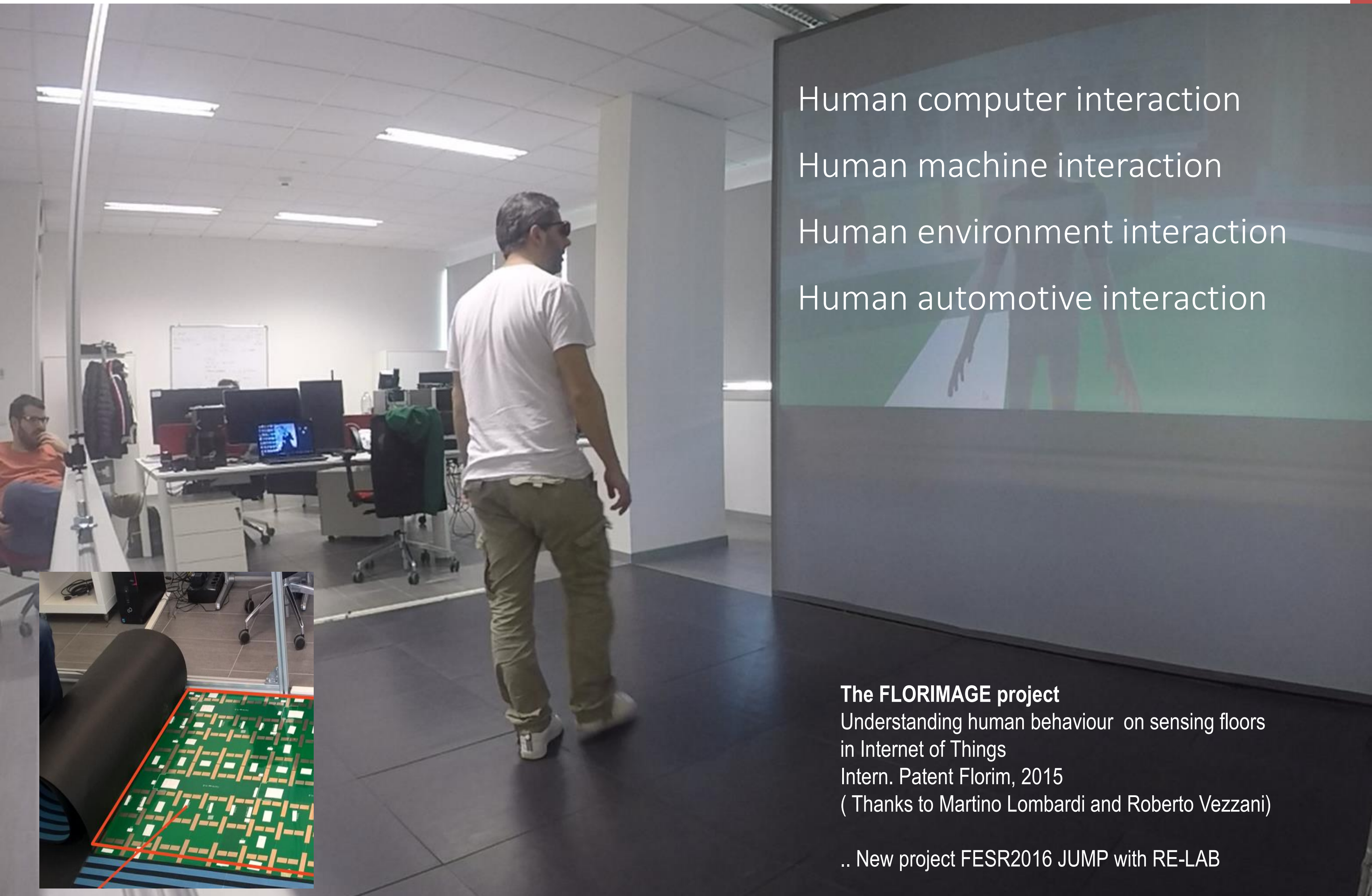
AUTOMOTIVE

DR(eye)VE: a Dataset for Attention-Based Tasks with Applications to Autonomous and Assisted Driving

Thanks to Francesco Solera, Stefano Aletto, Andrea Palazzi, Simone Calderara
Imagelab @WCVPR 2016



HUMAN-X-INTERACTION

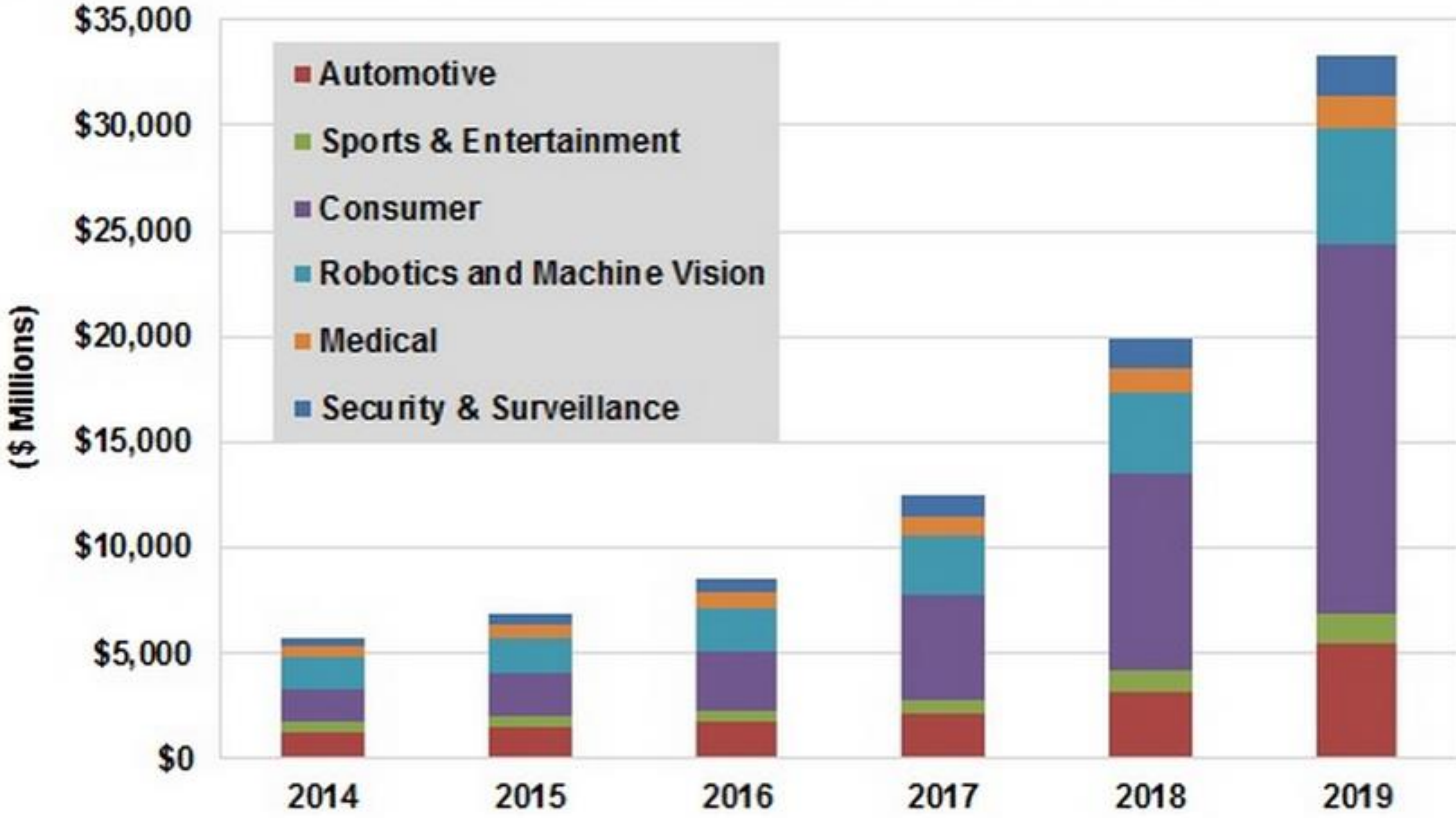


EGOCENTRIC VIEW

Tractica

More in large!

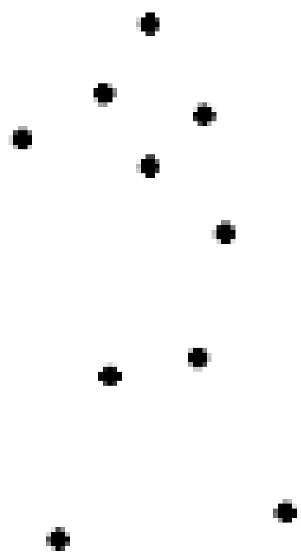
Computer Vision Revenue by Vertical Market, World Markets: 2014-2019



Source: Tractica

2. THE PIPELINE: THE TRACKING MODELS

WHAT CAN WE NEED

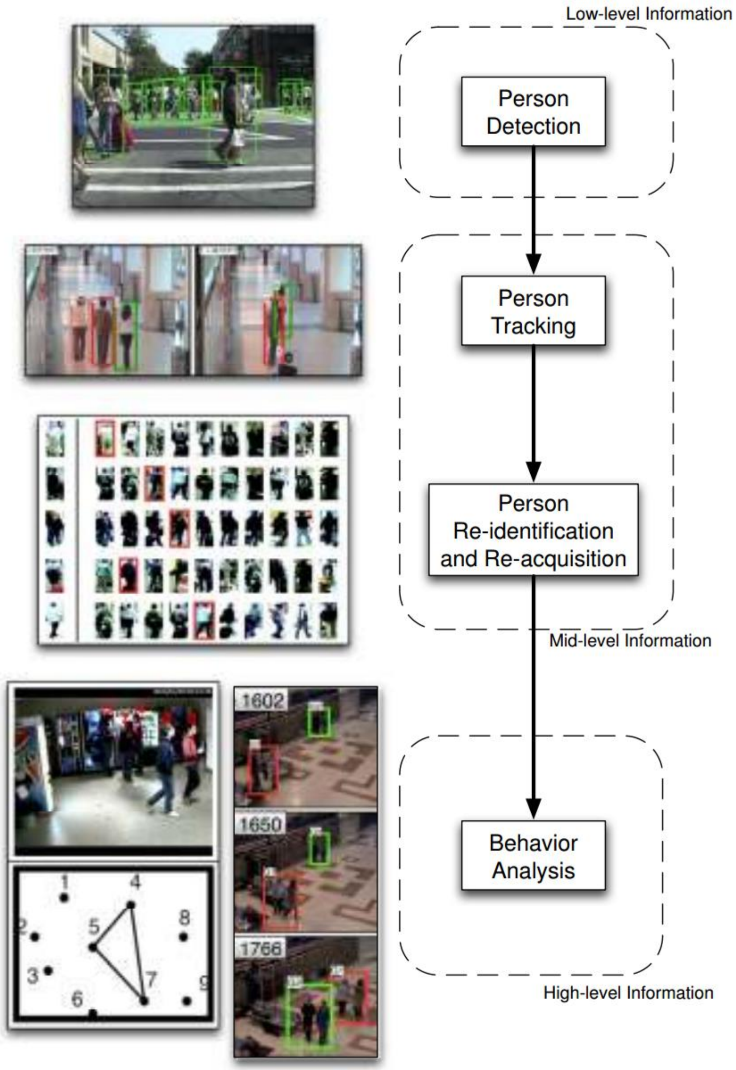


Humans recognize motion and recognize by motion

We need

- Computer vision for 2D to 3D space
- Frame spatial analysis (localization)
- Video motion analysis for Temporal coherency
- Learning
- Recognizing similar patterns in the space and time

THE CLASSIC PIPELINE (IN SURVEILLANCE)

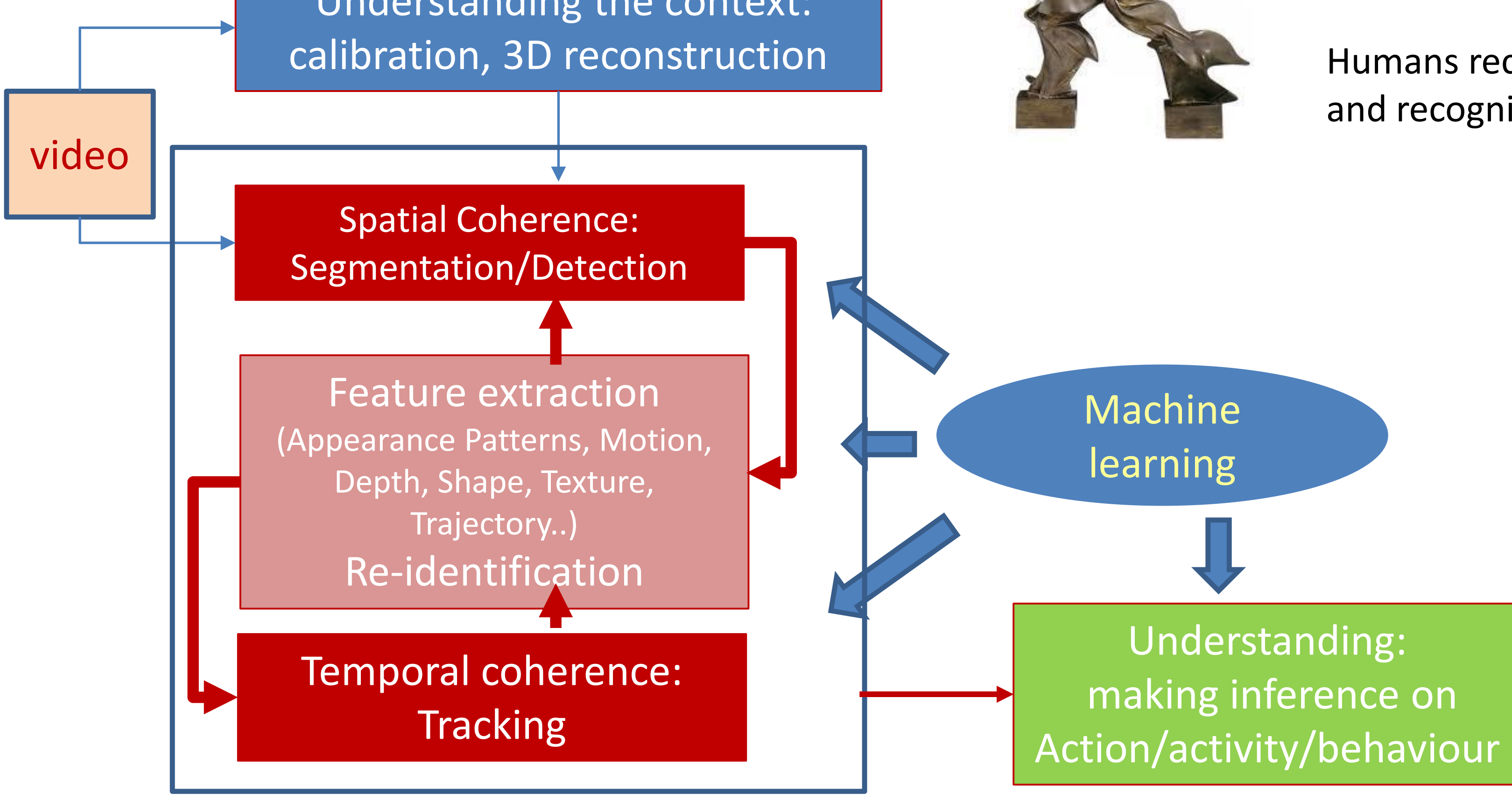


Humans recognize motion and recognize by motion

From *Beyond Multiple target tracking* Loris Bazzani phd thesis Verona 2012

PROBABLY IT'S NOT A PIPELINE

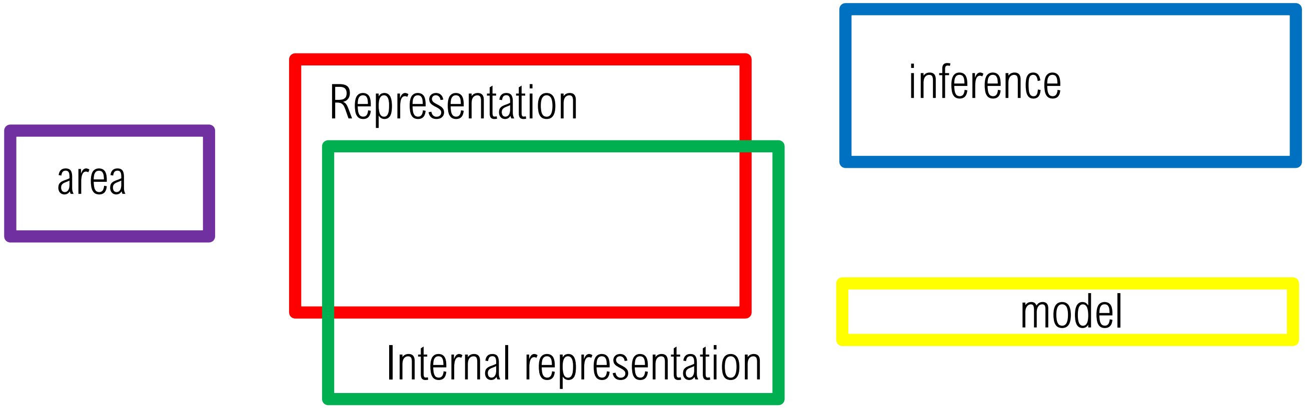
Humans recognize motion and recognize by motion



THE FIVE COMPONENTS IN TRACKING

The (unsolved) questions in tracking a single or multiple targets

1) Which **area** to track?
2) Which **visual/ motion features and representation** to extract and how?
3) Which **model/status** to update and keep in an internal representati9on short and long term memory?
4) Which **inference** to provide?
5) Which model **prediction** for the spatial and temporal coherence?



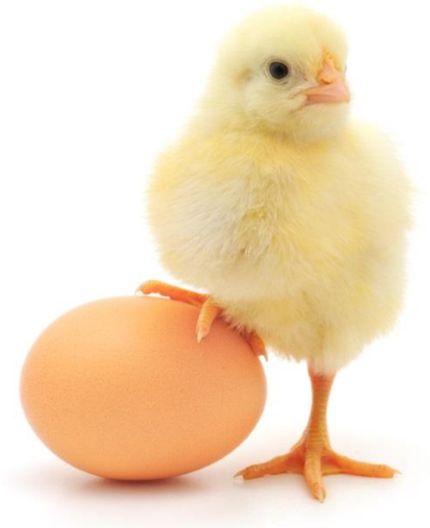
DETECTION AND TRACKING

Two connected dichotomies

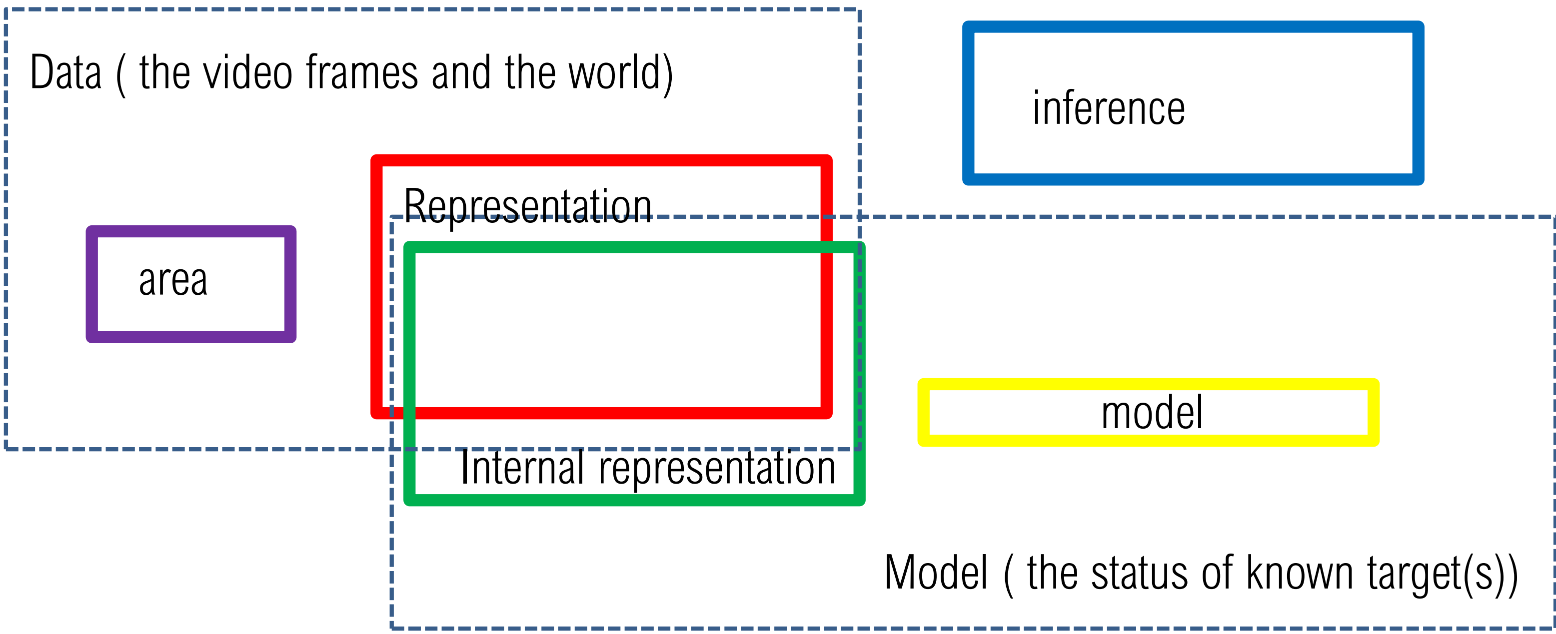
1) DATA DRIVEN vs MODEL DRIVEN

1) TRACKING BY DETECTION vs DETECTION BY TRACKING

The connection between tracking and detection is debated since the famous 2000's PAMI special issues…



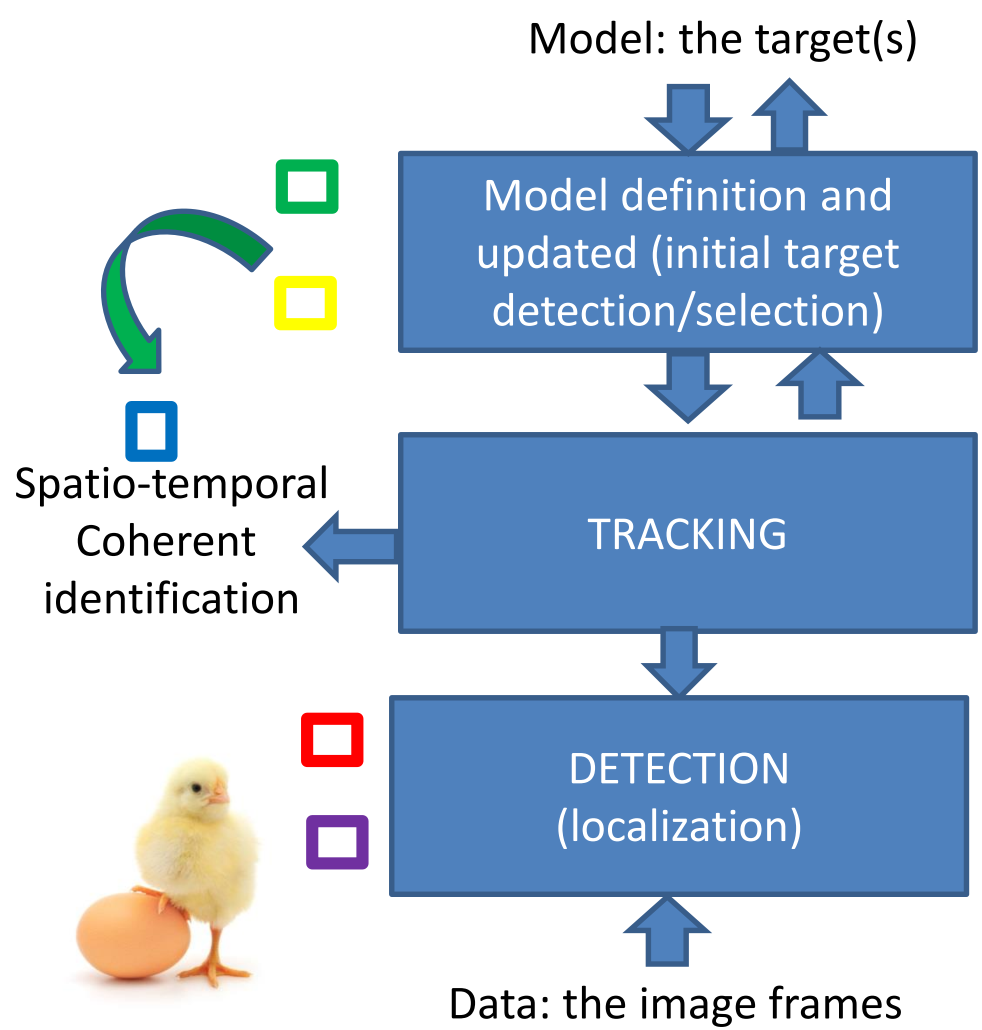
THE FIVE COMPONENTS



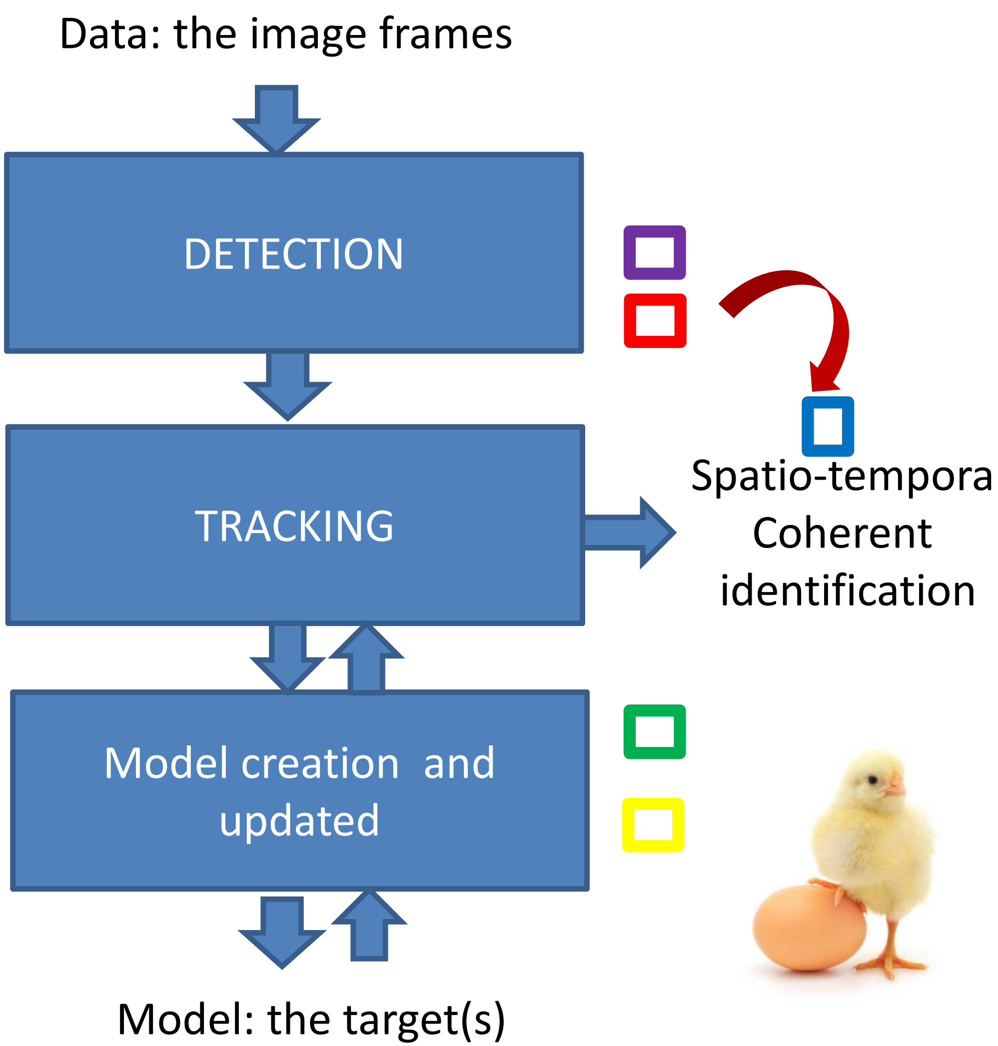
(DETECTION BY) TRACKING

Detection by tracking
or tracking without detection

- For single target tracking
- When an initialization is given (multimedia)
- When users are involved (HCI)
- In disjoined multi-camera (re-identification)

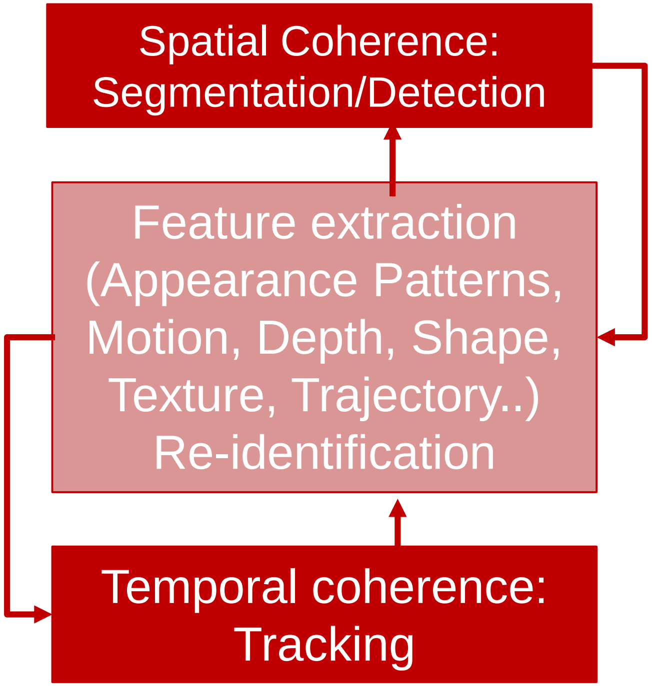


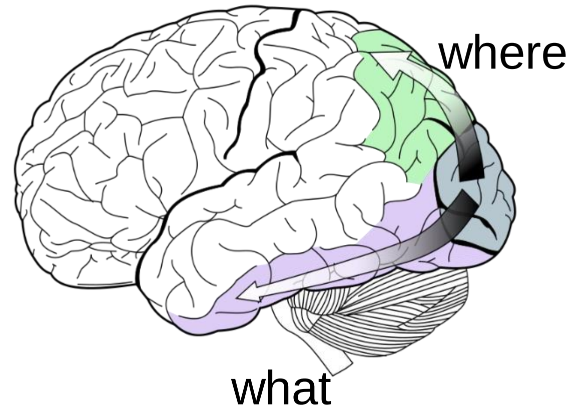
TRACKING BY DETECTION



Tracking by detection

- When detection is easy (video-surveillance)
- When detection is easier than prediction (MTT, crowd)
- In overlapped multicamera





Our visual behavior

Is not so different…

The path:

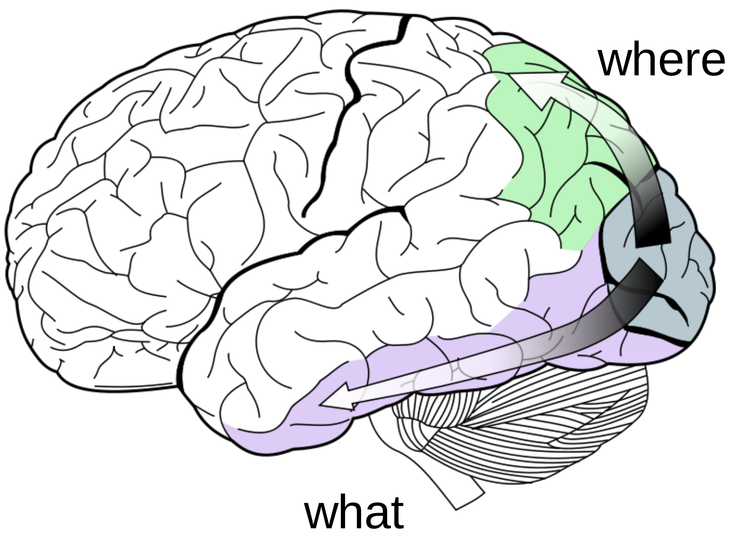
The stimuli from retinae through two parallel path reach the lateral geniculate nucleus in thalamus, then to the cortex in the occipital lobe and then

1) Two parallel paths
    1) The way of WHAT in the temporal lobe perceives color, shape of the object, the face..
    2) The way of WHERE in parietal lobe provides localization during the time of such objects
2) Centers hierarchically connected, process information and work together

FOR HUMANS TOO

The path: (E. Kendall, 2008)

1) The stimuli from retinae through two parallel path reach the lateral geniculate nucleus in thalamus, then to the cortex in the occipital lobe and then in the temporal and frontal lobes.
2) Two parallel paths
    1) The way of WHAT in the temporal lobe perceives color, shape of the object, the face..
    2) The way of WHERE in parietal lobe provides localization during the time of such objects
3) Centers hierarchically connected, process information and than come back to the WHAT area and work together

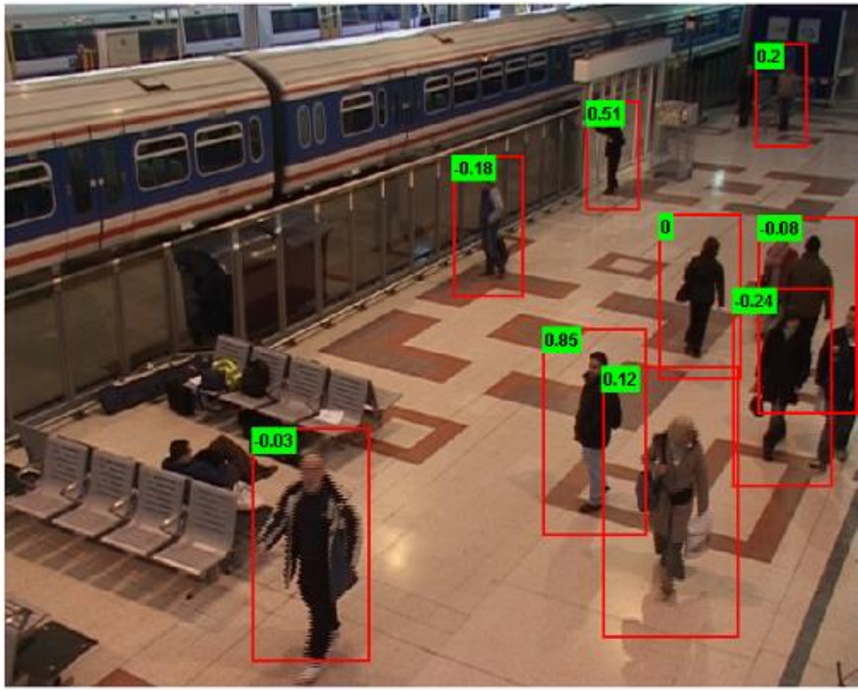


Based on attention and purpose

A FEW WORDS ON DETECTION

Target detection:

A. No target model (implicit in the context)
B. A given target model (humans, vehicle, animals…)
C. Learning target model by many many examples



A) NO TARGET MODEL

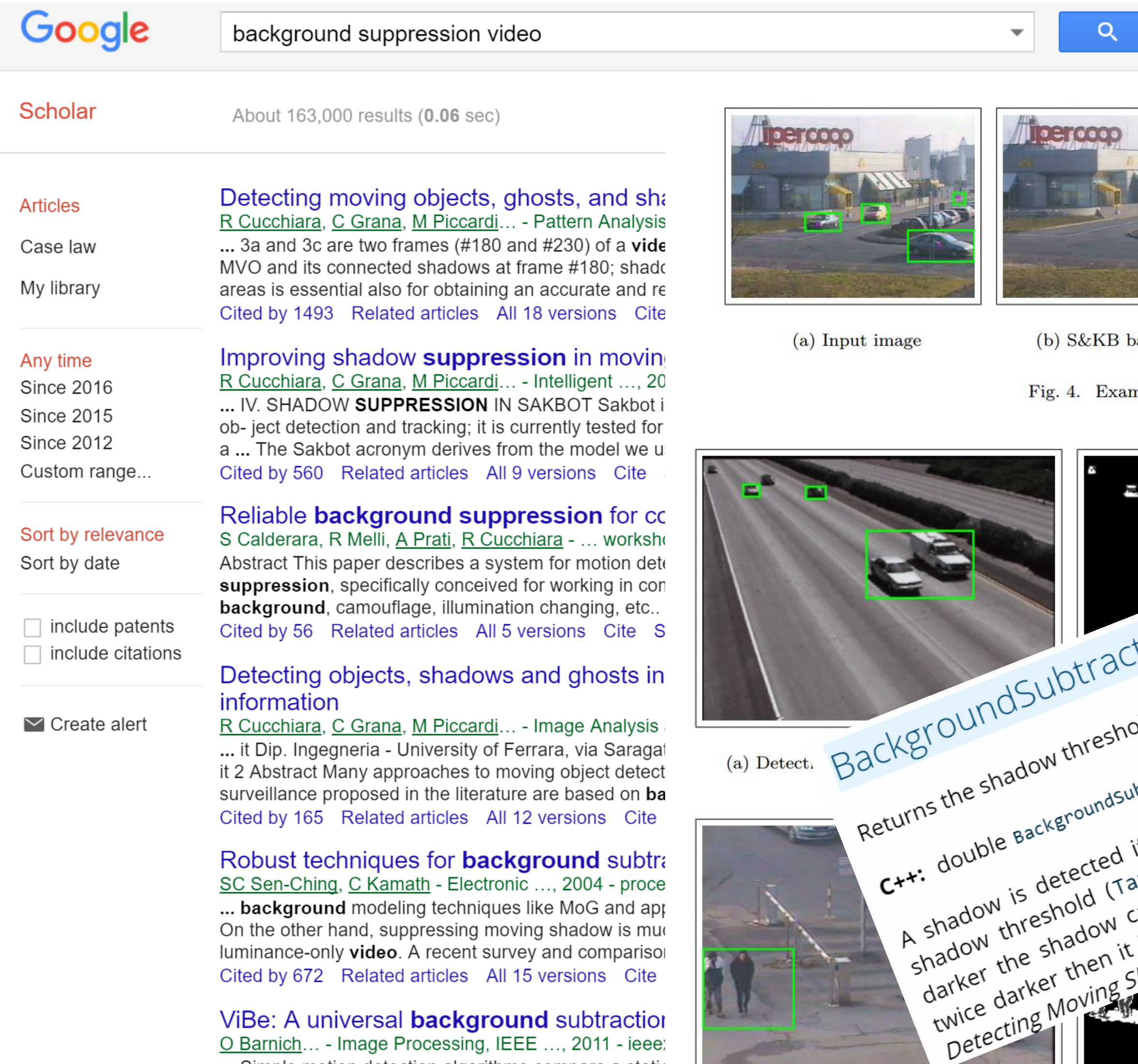
Detection by background suppression

- For static camera(s)
- For cameras with constrained motion (Ptz)
- In surveillance

Still open (?) questions

- Background initialization
- Background model update
- Background suppression

BACKGROUND SUPPRESSION– AN OLD RESEARCH



(a) Input image (b) S&KB background (c) Shadow detection (d) MVOs extracted

Fig. 4. Examples of Sakbot system in urban traffic scene

(a) Detect... ...suppression

(d) Detection w/o shadow suppression (e) Shadow detection (f) Detection with shadow suppression

MOG and shadows MOG2- OpenCv

BACKGROUND INITIALIZATION

Fast Background Initialization with Recursive Hadamard Transform

AVSS 2010

Davide Baltieri, Roberto Vezzani, Rita Cucchiara

2010 single core processor Results

Table 3: Timing results

	Frame Update (ms)	Background Estimation (ms)
Median	46	3133
DCT-based Method [14]	29	1506
RHT	**36**	**97**

Table 4: Averaged results using CAVIAR dataset

	Average error	Clustered error pixels
Median	16.00	5451
DCT-based Method [14]	14.12	3822
RHT	**12.55**	**2334**

Table 5: Averaged results using ViSOR dataset

	Average error	Clustered error pixels
Median	11.080	1929
DCT-based Method [14]	13.55	1807
RHT	**12.62**	**968**

A B C D E

Figure 4: Examples from two VISOR and two CAVIAR videos: (A,B) two random frames, (C) Estimated background us the median filter, (D) using the DCT based method of Reddy *et al.* ([14]), (E) Our proposed enhanced method

BACKGROUND SUPPRESSION NEWS

very few news

A modified Gaussian mixture **background** model via spatiotemporal distribution with shadow detection H Xia, S Song, L He - Signal, Image and **Video** Processing, 2016 – Springer

Background Subtraction Methods in Video Streams: A Review
Saba Joudaki, Mohd Shahrizal bin Sunar, Hoshang Kolivand, Dzulkifli bin Mohamad JSCDSS 2016 Malaysia (!)

Many improvements for technical engineering applications
Very few top-rank publications in the last five-years
(See google scholar)
Good commercial solutions
If you work we static cameras…. Use it!

B) WHEN OBJECT MODEL IS KNOWN…
C) WHEN IT IS LEARNED

Detectors

People detectors (and other targets)

Pedestrian detectors a long story…

- Detectors: Dalal, Triggs CVPR05, Felzenszwalb, CVPR08, Gavrila et al IJCV08, PAMI09
- Benchmarks: Dollar et al CVPR09
- Search modes Lampert et al CVPR08
- Detection in crowd Ge Collins PETS09
- Detection and tracking in crowd Rodriguez ICCV11
- Survey Dollar et al TPAMI11

Improving speed and accuracy

Multi-Stage Particle Windows for Fast and Accurate Object Detection

[Gualdi, Prati Cucchiara TPAMI11]

form siliding windows to particle window search for people and other targets

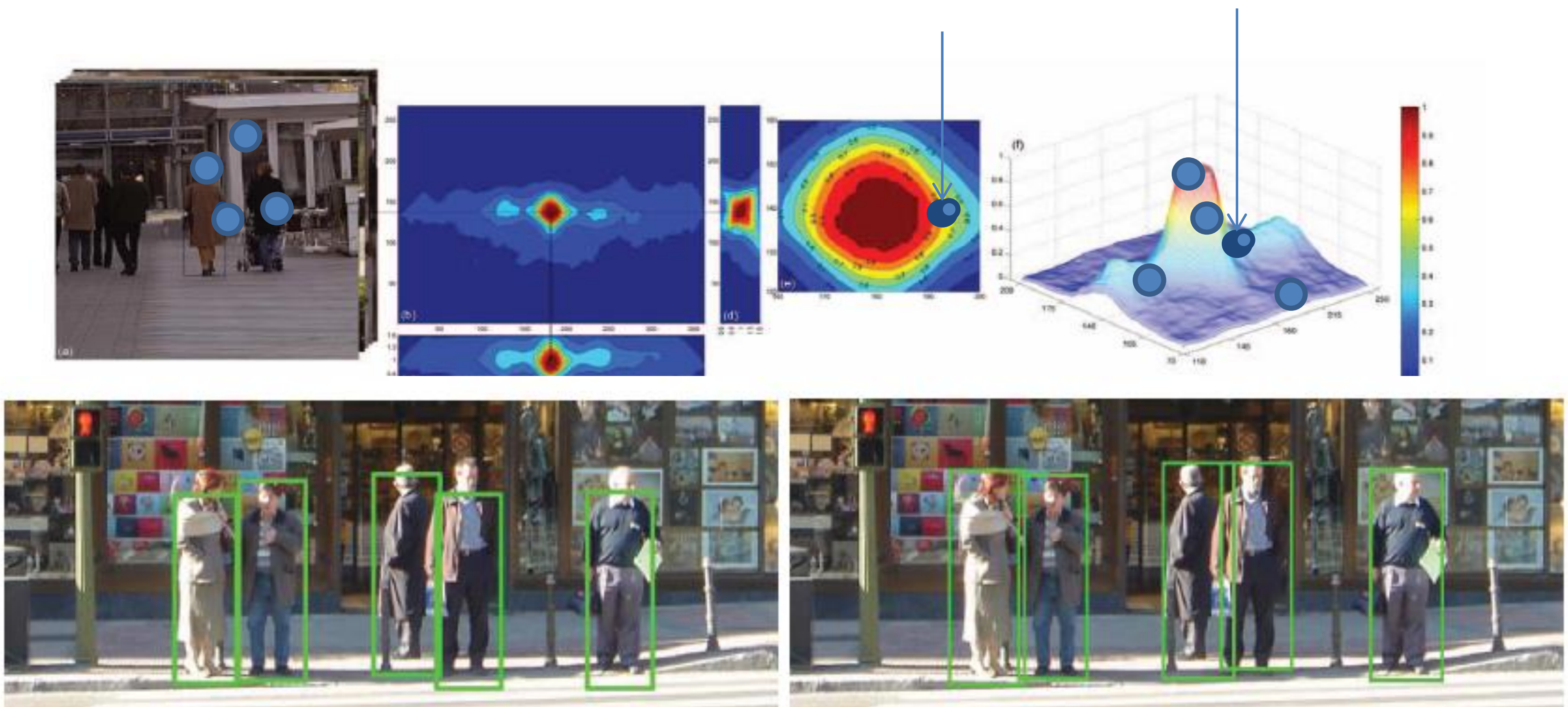
DETECTING PEOPLE

Improving speed and accuracy

Multi-Stage Particle Windows for Fast and Accurate Object Detection

[Gualdi, Prati Cucchiara TPAMI 2011]

from siliding windows to particle window search for people and other targets



(a) (b)

AN EGOCENTRIC VIEW
PEOPLE DETECTION IN AUTOMOTIVE

Complex task with very limited performance in real scenarios

Reasonable automotive dataset (Caltech USA)

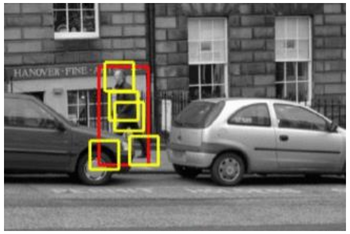
<u>4 Challenges</u>

- **Figure size:**
  Far pedestrians **appear very small in the image**. EG, with VGA resolution and 36deg vertical FOV, the figure of a 1m height child at 30 meters is only 25 pixels long.
- **Fast dynamics:**
  The detection latency must be small, and decisions must be obtained within a few frames.
- **Heavy clutter:**
  Pedestrian detection is typically taking place at urban scenes with a lot of background texture.
- **Articulation:**
  Pedestrians are non-rigid objects, spanning high variability in appearance and cause tracking difficulties.

PEOPLE DETECTION EXPERIMENTS

3 main approaches in literature

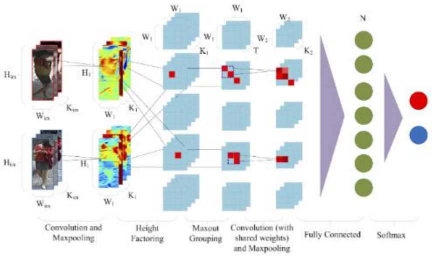
Part Based Models: People body is a collection of part dteected separately al latent variables [Felzenszwalb 2010]



Pros: Robust to occlusions, Accurate <15 % miss rate per-image

Cons: Slow approx 10fps, Need retrain, Need high resolution

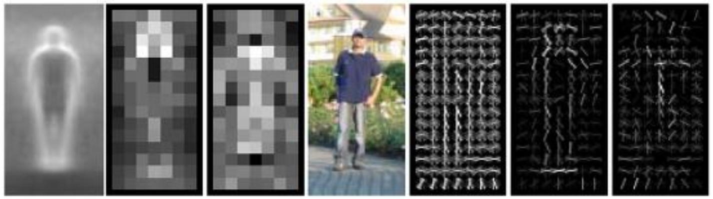
Deep network Models: A conv-net is usuallly trained learning both features and classification functions [Anelia 2015]



Pros: Fast 25fps, Flexibles, Accurate <20% miss rate per-image, easy retrain, hardware implementation

Cons: Scale dependent, Need many data for training, No control over the model, High resolution

Standard feature/classification holistic models [Dollar 2014]



Pros: Scale invariant, Fast approx 30 fps, Features are handcrafted, Flexible and controllable model performances

Cons: Less accurate approx 30% miss rate per-image, must select features, speed depend on features extraction and number of scales

RESEARCH PERFORMANCE ON CALTECH DATASET

Performance Report from [Benenson2014]

No perfect method

- Still impossible to have NO FP at satisfying miss rate
- Still impossible to achieve 0 MR per image
- Performance are evaluated on image independently.

Google requirement for self driving car is a 0,07 sec response of the system

Current fastest method SDN 10fps accuracy 60%

Current most accurate method Katamari accuracy 80% 5fps

<u>GOOGLE latest method DEEPCASCADE[Anelia15]
Accuracy 70% 15 fps</u>

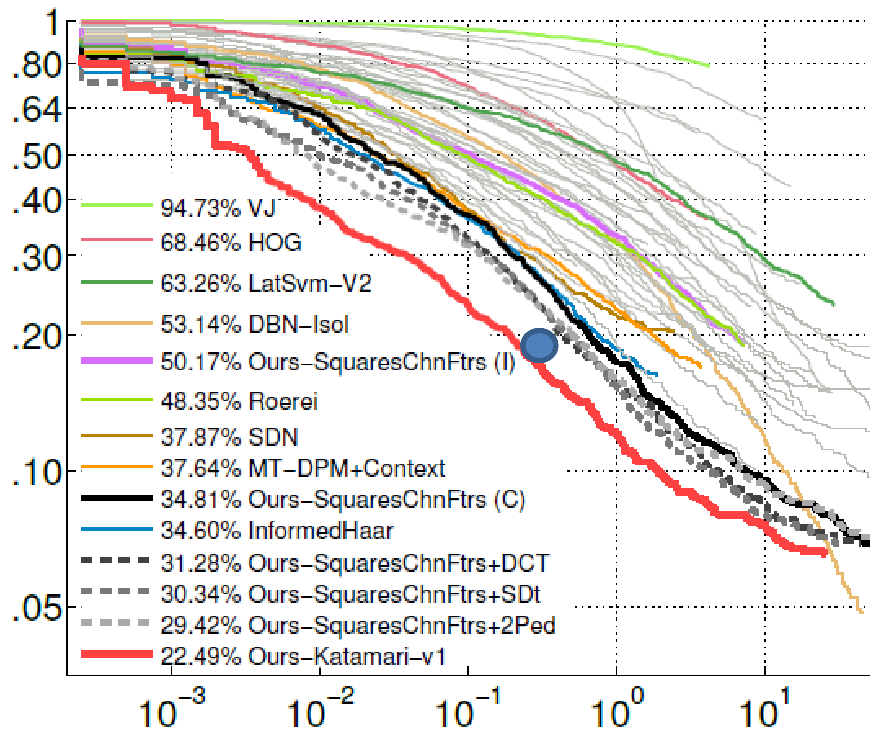
Ten Years of Pedestrian Detection, What Have We Learned?

Rodrigo Benenson   Mohamed Omran   Jan Hosang   Bernt Schiele

Max Planck Institut for Informatics
Saarbrücken, Germany
firstname.lastname@mpi-inf.mpg.de

Eccv 2014

Miss Rate MR



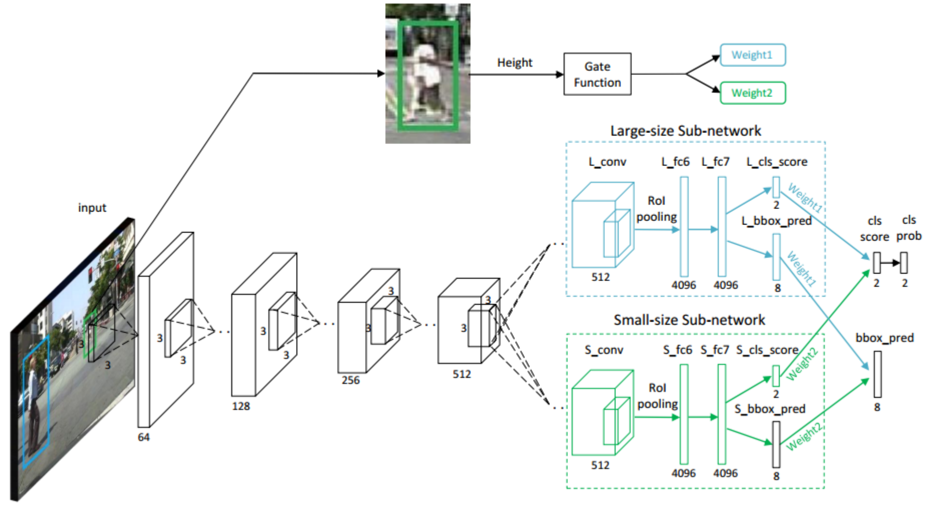
False Positive per image FP

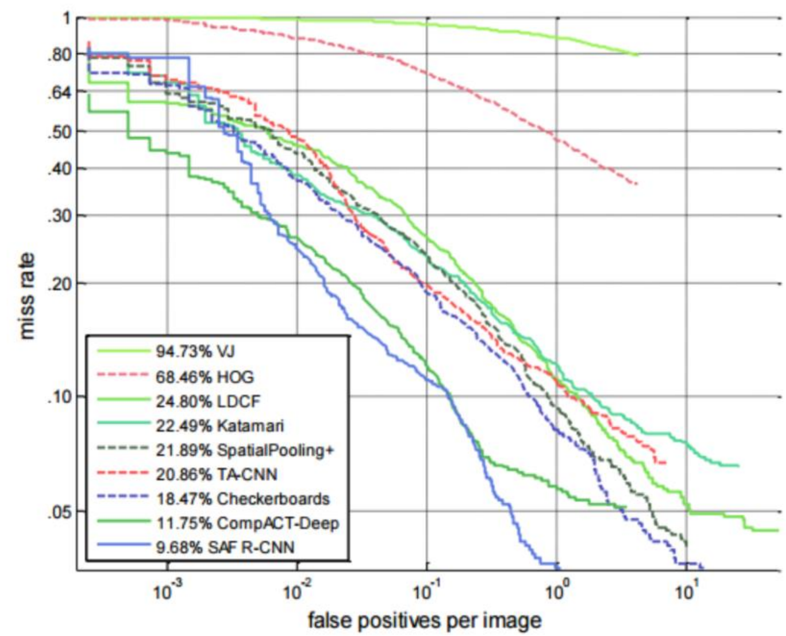
Arxiv 2016

Best now on caltech

Scale-aware Fast R-CNN for Pedestrian Detection

Jianan Li, Xiaodan Liang, ShengMei Shen, Tingfa Xu, and Shuicheng Yan

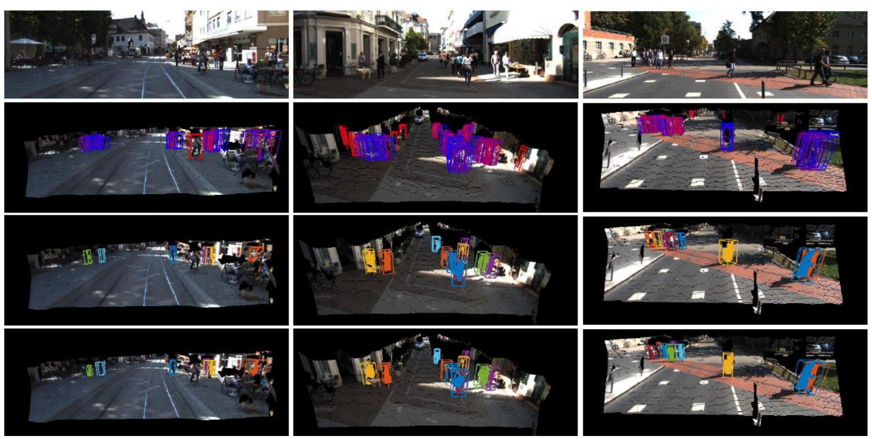




NEW DNN APPROACHES FOR AUTOMOTIVE

Top scorers on Kitty Benchmark:

- Combine CNN and Region proposals
- DO not treat People explicitly. Detect People Cars and Objects simoultaneously -> Exploits generalization capability of CNN filters
- Use 3D when available



3D Object Proposals for Accurate Object Class Detection@ NIPS2015

Exploit All the Layers: Fast and Accurate CNN Object Detector with Scale Dependent Pooling and Cascaded Rejection Classifiers @CVPR16

PEOPLE DETECTION SOLVED?

ICCV 2013

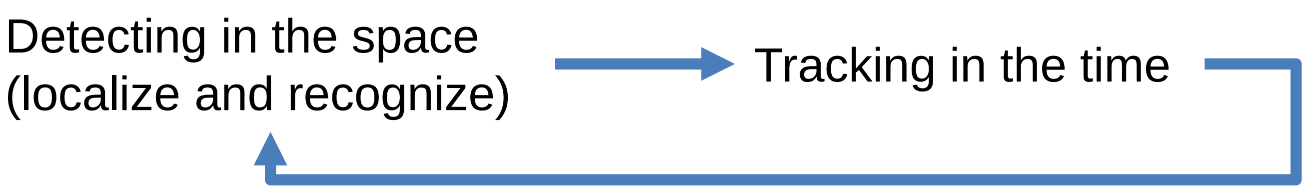
Learning People Detectors for Tracking in Crowded Scenes.
S. Tang, M. Andriluka, A. Milan, K. Schindler, S. Roth, B. Schiele

CVPR 2016

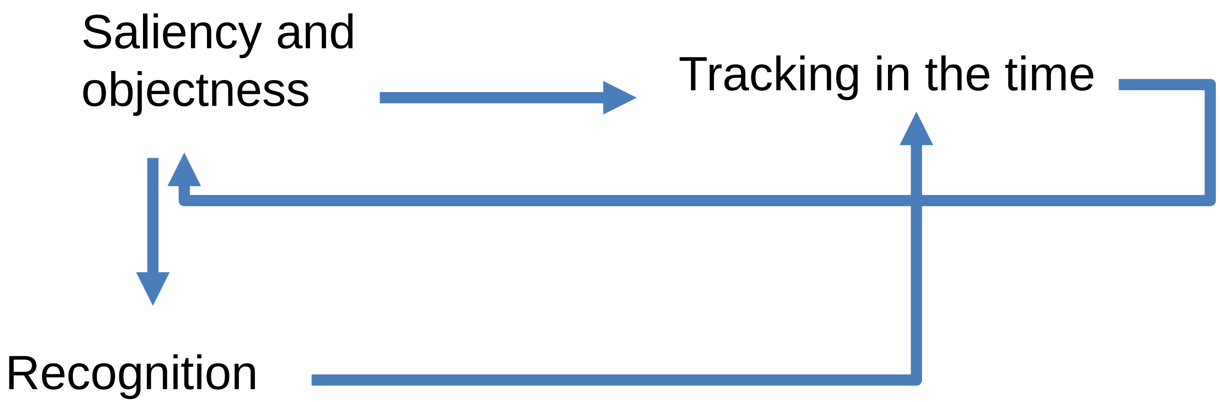
How Far Are We From Solving Pedestrian Detection?.
Shanshan Zhang, Rodrigo Benenson, Mohamed Omran, Jan Hosang, Bernt Schiele CVPR2016

NEXT TRENDS

FROM



TO



OBJECT DISCOVERY AND TRACKING

A co-working approach

Object discovery

And tracking

ICCV2015

Unsupervised Object Discovery and Tracking in Video Collections

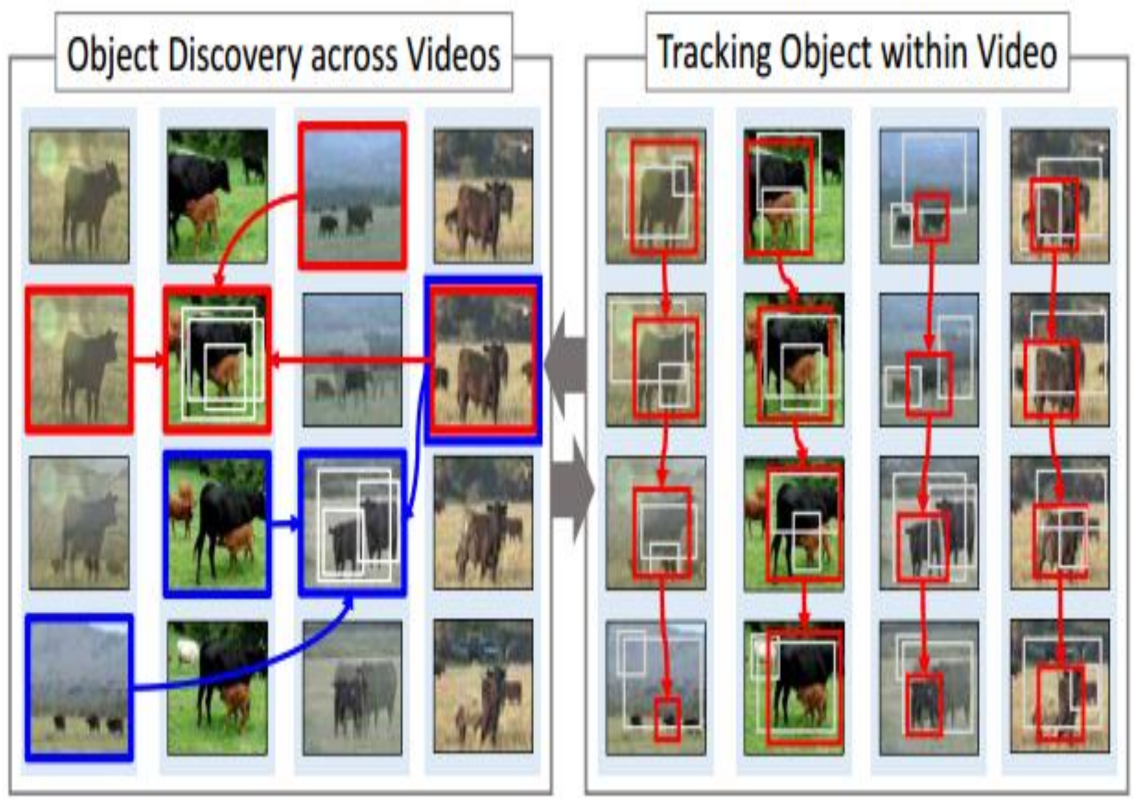
Suha Kwak[1,*] Minsu Cho[1,*] Ivan Laptev[1,*] Jean Ponce[2,*] Cordelia Schmid[1,†]

[1]Inria [2]École Normale Supérieure / PSL Research University

nating in r_{t+1}. A *spatio-temporal tube* is any sequence $r = [r_1, \ldots, r_T]$ of temporal neighbors in the same video. Our goal is to find, for every video v in the input collection, the top tube r according to the criterion

$$\Omega_v(r) = \sum_{t=1}^{T} \varphi[r_t, v_t, N(v_t)] + \lambda \sum_{t=1}^{T-1} \psi(r_t, r_{t+1}), \quad (1)$$

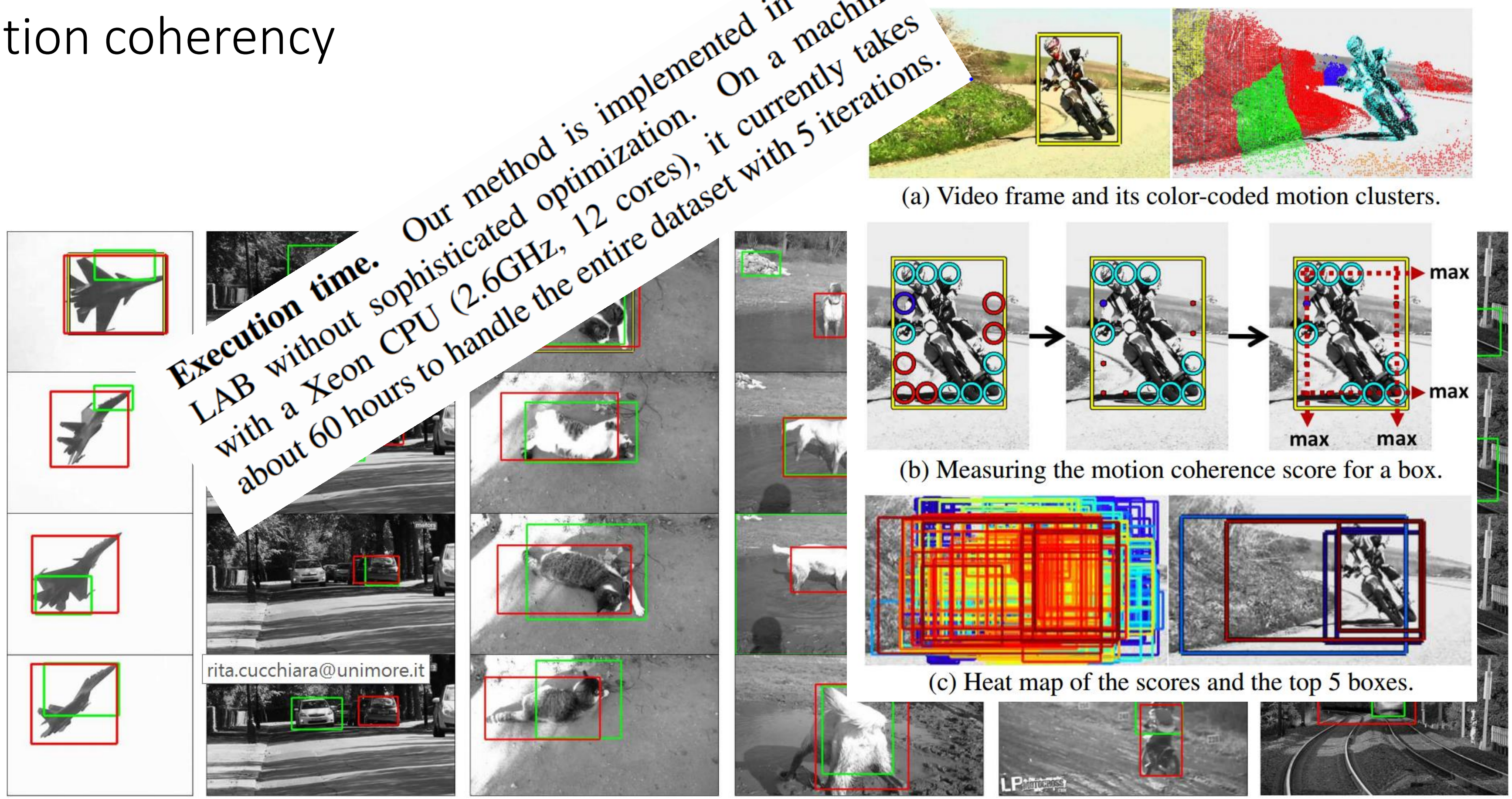
where $\varphi[r_t, v_t, N(v_t)]$ is a measure of confidence for r_t being an object (foreground) region, given v_t and its matching neighbors, and $\psi(r_t, r_{t+1})$ is a measure of temporal consistency between r_t and r_{t+1}; λ is a weight on temporal consistency.



OBJECT DISCOVERY AND TRACKING (CONT)

Motion coherency

Execution time. Our method is implemented in MAT-LAB without sophisticated optimization. On a machine with a Xeon CPU (2.6GHz, 12 cores), it currently takes about 60 hours to handle the entire dataset with 5 iterations.



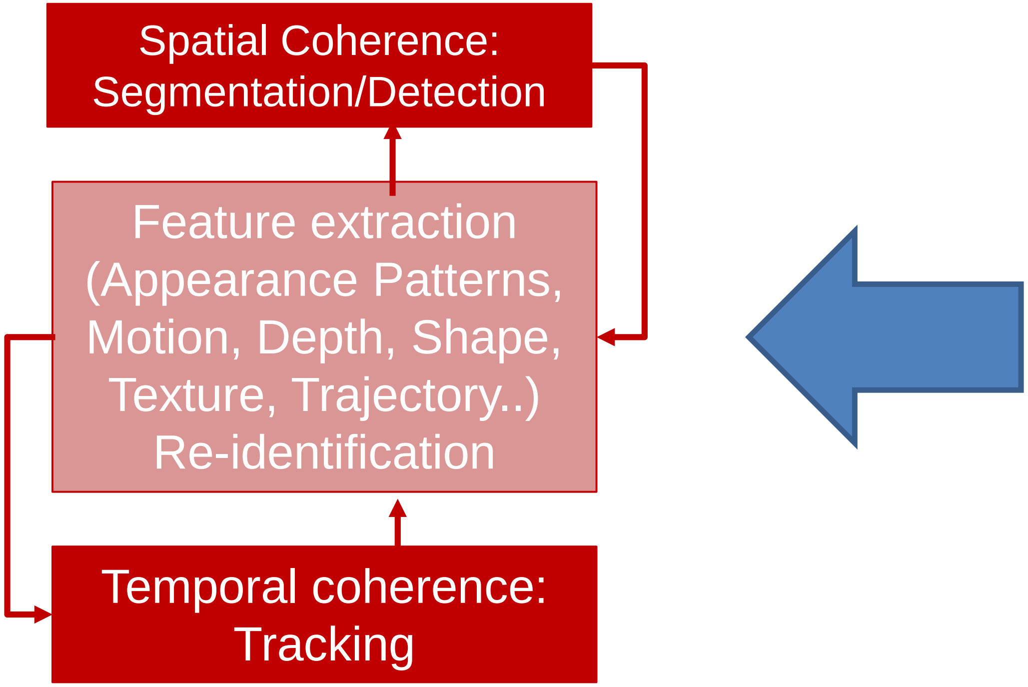
(a) Video frame and its color-coded motion clusters.

(b) Measuring the motion coherence score for a box.

(c) Heat map of the scores and the top 5 boxes.

(a) (b) (c) (d) (e) (f)

Figure 5. Visualization of examples that are correctly localized by our full method: (*red*) our full method, (*green*) our method without motion information, (*yellow*) ground-truth localization. The sequences come from (a) "aeroplane", (b) "car", (c) "cat", (d) "dog", (e) "motorbike", and (f) "train" classes. Frames are ordered by time from top to bottom. The localization results of our full method are spatio-temporally consistent. On the other hand, the simpler version often fails due to pose variations of objects (a, c–f) or produces inconsistent tracks when multiple target objects exist (b). More results are included in the supplementary file. (Best viewed in color.)



4.RE-IDENTIFICATION

PEOPLE RE-IDENTIFICATION

People re-ID two scopes:

1) **Long-time memory**: Search in galleries/whatching list etc: soft-biometry
2) **Short-time memory**: used in multi-target tracking with not overlapped cameras or occlusions or if the frames are not continuous?

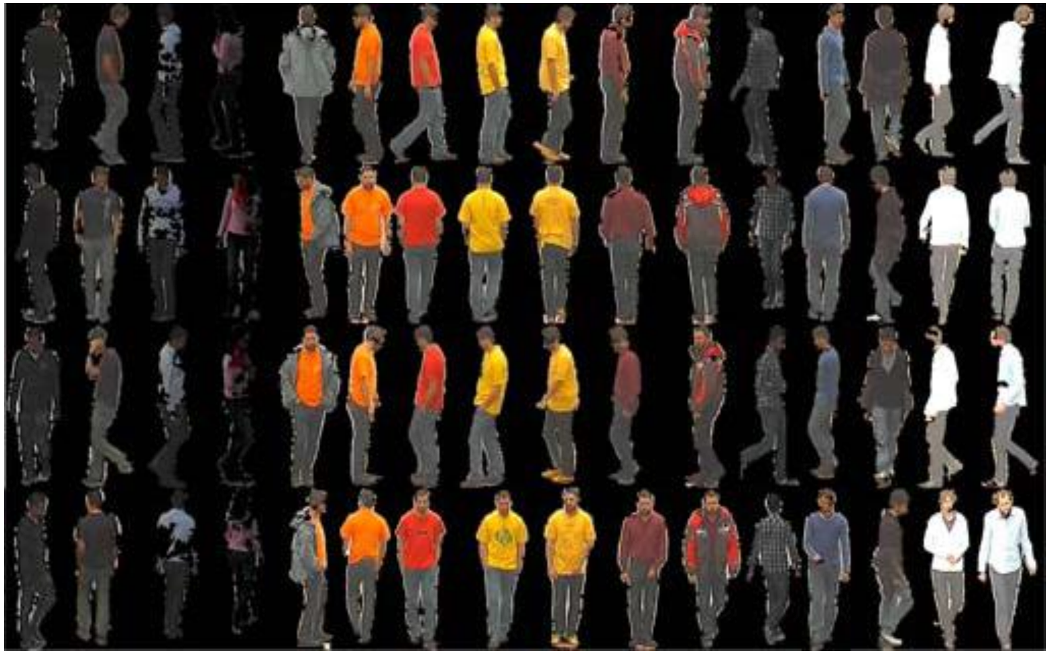
Answer to many questions

Where i've just seen this person?

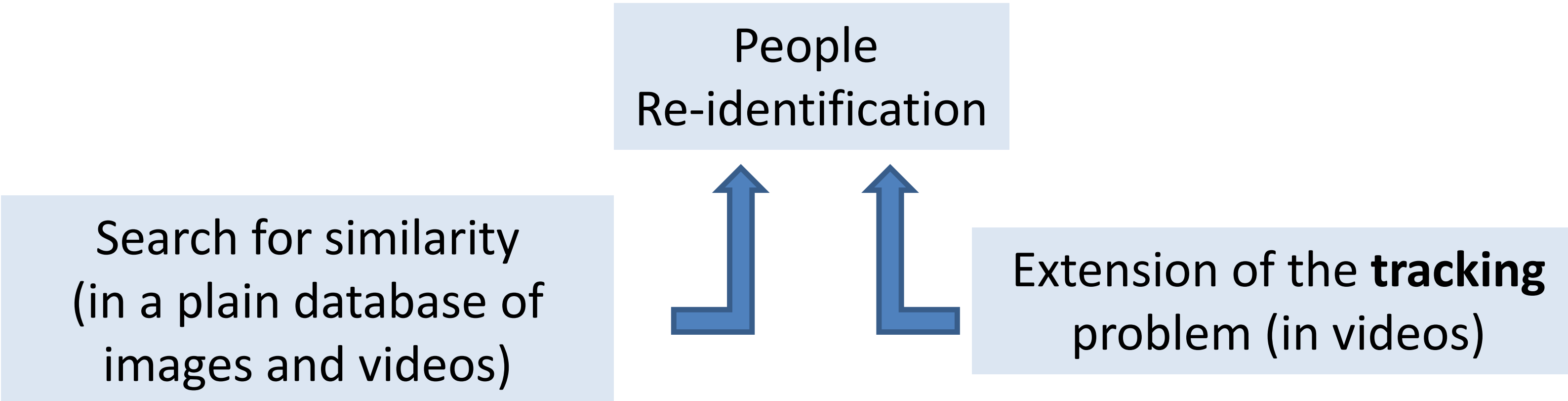
Where is he/she going?

Is this people appeared more time?

[R.Vezzani, D.Baltieri, and R.Cucchiara. People reidentification in surveillance and forensics: A survey. *ACM Comput. Surv.*46, 2, Article 29 (December 2013)]



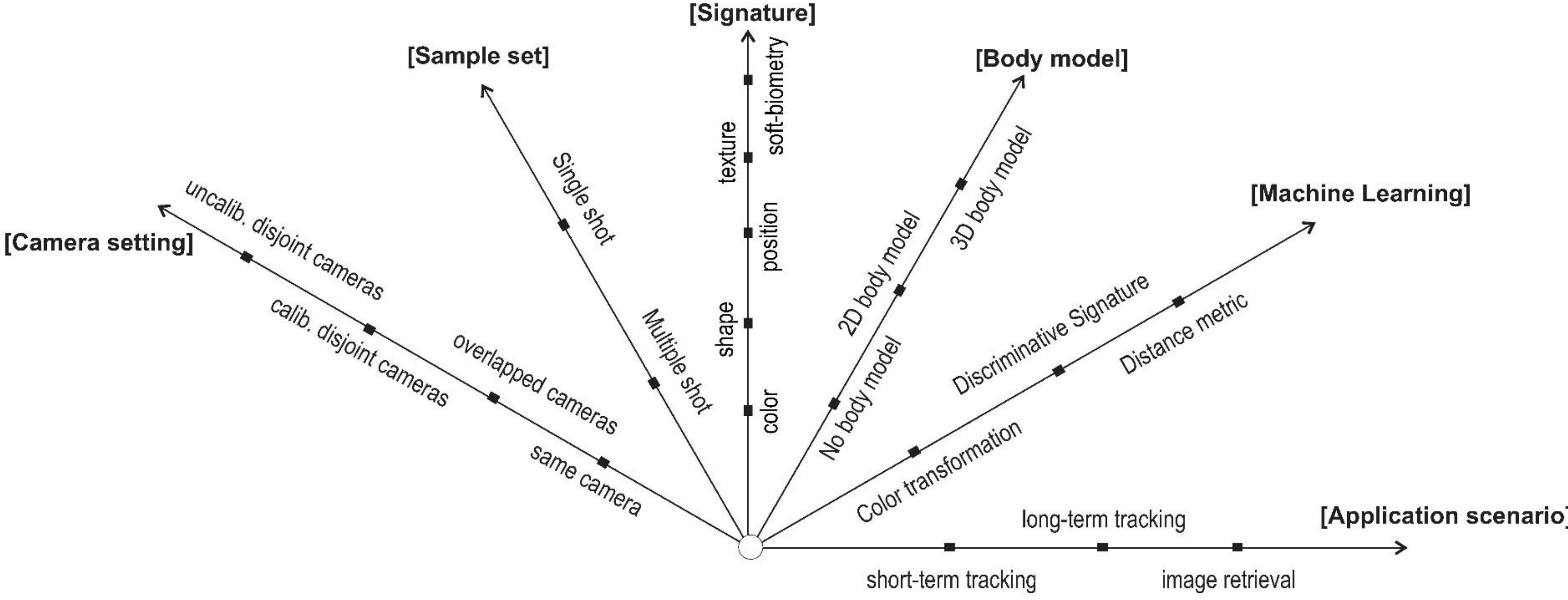
PEOPLE RE-IDENTIFICATION



As a component in the tracking problem,

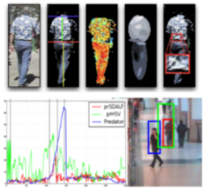
- re-identification aims at finding an **association** between **prediction** and **observation**.
- It supposes that a spatio-temporal coherence of the target position and appearance is satisfied, but there are **some blind spatio-temporal area**.
- it matches a previously seen target if it appears again in the **same camera**, after a **short time**, in a **position close** to the previous one, and with a **similar appearance**.

A MULTI-DIMENSIONAL SPACE



FEATURES

SDALF for person re-identification



Symmetry-driven accumulation of local features for human characterization and re-identification
L. Bazzani, M. Cristani, V. Murino
Computer Vision and Image Understanding (CVIU), 2013.
SDALF code / bibtex

Person re-identification by symmetry-driven accumulation of local features
M. Farenzena, **L. Bazzani**, A. Perina, M. Cristani, V. Murino
In Conference on Computer Vision and Pattern Recognition (CVPR), 2010
SDALF code / video / bibtex

SALF-based features

Now reference method for approaches using color and shape

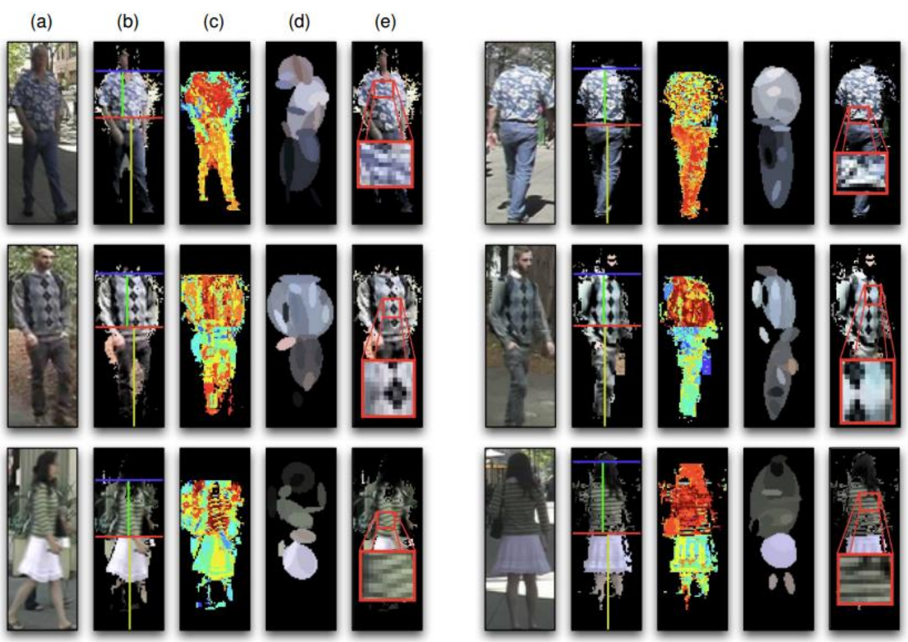
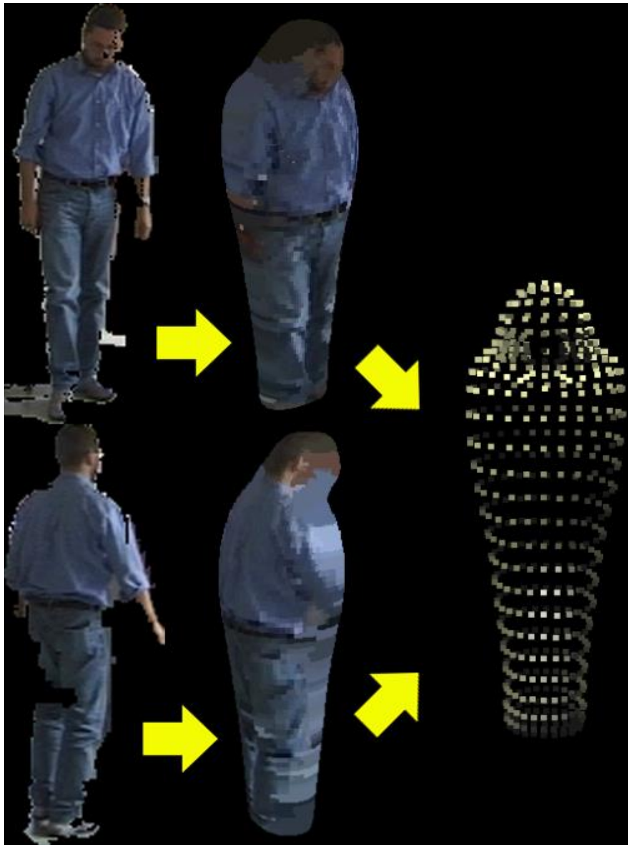
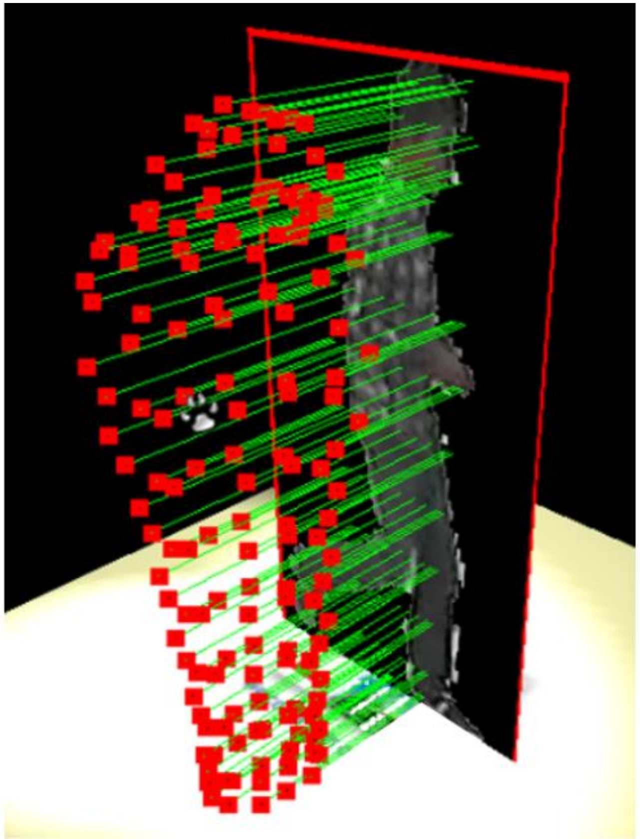


Figure 1: Sketch of the proposed descriptor. (a) Given an image or a set of images, (b) SDALF localizes meaningful body parts. Then, complementary aspects of the human body appearance are extracted: (c) weighted HSV histogram, represented here by its (weighted) back-projection (brighter pixels mean a more important color), (d) Maximally Stable Color Regions [1] and (e) Recurrent Highly Structured Patches. The objective is to correctly match SDALF descriptors of the same person (first column vs. sixth column).

3D RE-IDENTIFICATION

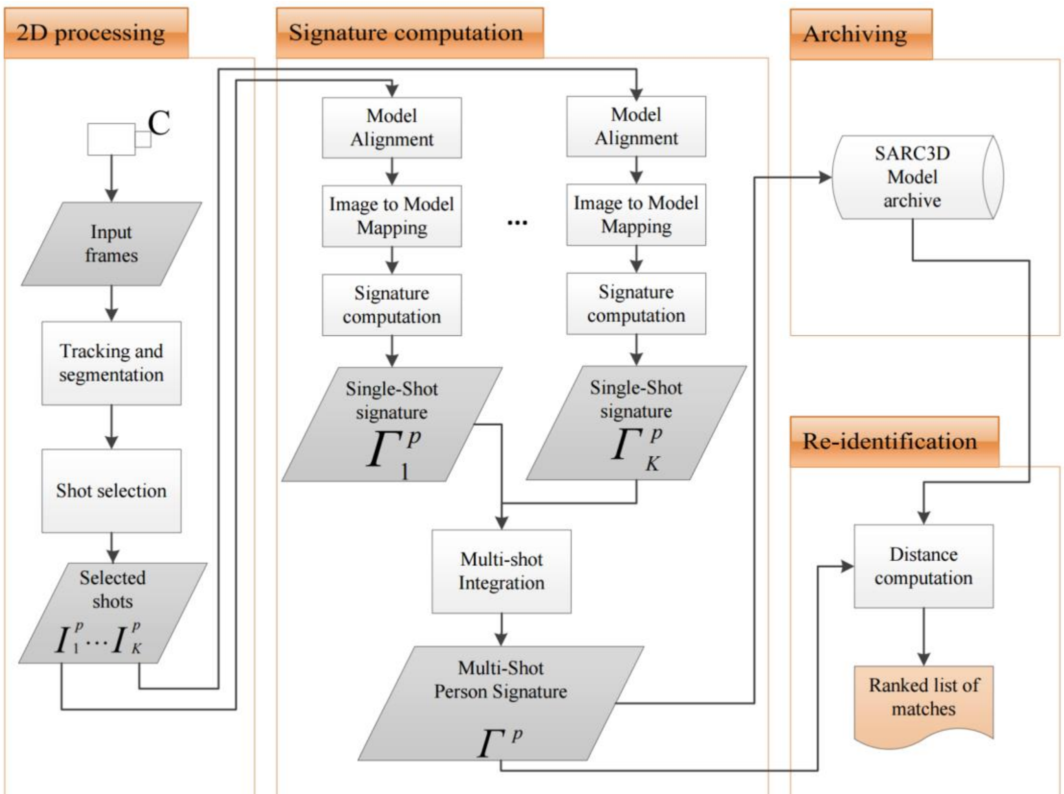
Data driven model: 3D to 3D or 2D to 3D Model match

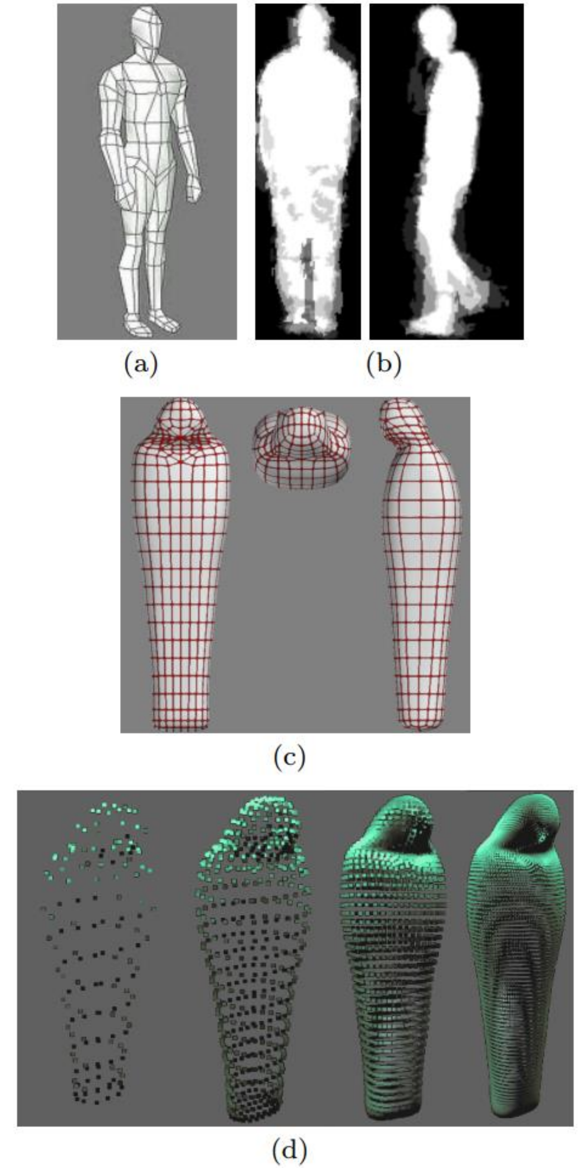




Baltieri, Davide; Vezzani, Roberto; Cucchiara, Rita "Mapping Appearance Descriptors on 3D Body Models for People Re-identification"*International Journal of Computer Vision*, *INTERNATIONAL JOURNAL OF COMPUTER VISION*, vol. 111, pp. 345 -364 , 2014

Fig. 2 Different snapshots of the same pedestrian viewed by a network of cameras, under varying light conditions

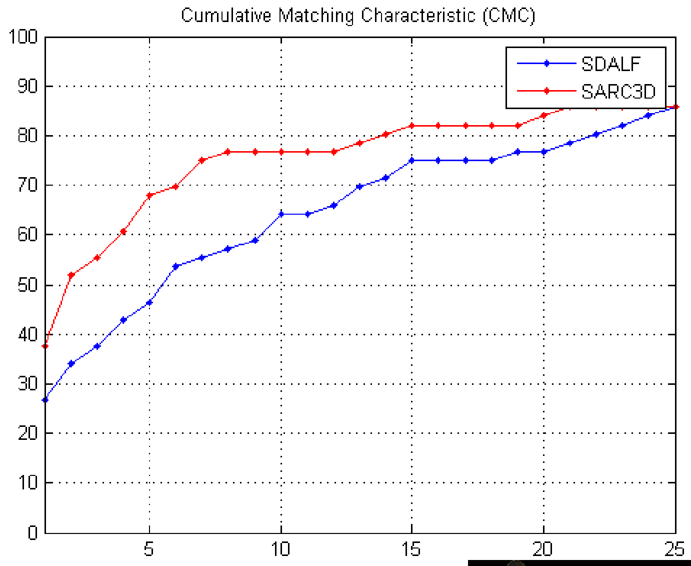




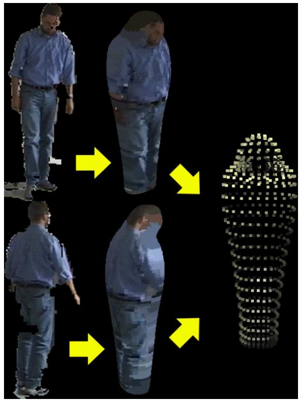
ig. 4 Genesis of SARC3D: (a) a human 3D model, (b) av-
·age silhouettes used for model creation, (c) our simplified
uman model, and (d) different sampling densities of the
ARC3D model used in our tests

RE-IDENTIFICATION: EXAMPLES

Better (only if you have tracking, already)



Models: 2D vs 3D



Davide Baltieri, Roberto Vezzani, Rita Cucchiara, 3DPeS: 3D People Dataset for Surveillance and Forensics
Ws MA3CHO, ACM Multimedia 2011

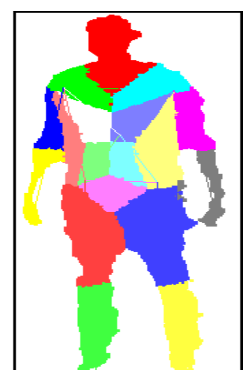
3D IS BETTER

For typically 3D objects with different shape in different views, as persons are.

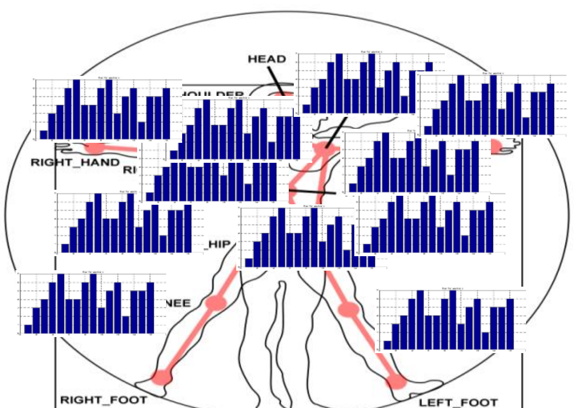
Also with Kinect based or camera based re-identification

3D FEATURES

Using RGB-D sensors



Point cloud partitioning into bone-wise clouds



NEW FEATURES AND APPROACHES

CVPR 2016

Human ID

50 **Recurrent Attention Models for Depth-Based Person Identification.**
Albert Haque, Alexandre Alahi, Li Fei-Fei

51 **Learning a Discriminative Null Space for Person Re-Identification.**
Li Zhang, Tao Xiang, Shaogang Gong

52 **Learning Deep Feature Representations With Domain Guided Dropout for Person Re-Identification.**
Tong Xiao, Hongsheng Li, Wanli Ouyang, Xiaogang Wang

53 **How Far Are We From Solving Pedestrian Detection?.**
Shanshan Zhang, Rodrigo Benenson, Mohamed Omran, Jan Hosang, Bernt Schiele

54 **Similarity Learning With Spatial Constraints for Person Re-Identification.**
Dapeng Chen, Zejian Yuan, Badong Chen, Nanning Zheng

55 **Sample-Specific SVM Learning for Person Re-Identification.**
Ying Zhang, Baohua Li, Huchuan Lu, Atshushi Irie, Xiang Ruan

56 **Joint Learning of Single-Image and Cross-Image Representations for Person Re-Identification.**
Faqiang Wang, Wangmeng Zuo, Liang Lin, David Zhang, Lei Zhang

57 **A Multi-Level Contextual Model For Person Recognition in Photo Albums.**
Haoxiang Li, Jonathan Brandt, Zhe Lin, Xiaohui Shen, Gang Hua

58 **Unsupervised Cross-Dataset Transfer Learning for Person Re-Identification.**
Peixi Peng, Tao Xiang, Yaowei Wang, Massimiliano Pontil, Shaogang Gong, Tiejun Huang, Yonghong Tian

59 **Pedestrian Detection Inspired by Appearance Constancy and Shape Symmetry.**
Jiale Cao, Yanwei Pang, Xuelong Li

60 **Recurrent Convolutional Network for Video-Based Person Re-Identification.**
Niall McLaughlin, Jesus Martinez del Rincon, Paul Miller

61 **Person Re-Identification by Multi-Channel Parts-Based CNN With Improved Triplet Loss Function.**
De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, Nanning Zheng

62 **Top-Push Video-Based Person Re-Identification.**
Jinjie You, Ancong Wu, Xiang Li, Wei-Shi Zheng

63 **Improving Person Re-Identification via Pose-Aware Multi-Shot Matching.**
Yeong-Jun Cho, Kuk-Jin Yoon

64 **Hierarchical Gaussian Descriptor for Person Re-Identification.**
Tetsu Matsukawa, Takahiro Okabe, Einoshin Suzuki, Yoichi Sato

Learning Deep Feature Representations with Domain Guided Dropout for Person Re-identification

Tong Xiao Hongsheng Li Wanli Ouyang Xiaogang Wang
Department of Electronic Engineering, The Chinese University of Hong Kong
{xiaotong,hsli,wlouyang,xgwang}@ee.cuhk.edu.hk

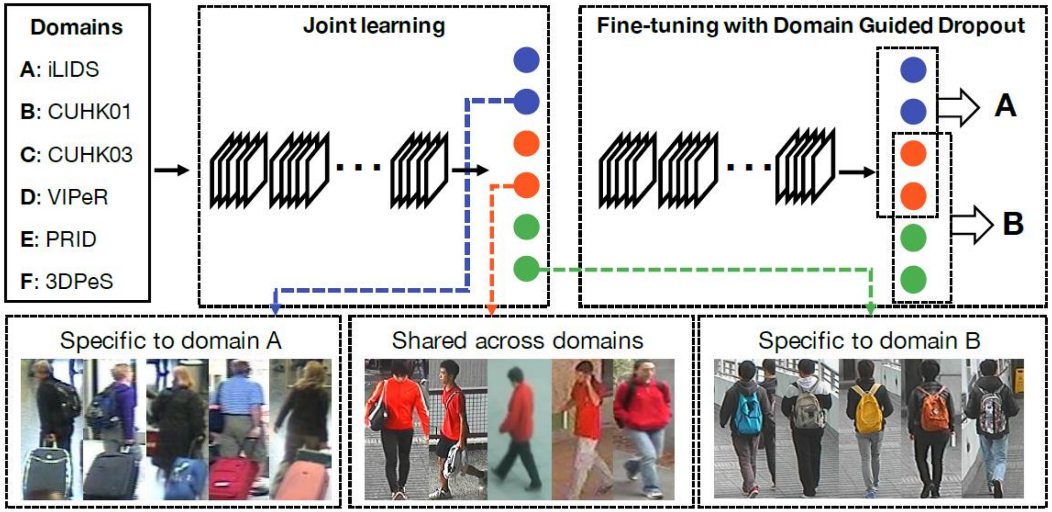


Figure 2. Overview of our pipeline. For the person re-identification problem, we first train a CNN jointly on all six domains. Then we analyze the effectiveness of each neuron on each domain. For example, some may capture the luggages that only appear in domain A, while some others may capture the red clothes shared across different domains. We propose a Domain Guided Dropout algorithm to discard useless neurons for each domain during the training process, which drives the CNN to learn better feature representations on all the domains simultaneously.

name	patch size/ stride	output size	#1×1	#3×3 reduce	#3×3	double #3×3 reduce	double #3×3	pool+proj
input		$3 \times 144 \times 56$						
conv1 – conv3	$3 \times 3/2$	$32 \times 144 \times 56$						
pool3	$2 \times 2/2$	$32 \times 72 \times 28$						
inception (4a)		$256 \times 72 \times 28$	32	32	32	32	32	avg + 32
inception (4b)	stride 2	$384 \times 72 \times 28$	32	32	32	32	32	max + pass through
inception (5a)		$512 \times 36 \times 14$	64	64	64	64	64	avg + 64
inception (5b)	stride 2	$768 \times 36 \times 14$	64	64	64	64	64	max + pass through
inception (6a)		$1024 \times 36 \times 14$	128	128	128	128	128	avg + 128
inception (6b)	stride 2	$1536 \times 36 \times 14$	128	128	128	128	128	max + pass through
fc7		256						
fc8		M						

Table 1. The structure of our proposed CNN for person re-identification

Current dataset used

Dataset	#ID	#Trn. images	#Val. images	#Prb. ID	#Gal. ID
CUHK03 [23]	1467	21012	5252	100	100
CUHK01 [21]	971	1552	388	485	485
PRID [15]	385	2997	749	100	649
VIPeR [13]	632	506	126	316	316
3DPeS [5]	193	420	104	96	96
i-LIDS [50]	119	194	48	60	60
Shinpuhkan [18]	24	18004	4500		

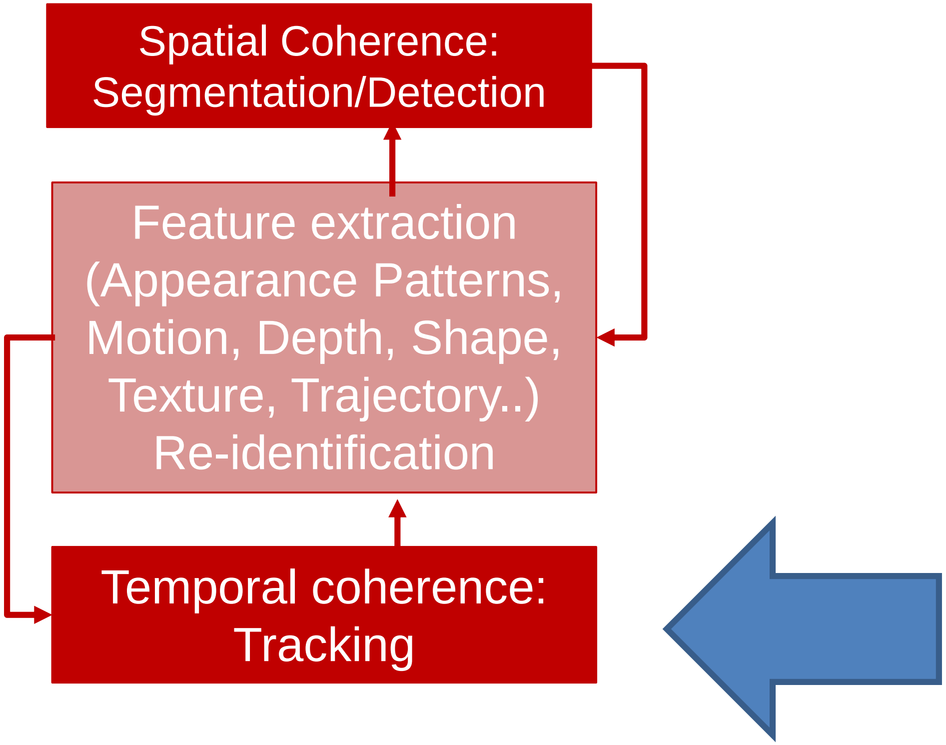
Table 2. Statistics of the datasets and evaluation protocols

Method	CUHK03	CUHK01	PRID
Best	62.1 [32]	53.4 [32]	17.9 [32]
Individually	72.6	34.4	37.0
JSTL	72.0	62.1	59.0
JSTL+DGD	72.5	63.0	60.0
FT-JSTL	74.8	66.2	57.0
FT-JSTL+DGD	**75.3**	**66.6**	**64.0**
Method	VIPeR	3DPeS	iLIDS
Best	**45.9** [32]	54.2 [41]	52.1 [9]
Individually	12.3	31.1	27.5
JSTL	35.4	44.5	56.9
JSTL+DGD	37.7	45.6	59.6
FT-JSTL	37.7	54.0	61.1
FT-JSTL+DGD	38.6	**56.0**	**64.6**

Table 3. CMC top-1 accuracies of different methods

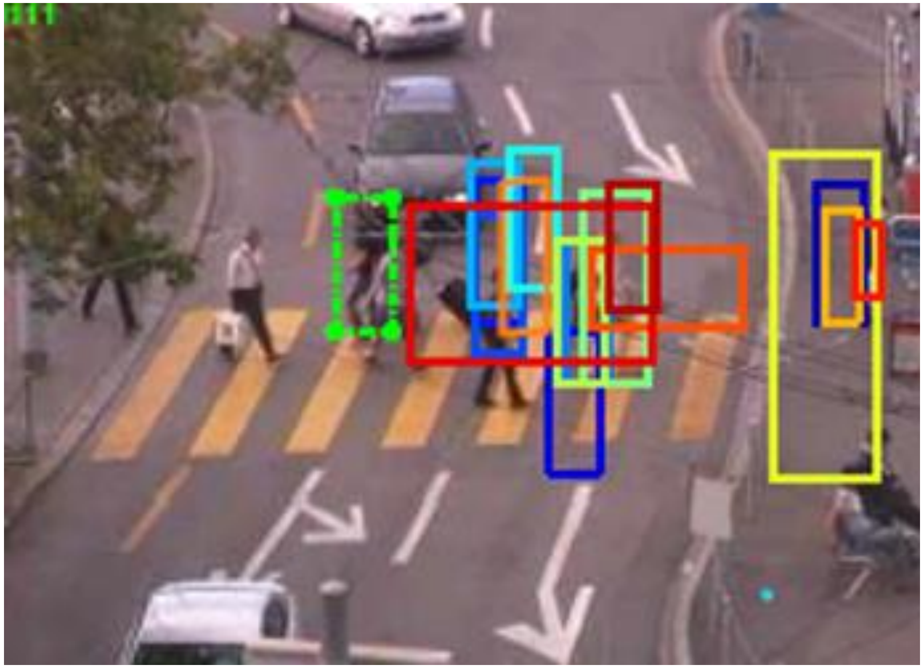
Learning Deep Feature Representations with Domain Guided Dropout for Person Re-identification

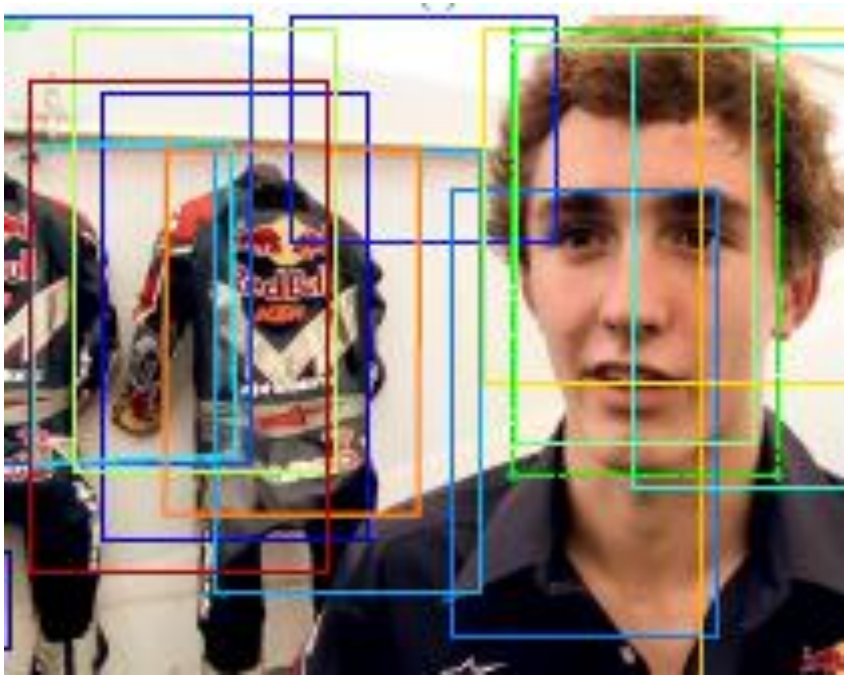
Tong Xiao Hongsheng Li Wanli Ouyang Xiaogang Wang
Department of Electronic Engineering, The Chinese University of Hong Kong
{xiaotong,hsli,wlouyang,xgwang}@ee.cuhk.edu.hk



5.LET'S GO TO TRACKING

ENVIRONMENTAL VS EGOCENTRIC VIEWS





Surveillance

Building automation

...

Eyewear cameras

Automotive

...

AFTER DETECTION... TRACKING!

Environmental

View

Static (multiple) cameras

Large view

Large resolution

Small target size

Crowd situation

Total/partial occlusions

Re-identification in multicamera

Id switch problem

Egocentric

View

Moving (single) camera

Short-distrorted view

Large/small resolution

Large target size

Speed real-time constraints

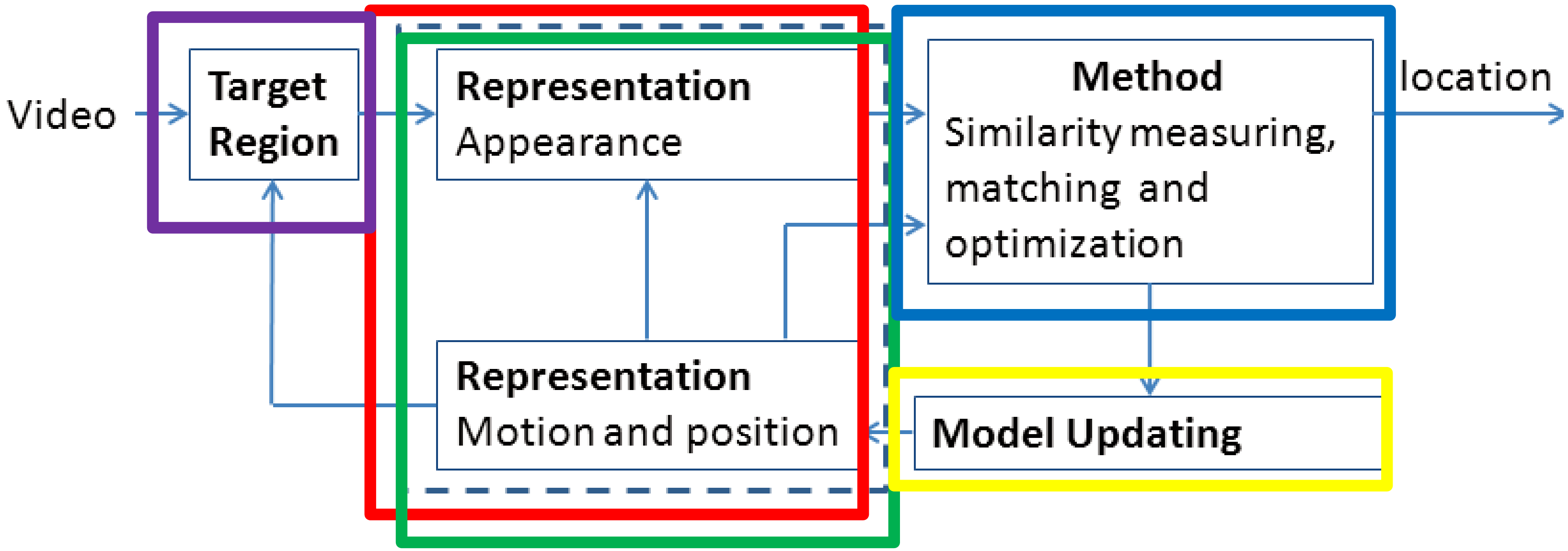
Frequent partial occlusions

Re-identification after occlusions

Frag problem

SINGLE TARGET TRACKING

1. Region of interest
2. Data Representation: how to observe invariant and variant features in the frame and
3. Model Representation how to hold them in an internal representation
4. Inference Method
5. Model Update

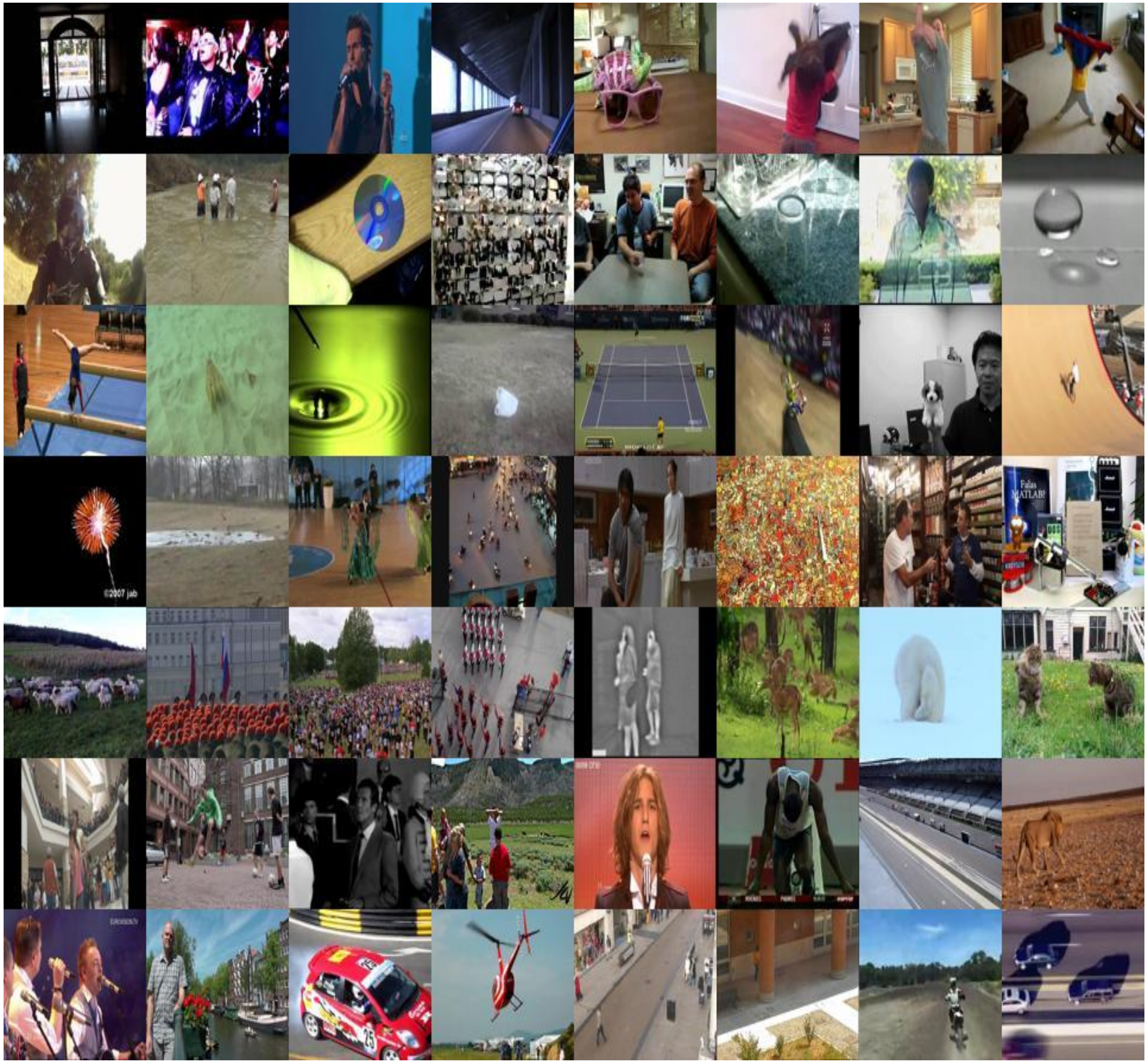


Arnold W. M. Smeulders, Dung M. Chu, Rita Cucchiara, Simone Calderara, Afshin Deghghan and, and Mubarak Shah, Visual Tracking: an Experimental Survey, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014.

ALOV300++ DATASET

http://imagelab.ing.unimore.it/dsm/

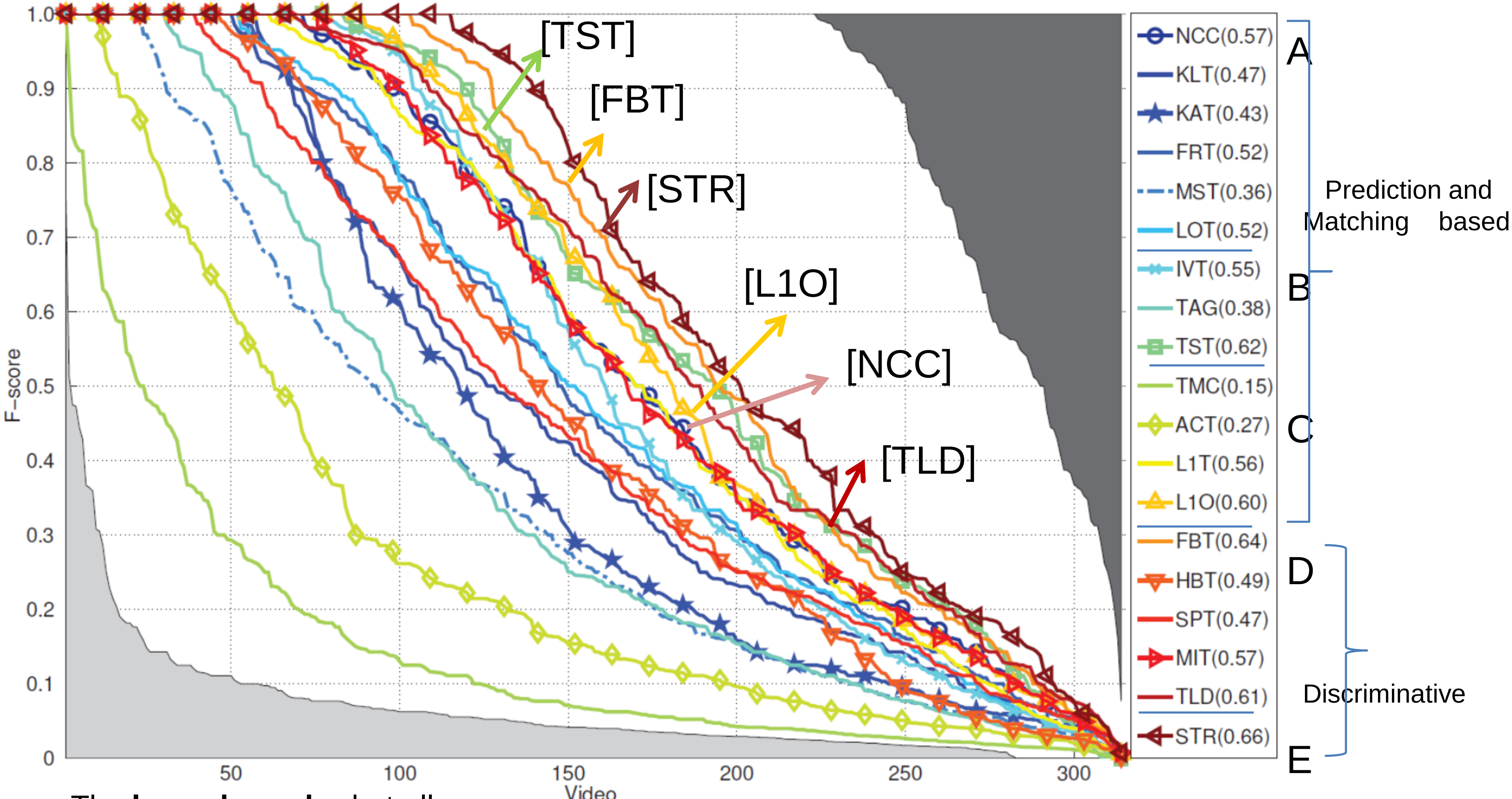
- 01-LIGHT
- 02-SURFACECOVER
- 03-SPECULARITY
- 04-TRANSPARENCY
- 05-SHAPE
- 06-MOTIONSMOOTHNESS
- 07-MOTIONCOHERENCE
- 08-CLUTTER
- 09-CONFUSION
- 10-LOWCONTRAST
- 11-OCCLUSION
- 12-MOVINGCAMERA
- 13-ZOOMINGCAMERA
- 14-LONGDURATION



PERFORMANCE EVALUATION SINGLE TARGET

About the 30%, correctly tracked only

The **upper bound,** taking the best of all trackers at each frame 10%



The **lower bound,** what all trackers can do 7%

Survival curves by Kaplan-Meijer

COMPARED METHODS

A. Tracking by Matching

- **[NCC] Normalized Cross-Correlation**
  K. Briechle and U. Hanebeck, SPIE 2001
- [KLT] Lucas-Kanade Tracker
  S. Baker and I. Matthews, IJCV2004
- [KAT] Kalman Appearance Tracker
  H. Nguyen and A. Smeulders, TPAMI 2004
- [FRT] Fragments-based Robust Tracking
  A. Adam, E. Rivlin, and I. Shimshoni, CVPR2006
  [MST] Mean Shift Tracking
  D. Comaniciu, V. Ramesh, and P. Meer, CVPR2000
- [LOT] Locally Orderless Tracking
  S. Oron, A. Bar-Hillel, D. Levi, S. Avidan, CVPR2012

B. Tracking by Matching with extended model (ST memory)

- [IVT] Incremental Visual Tracking
  D. Ross, J. Lim, and R.S.Lin, IJCV2008
- [TAG] Tracking on the Affine Group
  J. Kwon and F.C. Park, CVPR2009
- [TST] Tracking by Sampling Trackers
  J. Kwon, K.M. Lee, 2ICCV 011

C. Tracking by Matching with constraints

- [TMC] Tracking by Monte Carlo sampling
  J. Kwon, K.M. Lee,CVPR 2009
- [ACT] Adaptive Coupled-layer Tracking
  L. Cehovin, M. Kristan, A. Leonardis, ICCV2011
- [L1T] L1-minimization Tracker
  X. Mei and H. Ling, ICCV2009
- **[L1O] L1 Tracker with Occlusion detection**
  X. Mei, H. Ling, Y. Wu, E. Blasch, L. Bai, CVPR2011

D. Tracking by Discriminant Classification

- [MIT] Multiple Instance learning Tracking
  B. Babenko, M.H. Yang, and S. Belongie, CVPR2009
- **[TLD] Tracking, Learning and Detection**
  Z. Kalal, J. Matas, and K. Mikolajczyk, CVPR2010
- [FBT] Foreground-Background Tracker
  H. Nguyen and A. Smeulders, 2006, IJCV2010
- [HBT] Hough-Based Tracking
  M. Godec, P.M. Roth, H.Bischof, ICCV2011
  [SPT] Super Pixel tracking
  S. Wang, H. Lu, F. Yang, M.H. Yang, ICCV2011

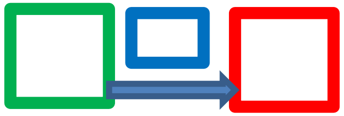
E. Tracking by discriminant Classification with constraints

- [STR] STRuck
  S. Hare, A. Saffari, P. Torr, ICCV2011

INFERENCE METHODS

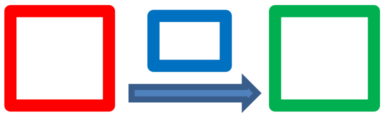
1) Tracking as an inference task (with a statistical model)

- Define the **object model** as the object **status**; that is the appearance and motion representation
- Define the **status evolution** and the status **prediction** (linear, non linear unknown etc) during the time
- Define the data **matching**, i.e. the **measurement** of the prediction against the current data and the status correction

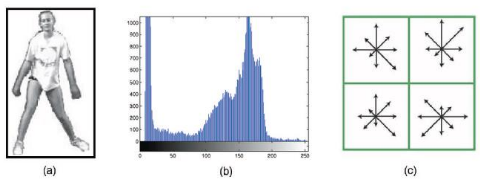


2) Tracking as a model based search and a pattern recognition task

- Define the **object model** (appearance, segmentation, with foreground, texture..) and possibly a non object model
- Define the **search space** (everywhere or according with a prediction)
- Define the **discriminative classifier for the association method** and the memory update



CLASSICAL REPRESENTATION



Appearance representation: (a) 2D-Array ([10]); (b) Histogram; (c) Feature vector.

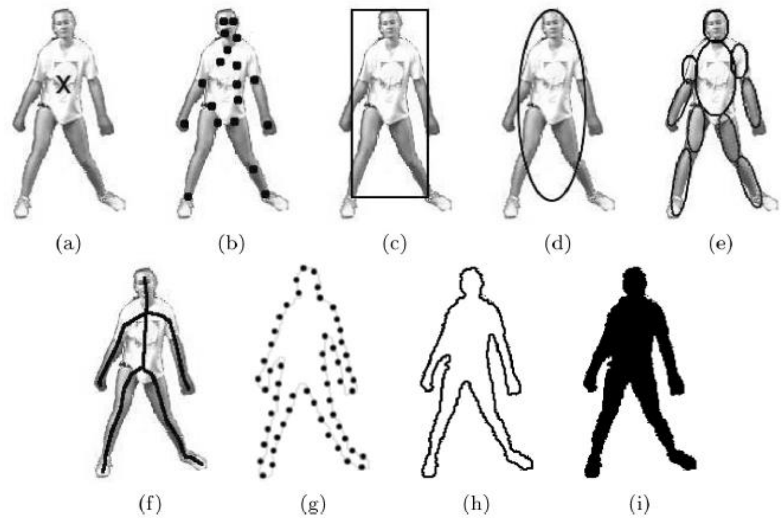


Fig. 1. Object representations. (a) Centroid, (b) multiple points, (c) rectangular patch, (d) elliptical patch, (e) part-based multiple patches, (f) object skeleton, (g) complete object contour, (h) control points on object contour, (i) object silhouette.

Appearance representation
Motion Models

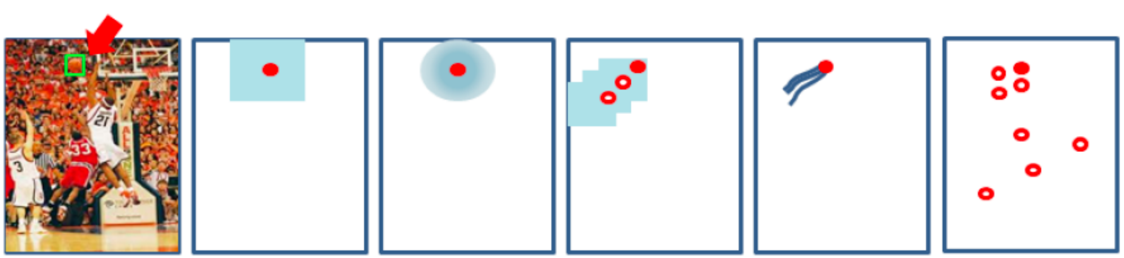
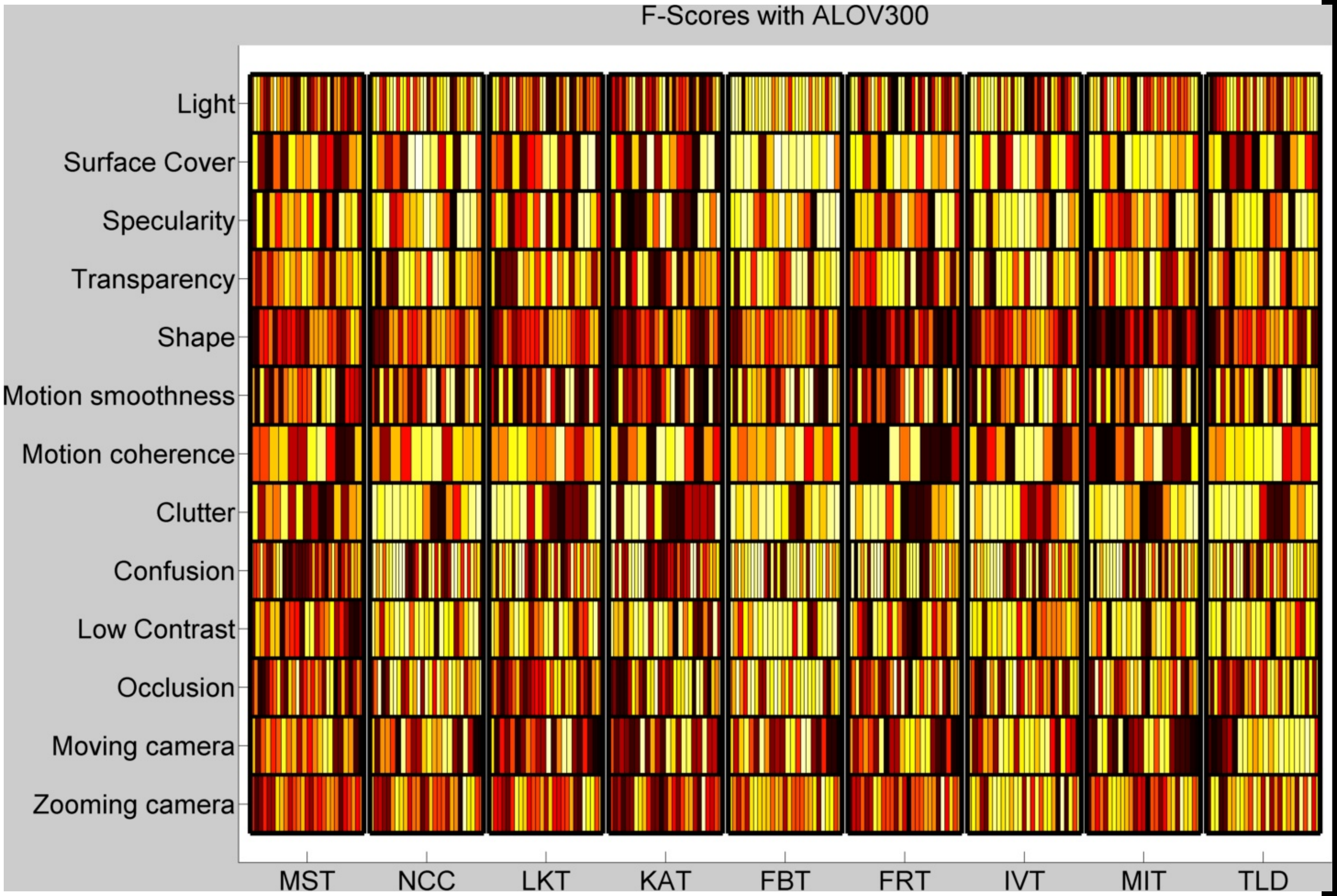


Fig. 4: Motion models used in tracking. From left to right: uniform search, Gaussian motion model, motion prediction, implicit motion model, and tracking and detection.

MOVING CAMERAS AND EGOCENTRIC

TLD is
very robust



EGOCENTRIC VISION

Tracking and re-identification

fast head movement, blur, illumination changes, the trackers results in being extremely short-living.

Egocentric Object Tracking: an Odometry-Based Solution, *Stefano Alletto, Giuseppe Serra, Rita Cucchiara,* International Conference on Image Analysis and Processing, 2015

IMPROVING TLD IN EGOCENTRIC VIEWS

NEW APPROACHES

Motion model with instance proposal CVPR 2016

(a) Frame t (b) Frame $t+1$

Figure 1: **Top row:** Most existing tracking-by-detection methods examine hypothesis locations within a local and heuristically defined search window around the last detected location. **Bottom row:** Our tracker seeks high-quality hypotheses over the entire image using instance-specific edgebox locations.

Beyond Local Search: Tracking Objects Everywhere with Instance-Specific Proposals

Gao Zhu[1], Fatih Porikli[1,2,3], and Hongdong Li[1,3]
Australian National University[1] and NICTA[2]
ARC Centre of Excellence for Robotic Vision[3]
{gao.zhu,fatih.porikli,hongdong.li}@anu.edu.au *

6. ..TO MULTIPLE-TARGET TRACKING…

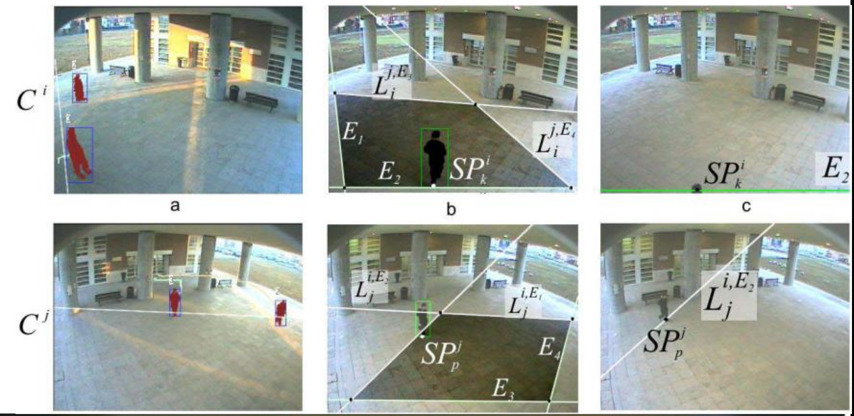
Single target tracking is difficult.

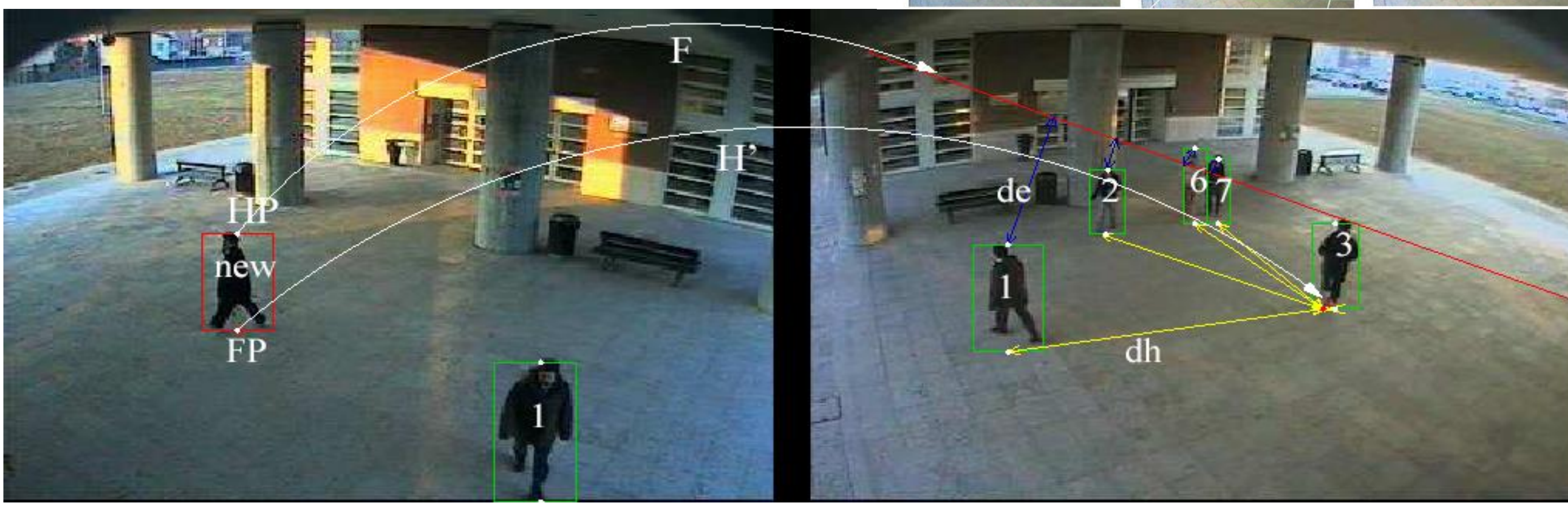
Multi-target tracking is MORE difficult

«CLASSIC APPROACHES» MULTI-SINGLE TARGET TRACKING

Multiple overlapped cameras, multiple target (static cameras)

1. Tracking single objects in each FoV
2. Defining overlapped field of views
3. Using geometric contraints (epipolar lines)
4. Improve with statistical inference

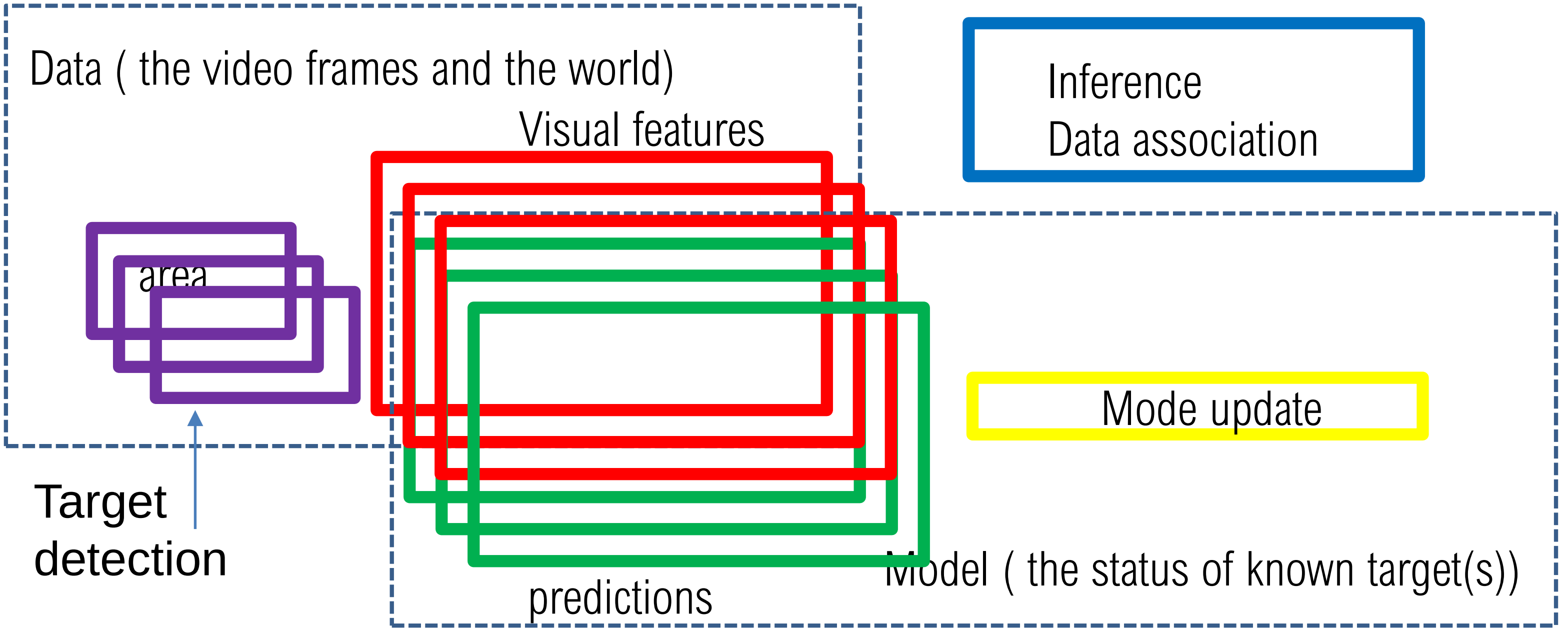




S. Calderara, R. Cucchiara, A. Prati,**"Bayesian-competitive Consistent Labeling for People Surveillance«** in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, n. 2, pp. 354-360, 2008

MULTIPLE SINGLE TARGET TRACKER

MULTIPLE TARGET TRACKING



1) Stocastic MTT
2) Tracking-by-detection: data association
3) SoA methods

MULTI-TARGET TRACKING IS DIFFERENT!

STT: the state of only one target is modelled:

Constraints:

- detections from other targets are assumed to be false alarms
- problems are occlusions

MTT: it takes the existence of more than one target into account simultaneously in their measurement association processes for closely-spaced and crossing targets.

Constraints:

- only one measurement is assumed to be produced by each target at a given time
- the targets are assumed to have independent dynamics.

THE CLASSIC METHOD: STOCHASTIC MTT

If tracking is considered a stochastic prediction of the target state

MTT can be an extension of STT

The «classic» tracking

The target is represented by

- the target state
- $\boldsymbol{x}_k = [x_k, y_k, vx_k, vy_k]^T$ the (motion) state
- $\boldsymbol{xapp}_k = [h_k, w_k, F_k, \ldots.]^T$ the appearance model

- The global state **X** takes into account the union of **x**k

STOCHASTIC MTT

- The state evolves during the time
- The dynamic nature is a process model, an in particular a Hidden Markov process , normally of the first order Markov chain

$$\boldsymbol{x}_{k,t} = \boldsymbol{f}_k(\boldsymbol{x}_{k,t-1}, \boldsymbol{q}_{k,t-1})$$

First order Markov chain (k omitted)

$$p(\mathbf{x}_t|\mathbf{x}_{t-1}) = p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{x}_{t-2}, \ldots, \mathbf{x}_1)$$

The observations

$$\boldsymbol{z}_{k,t} = \boldsymbol{h}_k(\boldsymbol{x}_{k,t}, \boldsymbol{r}_{k,t})$$

It is a possibly non-linear function that translates from the state space to the observation space

The noise sequences qk,t−1 and rk,t are assumed to be mutually independent and identically distributed (i.i.d), and also independent of xk,t and zk,t respectively

STOCHASTIC MTT

Bayesian Model

posterior (what's the model?) $\rightarrow$ **likelihood** (what measurement would we expect to see if we knew the model?) **prior** (our knowledge about these parameters)

$$p(\mathbf{x}|\mathbf{z}) = \frac{p(\mathbf{z}|\mathbf{x})\, p(\mathbf{x})}{p(\mathbf{z})}$$

normalization

$$p(\mathbf{z}) = \int_{\mathbf{x}} p(\mathbf{z}|\mathbf{x})\ p(\mathbf{x}) d\mathbf{x}$$

Treat tracking problem as a first order Markov process

- Estimate $p(x_t \mid z_t, x_{t-1})$
- Combine Markov assumption with Bayes Rule

$$p(\mathbf{x}_t|\mathbf{z}_t, \mathbf{x}_{t-1}) \propto p(\mathbf{z}_t|\mathbf{x}_t)\ p(\mathbf{x}_t|\mathbf{x}_{t-1})$$

Inference is given by ***prediction*** *and* ***matching***

measurement likelihood (likelihood of seeing this measurement)

prediction (based on previous frame and motion model)

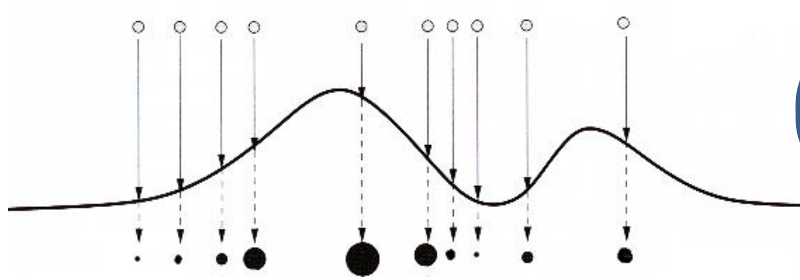
MONTECARLO

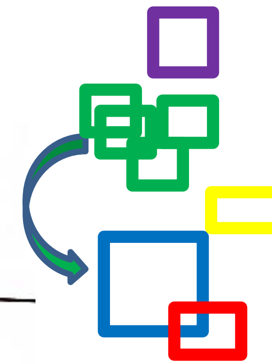
$$p\left(\mathbf{x}_k \middle| \mathbf{Z}_k\right) = \kappa\, p\left(\mathbf{z}_k \middle| \mathbf{x}_k\right) \int p\left(\mathbf{x}_k \middle| \mathbf{x}_{k-1}\right) p\left(\mathbf{x}_{k-1} \middle| \mathbf{Z}_{k-1}\right) d\mathbf{x}_{k-1}$$

Monte Carlo characterization of pdf:

- Represent posterior density by a set of random i.i.d. samples (**particles**) from the pdf $p(x_{0:t}|z_{1:t})$ (the priori in the previous frame)

- For larger number *N* of particles equivalent to functional description of pdf; For $N \rightarrow \infty$ approaches optimal Bayesian estimate





Regions of high density

- **Many particles**
- **Large weight of particles**

Uneven partitioning

Discrete approximation for continuous pdf

$$P_N(x_{0:t} \mid z_{1:t}) = \sum_{i=1}^{N} w_t^i \, \delta(x_{0:t} - x_{0:t}^i)$$

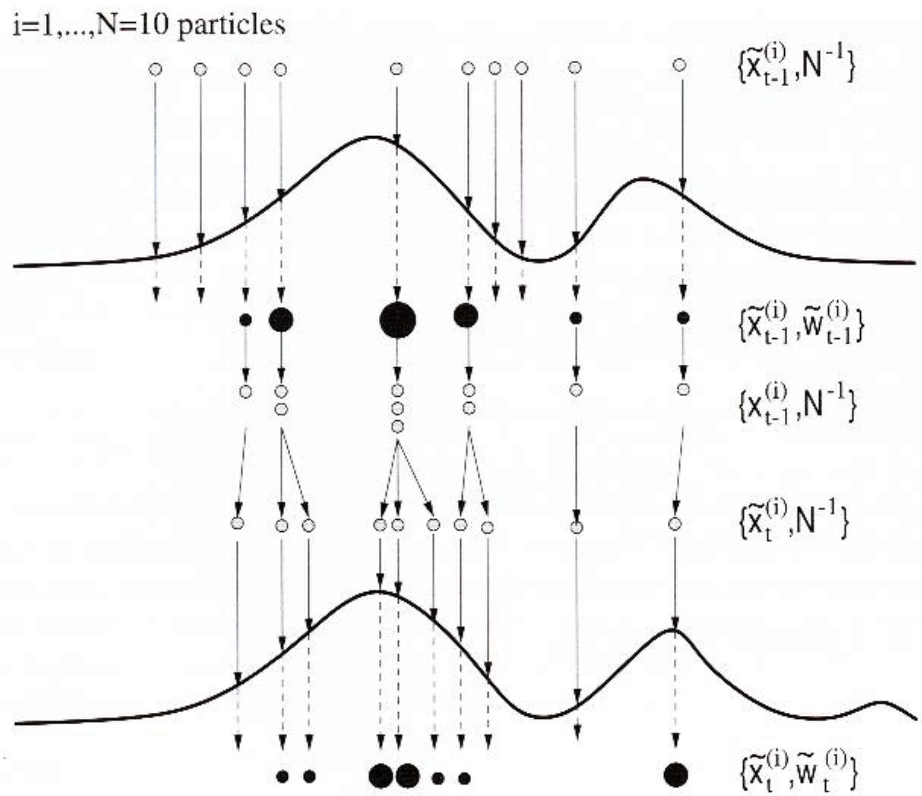
unweighted measure

compute importance weights $\Rightarrow p(x_{t-1} | z_{1:t-1})$

resampling

move particles

predict $p(x_t | z_{1:t-1})$



BRAMBLE

Extended to MTT

(Bramble , Bayesian Multiple-BLob Tracker, Misard, J MacCormick **2003**)

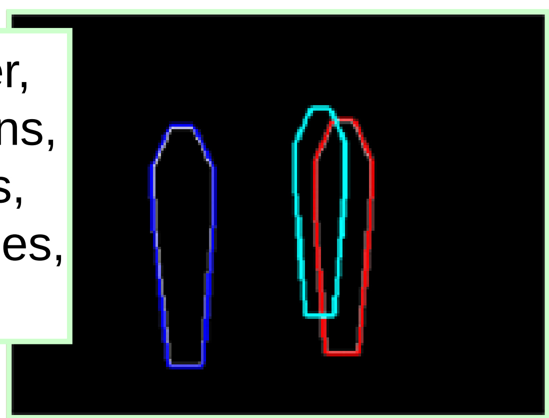
The state is the state of MULTIPLE TARGETS

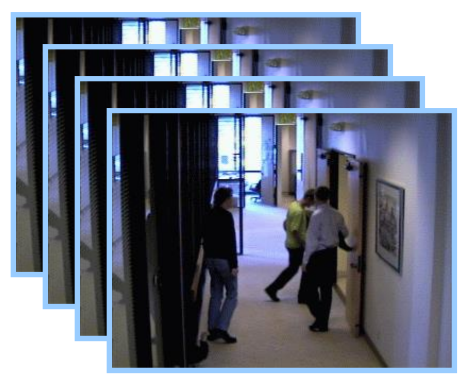
$$p(\mathbf{X}_t \mid \mathbf{Z}_t, \mathbf{Z}_{t-1}, \ldots, \mathbf{Z}_1)$$

State at frame t

Image Sequence

Number,
Positions,
Shapes,
Velocities,
…



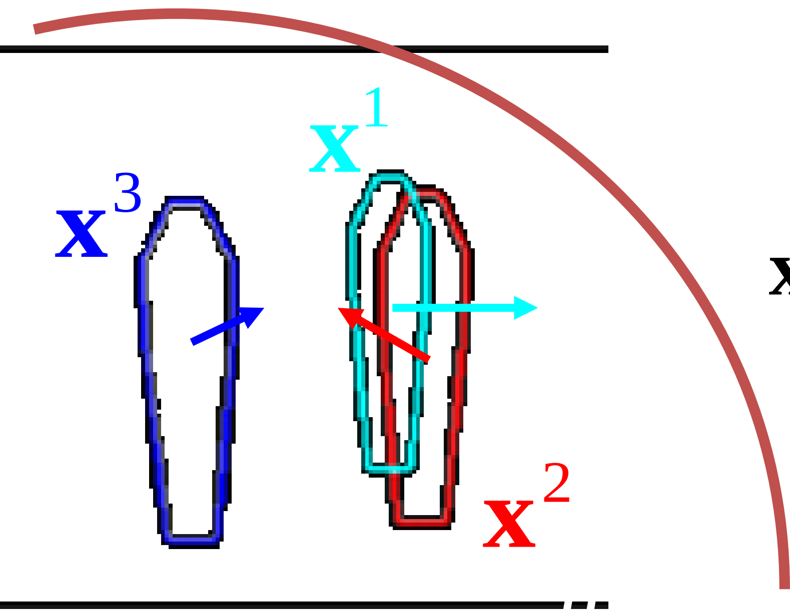


MTT BRAMBLE

Observations are conditionally independent (but is it true??)

$$p(\mathbf{Z} \mid \mathbf{X}) = \prod_k p(z_k \mid \mathbf{X})$$

The state is a single state with a fixed number of target



$$\mathbf{X} = \left(m, \quad \mathbf{x}^1, \quad \ldots, \quad \mathbf{x}^m\right)$$

$$\mathbf{x} = \left(i, x, y, vx, vy, shape\right)$$

PARTICLE FILTERING ON MTT

Bramble approach

$$p(\mathbf{X}_t \mid \mathbf{Z}_{1:t})$$
$$\{(\mathbf{X}_t^i, \pi_t^i)\}$$

N Points: $\mathbf{X}_t^1$ $\mathbf{X}_t^2$ ... $\mathbf{X}_t^{N-1}$ $\mathbf{X}_t^N$

N Weights: π_t^1 π_t^2 π_t^{N-1} π_t^N

BUT Stocastic models with a unique status are suitable for few targets

- Fixed number of target

- Many parameters, - Without detection, large possibility of drift

MULTIPLE SINGLE TARGET TRACKERS **VS** MULTI-TARGET TRACKER

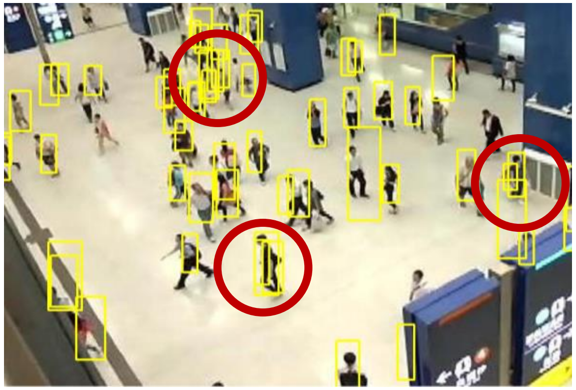
Multiple STT	MTT
Need a parallel implementation	Depends on the detector performances…
Keep a model for each pedestrian	Will get better as detectors will improve!
Terribly slow	Rarely seen online
Online	Association and optimization methods work really well

MULTIPLE SINGLE TARGET TRACKERS **VS** MULTI-TARGET TRACKER

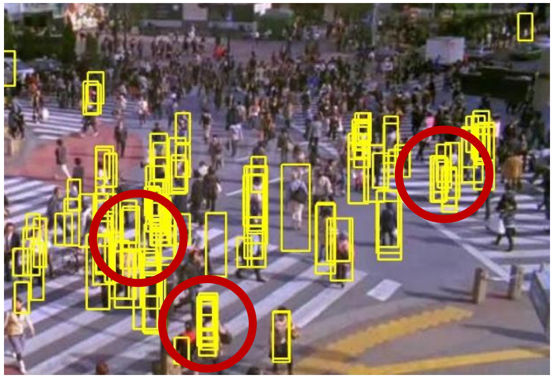
Multiple STT	MTT
is **LOCAL**	is **GLOBAL**

and after only a few frames

the single trackers concentrates around the most responsive pedestrians



1dawei1 (~60 ped.)



1japancross2 (~120 ped.)

MULTIPLE SINGLE TARGET TRACKERS **VS** MULTI-TARGET TRACKER

QUANTITATIVE EVALUATION (MOTA/MOTP)

	1dawei1	1japancross2	
CEM* 2014	96% / 0.25m	79% / 0.45m	MTT
CMPT*	94% / 0.15m	82% / 0.50m	MTT
TLD**	**68% / 0.40m**	**59% / 0.60m**	MSTT
STRUCT**	**44% / 0.70m**	**29% / 1.20m**	MSTT

The multiple single trackers rapidly drift as the crowd moves.
Many pedestrians are still tracked as the crowd often moves homogeneously, so nearby pedestrians help in keeping the bounding box near the target – **but what are we really tracking then?**

* we used Dollars' detector which, appropriately trained, yielded an error of about 20%
P. Dollár, R. Appel, S. Belongie and P. Perona
Fast Feature Pyramids for Object Detection, PAMI 2014
** initialization was done manually for each pedestrian

MTT BY DETECTION:
A DATA ASSOCIATION PROBLEM

MULTIPLE TARGET TRACKING

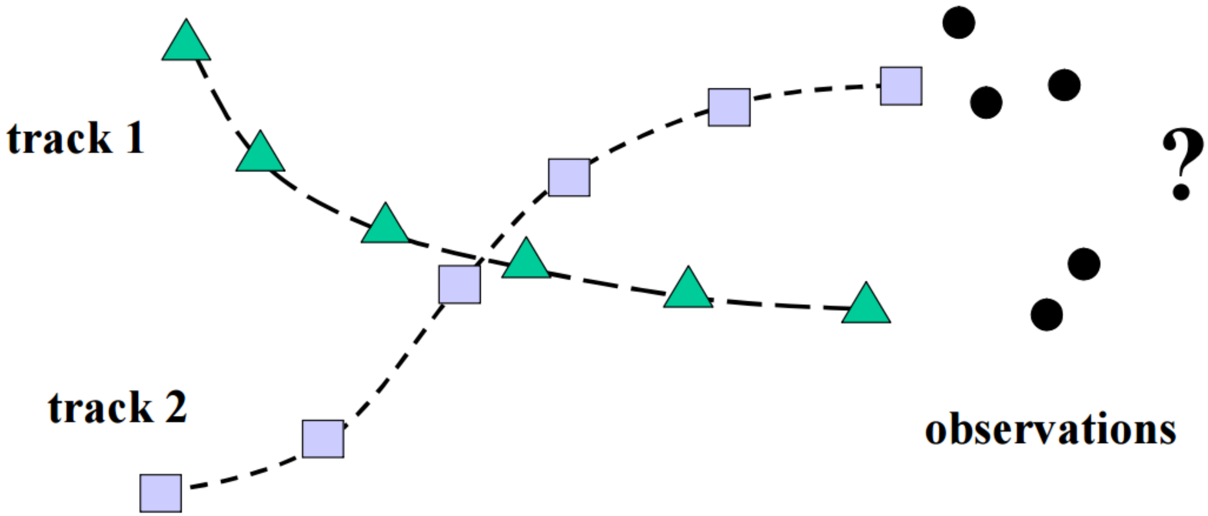
Most cited MT-Trackers published before 2014

- Discrete-Continuous Optimization for Multi-Target Tracking
  Anton Andriyenko, Konrad Schindler and Stefan Roth, 2012
  http://www.gris.informatik.tu-darmstadt.de/~aandriye/dctracking.html
- Global Multi-object Tracking Using Generalized Minimum Clique Graphs
  Amir Roshan Zamir, Afshin Dehghan and Mubarak Shah, 2012
  http://crcv.ucf.edu/projects/GMCP-Tracker/
- Continuous Energy Minimization for Multi-Target Tracking CEM
  Anton Andriyenko and Konrad Schindler, 2014
  http://www.gris.informatik.tu-darmstadt.de/~aandriye/contracking.html
- Multiple Object Tracking using K-Shortest Paths Optimization
  J. Berclaz, F. Fleuret, E. Türetken and P. Fua, 2011
  http://cvlab.epfl.ch/software/ksp
- Continuous Energy Minimization for Multi-Target Tracking
  A. Milan, S. Roth and K. Schindler , TPAMI 36(1), 2014

What do they all have in common?

They are data association techniques that work on already detected pedestrians.

IT'S A DATA ASSOCIATION PROBLEM



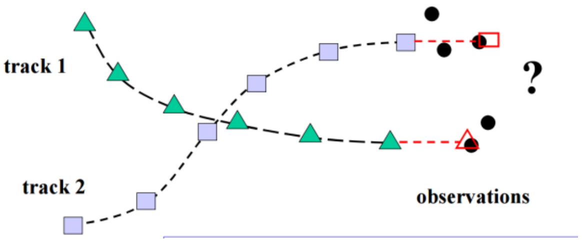
How to determine which observations to add to which track?

Courtesy of B.Collins

FILTERING, GATING AND ASSOCIATION

1) FILTERING

- Prediction: propagate state pdf forward in time, taking process noise into account (translate, deform, and spread the pdf)

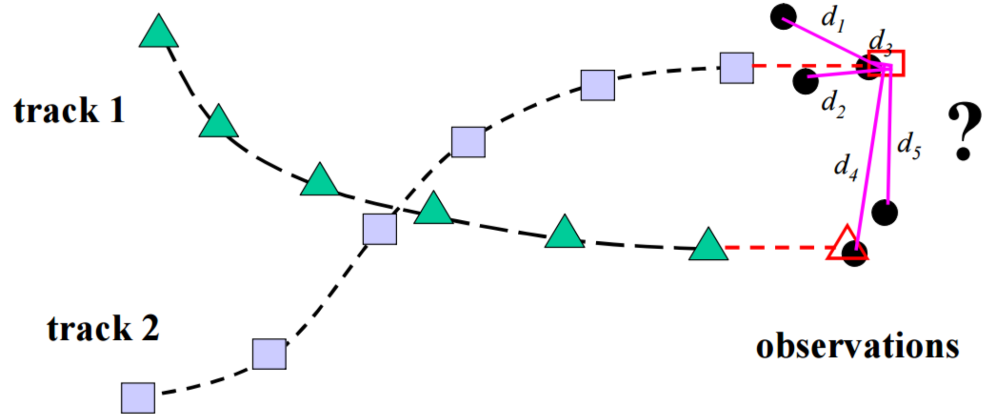


2) GATING AND ASSOCIATION

- Gating to determine possible matching observations
- Data association to determine best match
- Update: use Bayes theorem to modify prediction pdf based on current measurement

GATING

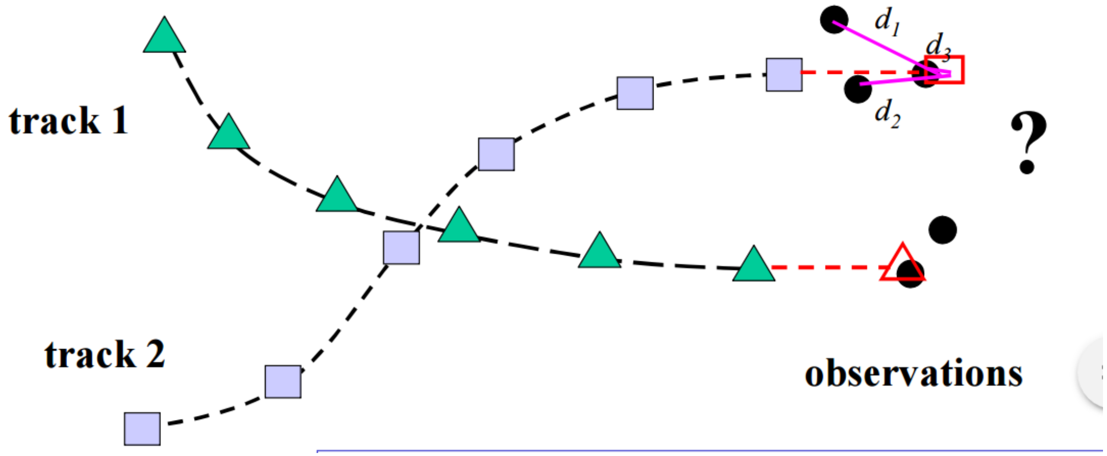
Match should be close to predicted values



some matches are highly unlikely

DIFFEREENT STRATEGIES

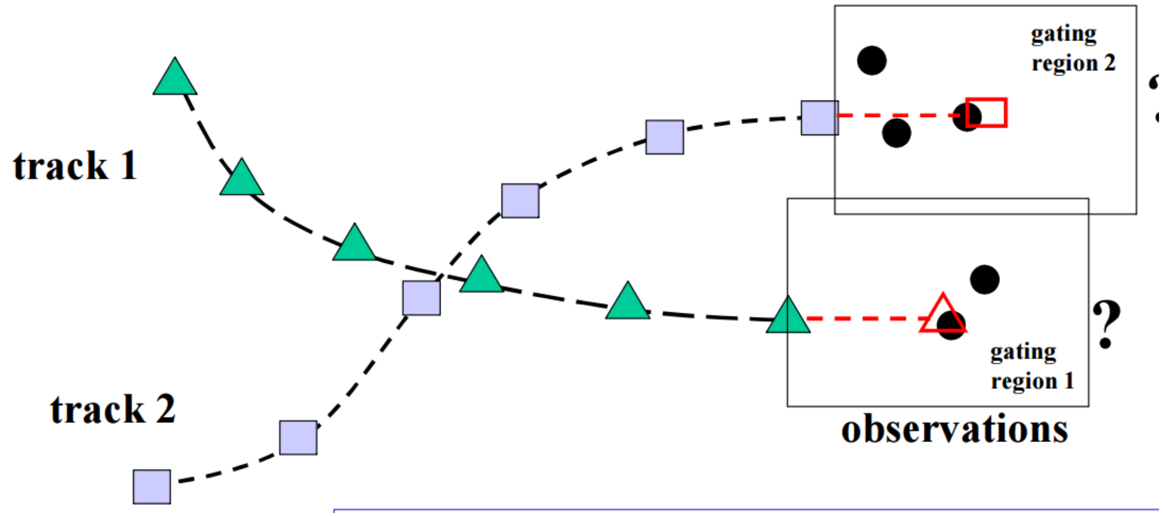
OF ASSOCIATION



GATING

A method for pruning matches that are geometrically unlikely from the start. Allows us to decompose matching into smaller subproblems.

→ Divide and conquer!



MTT AS A DATA ASSOCIATION MODEL

Studied in the field of radar technology 30 years ago

Three major categories

1. **Nearest neighbor (NN)** — **online**
2. **Joint probabilistic data association (JPDA)** — **on a gating window**
3. **Multiple hypothesis tracking (MHT)** — **on the overall data**

NN and JPDA work in a single scan of the dataset

Greedy approach: in each timestamp, every sample is associated with a single track

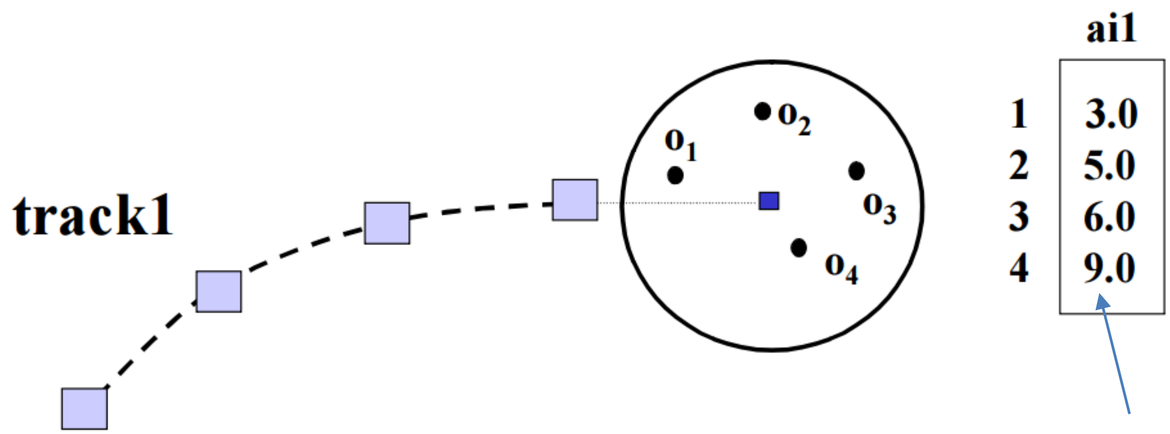
Objective: minimize the error across all associations in the current timestamp

Performance:

- Efficient – can work in polynomial time
- Greedy approach results in many false associations

1. NN NEAREST NEIGHBOR

Evaluate each observation in track gating region. Choose “best” one to incorporate into track

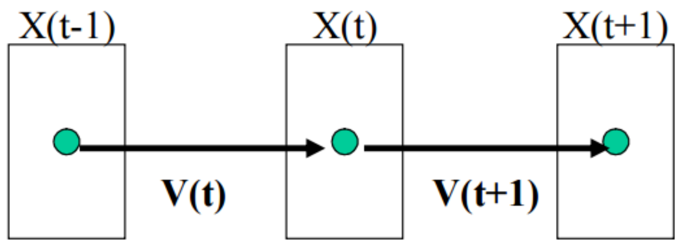


A1j is the score of matching j to track 1, based on:

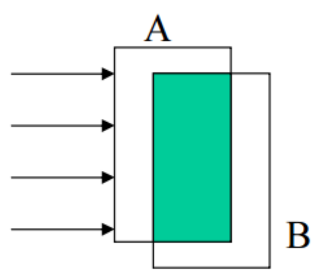
- Position
- similarity of appearance
- correlation scores…

Prediction

Model association



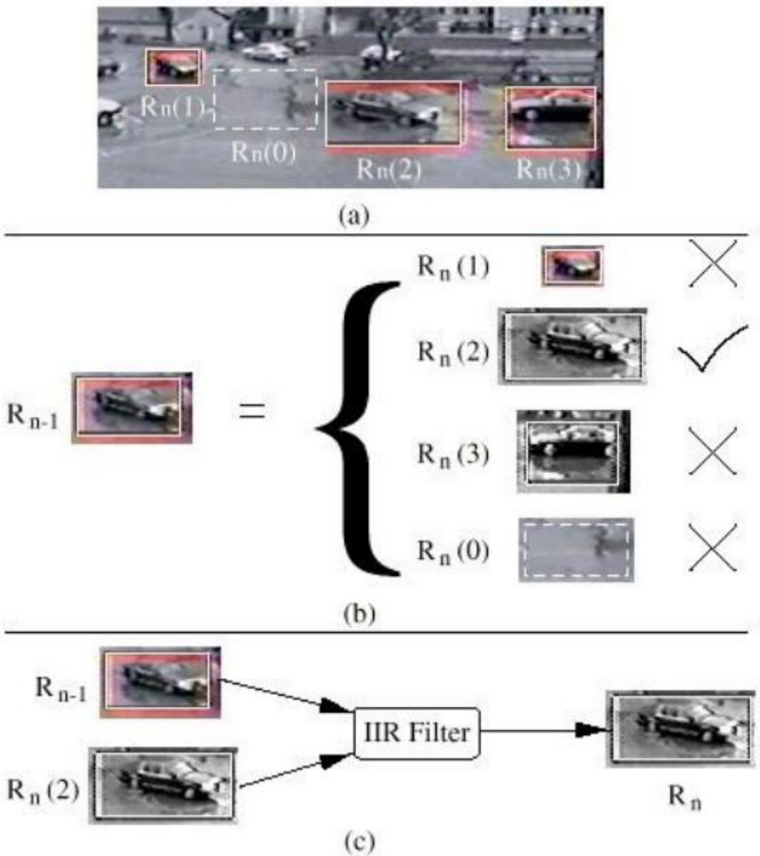
constant velocity assumes V(t) = V(t+1)



$$\text{score} = \frac{2 * \text{area}(A \text{ and } B)}{\text{area}(A) + \text{area}(B)}$$

A = bounding box at time t, adjusted by velocity V(t)
B = bounding box at time t+1

Correlation of image templates is an obvious choice (between frames)



Extract motion blobs

For object in previous frame, compute correlation score with all blobs in current fram Pick one with highest score (suboptimal strategy).

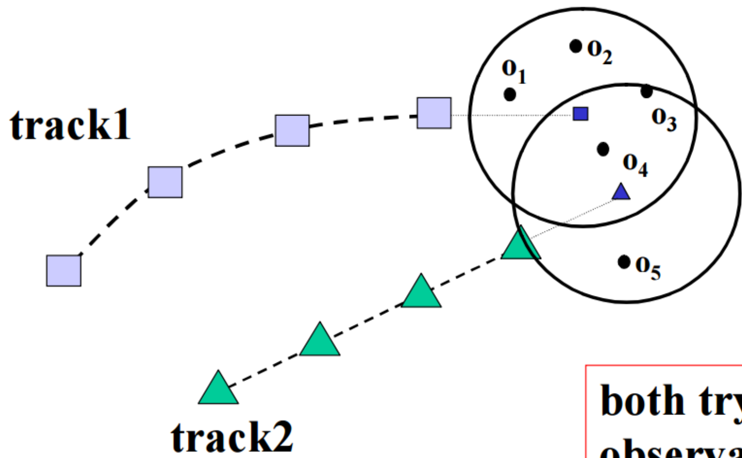
Update appearance template of blobs

However, cross correlation is computationally expensive.

PROBLEMS

BUT **after a switch no recover**

if do independently for each track, it could end up with contention for the same observations.



	ai1	ai2
1	3.0	
2	5.0	
3	6.0	1.0
4	9.0	8.0
5		3.0

both try to claim observation o_4

LINEAR ASSIGNMENT PROBLEM OR GLOBAL NEAREST NEIGHBOUR

Given N Target in a previous frame and M observation in the current frame

Choose a 1-1- correspondence

	1	**2**	**3**	**4**	**5**
1	0.95	0.76	0.62	0.41	0.06
2	0.23	0.46	0.79	0.94	0.35
3	0.61	0.02	0.92	0.92	0.81
4	0.49	0.82	0.74	0.41	0.01
5	0.89	0.44	0.18	0.89	0.14

Remember that there are 5x4x3x2c1=120 possibilities (N!)

Mathematical definition. Given an NxN array of benefits $\{X_{ai}\}$, determine an NxN permutation matrix M_{ai} that maximizes the total score:

maximize: $$E = \sum_{a=1}^{N} \sum_{i=1}^{N} M_{ai} X_{ai}$$

subject to:

$$\forall i \ \sum_{a=1}^{A} M_{ai} = 1$$

$$\forall a \ \sum_{i=1}^{I} M_{ai} = 1$$

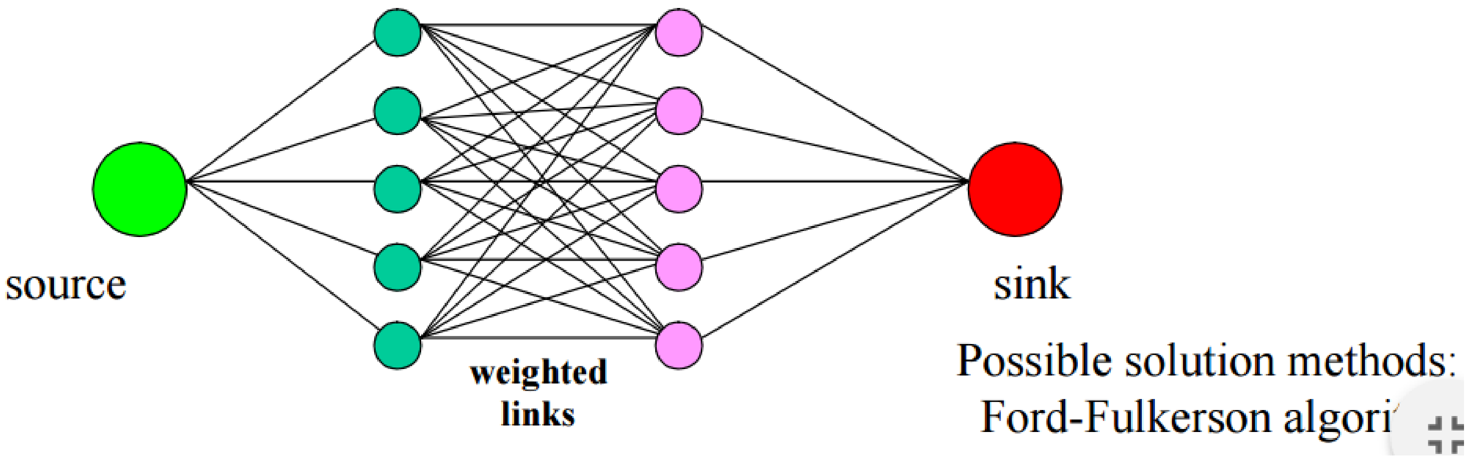
$$M_{ai} \in \{0, 1\}$$

constraints that say M is a permutation matrix

SOLUTIONS

Greedy strategy

The problem can also be viewed as a **weighted bipartite graph**, with nodes being row/col indices and edges being weighted by the matrix entries Xai. Perhaps this can be solved by mincut/maxflow? (polynomial complexity)



HUNGARIAN ALGORITHM

Hungarian algorithm

From Wikipedia, the free encyclopedia

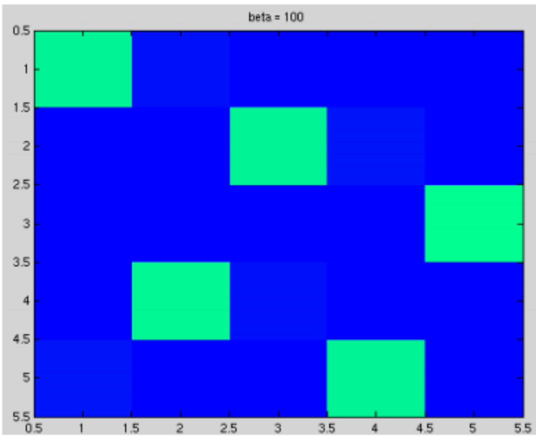
The **Hungarian algorithm** is a combinatorial optimization algorithm which solves assignment problems in polynomial time ($O(n^3)$). The first version, known as the **Hungarian method**, was invented and published by Harold Kuhn in 1955. This was revised by James Munkres in 1957, and has been known since as the **Hungarian algorithm**, the **Munkres assignment algorithm**, or the **Kuhn-Munkres algorithm**. In 2006, it was discovered that Carl Gustav Jacobi had solved the assignment problem in the early 19th century, and published posthumously in 1890 in the Latin language.[1]

The algorithm developed by Kuhn was largely based on the earlier works of two Hungarian mathematicians: Dénes König and Jenő Egerváry. The great advantage of Kuhn's method is that it is strongly polynomial (see Computational complexity theory for details). The main innovation of the algorithm was to combine two separate parts in Egerváry's proof into one.

hence the name

Again, courtesy of B.Collins

HUNGARIAN



permutation matrix computed by Hungarian Algorithm

0.95	0.76	0.62	0.41	0.06
0.23	0.46	**0.79**	0.94	0.35
0.61	0.02	0.92	0.92	**0.81**
0.49	**0.82**	0.74	0.41	0.01
0.89	0.44	0.18	**0.89**	0.14

score: 4.26

Improvements:

Murty's K-best algorithm

So far we know how to find the best assignment (max sum scores). But what if we also want to know the second best? Or maybe the top 10 best assignments?

2. PDAF

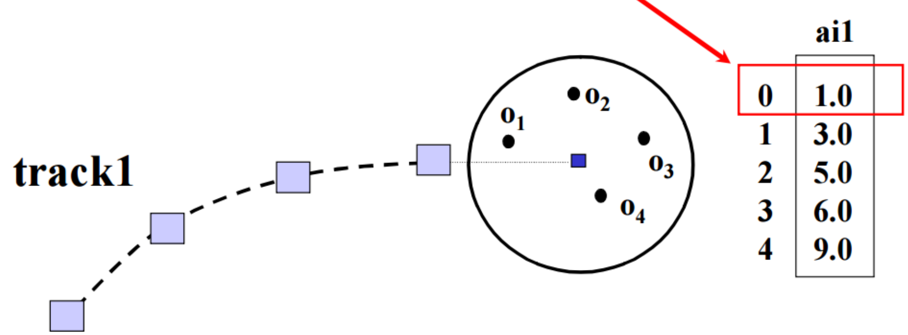
Probabilistic Data Association Filter

Updating single track based on new observations.

General idea: Instead of matching a single best observation to the track, we update based **on all observations (in gating window), weighted by their likelihoods.**

Use **Kalman for prediction**

Consider all points in gating window. Also consider the additional possibility that no observations match.

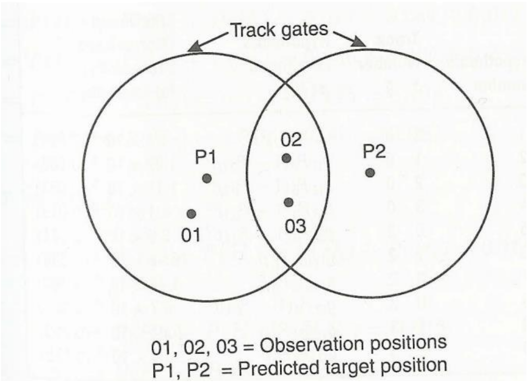


p_{i1} = "probability" of matching observation i to track 1

$$p_{i1} = \frac{a_{i1}}{\sum_{i=0}^{n} a_{i1}}$$

JPDAF

Joint Probabilistic Data Association Filter If maintaining multiple tracks, doing PDAF on each one independently is nonoptimal, since observations in overlapping gate regions will be counted more than once (contribute to more than one track). **JPDAF reasons over possible combinations of matche**s, in a principled way.



$$P(H) = \prod_{\substack{\text{Track i assigned} \\ \text{to observation j}}} g_{ij} P_D \prod_{\substack{\text{Tracks assigned} \\ \text{to no match (0)}}} (1-P_D) \prod_{\substack{\text{Unassigned} \\ \text{observations}}} B$$

Hypothesis Matrix for Example of Figure 6.3

Hypothesis Number	Track Number 1	2	Hypothesis Likelihood $p(H_i)$	Likelihood (Normalized Probability) for Example
1	0	0	$(1-P_D)^2\beta^3$	2.4×10^{-6} (0.011)
2	1	0	$g_{11}P_D(1-P_D)\beta^2$	1.82×10^{-5} (0.086)
3	2	0	$g_{12}P_D(1-P_D)\beta^2$	1.11×10^{-5} (0.053)
4	3	0	$g_{13}P_D(1-P_D)\beta^2$	4.1×10^{-6} (0.019)
5	0	2	$g_{22}P_D(1-P_D)\beta^2$	8.6×10^{-6} (0.041)
6	1	2	$g_{11}g_{22}P_D^2\beta$	6.47×10^{-5} (0.306)
7	3	2	$g_{13}g_{22}P_D^2\beta$	1.44×10^{-5} (0.068)
8	0	3	$g_{23}P_D(1-P_D)\beta^2$	6.7×10^{-6} (0.032)
9	1	3	$g_{11}g_{23}P_D^2\beta$	5.04×10^{-5} (0.239)
10	2	3	$g_{12}g_{23}P_D^2\beta$	3.06×10^{-5} (0.145)

JPDAF

Formally elegant but complex

using only a finite number of tracks

Called also **tracking before detection**, works on a fixed gating window (no online)

Tracking Multiple Interacting Targets Using a Joint Probabilistic Data Association Filter

Arsène Fansi Tchango*†, Vincent Thomas†, Olivier Buffet†, Alain Dutech† and Fabien Flacher*
*Thales Services SAS Company, Vélizy-Villacoublay, France
Email: firstname.lastname@thalesgroup.com
†INRIA / Université de Lorraine, Nancy, France
Email: firstname.lastname@loria.fr

Joint Probabilistic Data Association Revisited

ICCV 2015

Seyed Hamid Rezatofighi[1] Anton Milan[1] Zhen Zhang[2] Qinfeng Shi[1] Anthony Dick[1] Ian Reid[1]
[1]School of Computer Science, The University of Adelaide, Australia
[2]School of Computer Science and Technology, Northwestern Polytechnical University, Xian, China
hamid.rezatofighi@adelaide.edu.au

3. MHT MULTIPLE HYPOTHESIS TRACKING

Multiple hypotheses are maintained

- Joint probabilities are calculated recursively when new measurements are received

Each association is based on both previous and subsequent data (multiple scans)

Unfeasible hypotheses are eventually eliminated

Performance:

- Very accurate
- Computational and space complexity is exponential to the number of measurements

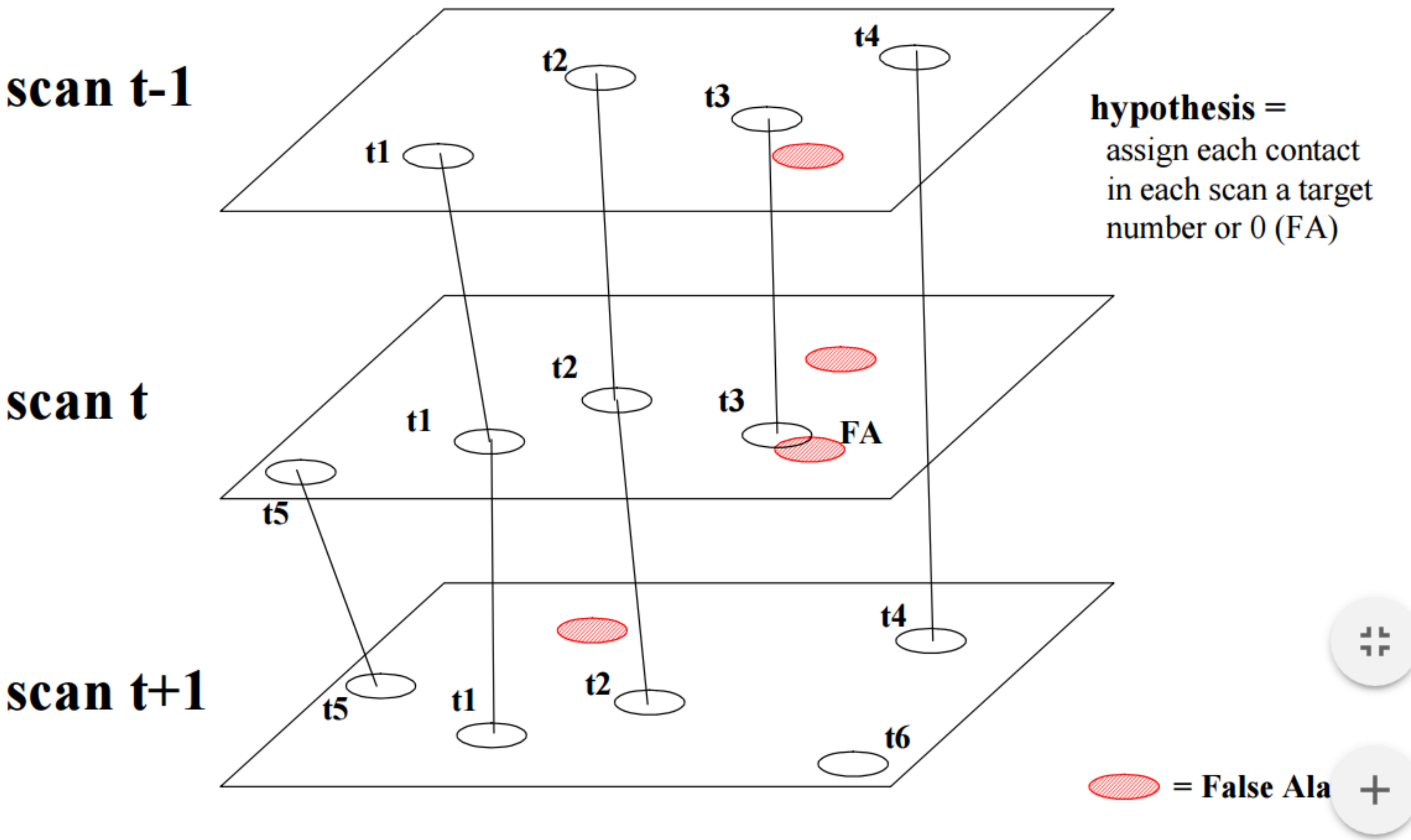
Benfold, B., Reid, I.: Stable multi-target tracking in real time surveillance video. In: CVPR. (2011)

Multiple Hypothesis Tracking Revisited

ICCV 2015

Chanho Kim † Fuxin Li ‡† Arridhana Ciptadi † James M. Rehg †
† Georgia Institute of Technology ‡ Oregon State University

MHT



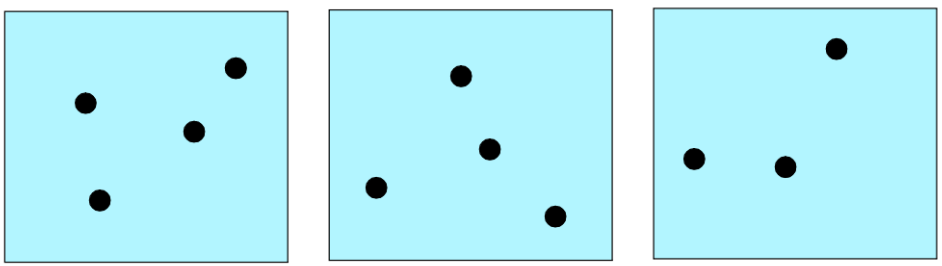
scan t-1
scan t
scan t+1
t1
t2
t3
t4
t5
t6
FA
hypothesis =
assign each contact
in each scan a target
number or 0 (FA)
= False Ala

MHT

Combiatorial explosion

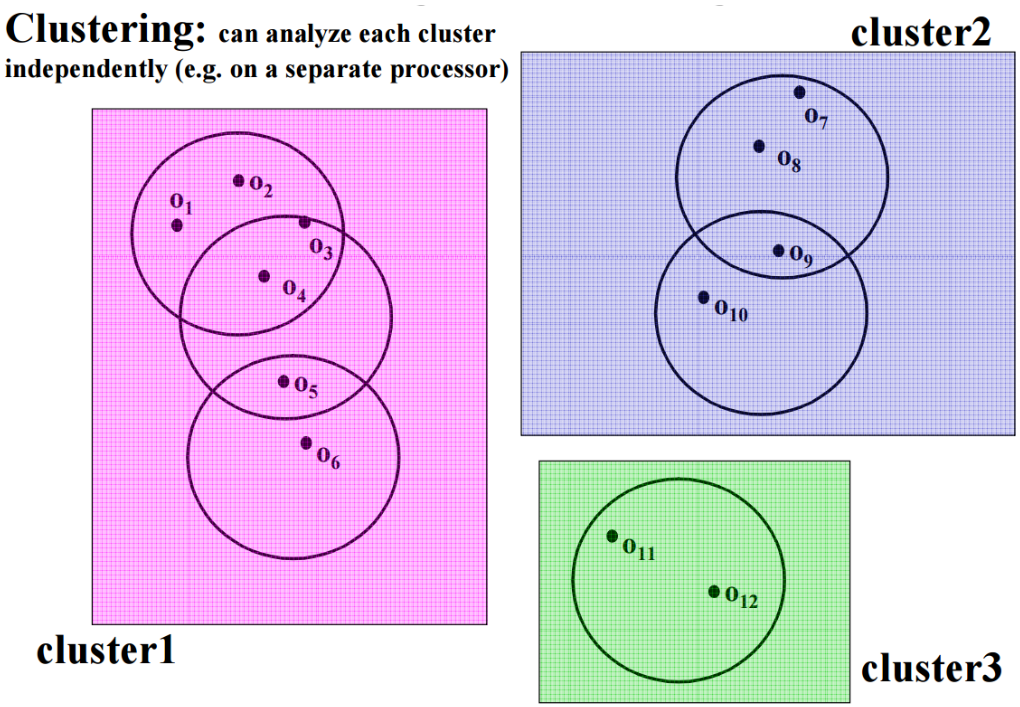
Rough order of magnitude on number of hypotheses:

Let’s say we have an upper bound N on number of targets and we can associate each contact in each scan a number from 1 to N. (we are ignoring false alarms at the moment)



$$\frac{N!}{(N-4)!} * \frac{N!}{(N-5)!} * \frac{N!}{(N-3)!}$$

MHT WITH MITIGATION STRATEGIES



combine MHT with Murty's k-best assignment algorithm to maintain a fixed set of k best hypotheses at each scan.
Cox et al TPAMI 96

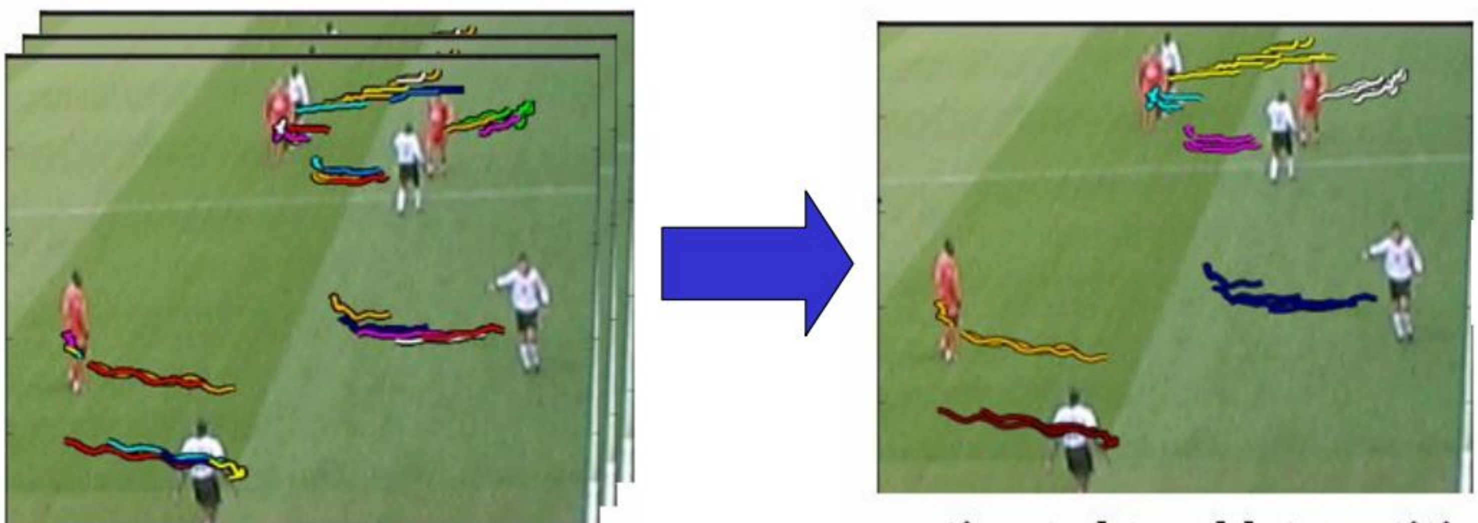
MCMCDA

Idea: use Markov Chain Monte Carlo (MCMC) to sample from / explore the huge combinatorial space of hypotheses.

- S. Oh, S. Russell, and S. Sastry, 2004. Markov Chain Monte Carlo data association for general multiple-target tracking problems. In Proc. IEEE Int. Conf. on Decision and Control, pages 735–742, 2004.
- Yu, G. Medioni, and I. Cohen, 2007. Multiple target tracking using spatio-temporal Markov Chain Monte Carlo data association. In Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition, pages 1–8, 2007.
- W.Ge and R.Collins, 2008, "Multi-target Data Association by Tracklets with Unsupervised Parameter Estimation," British Machine Vision Conference (BMVC'08), University of Leeds, September 2008, pp. 935-944.

MCDMA

Find a partition of the set of overlapping tracklets such that tracklets belonging to the same object are grouped together. They could obviously be merged after that by a postprocessing stage.

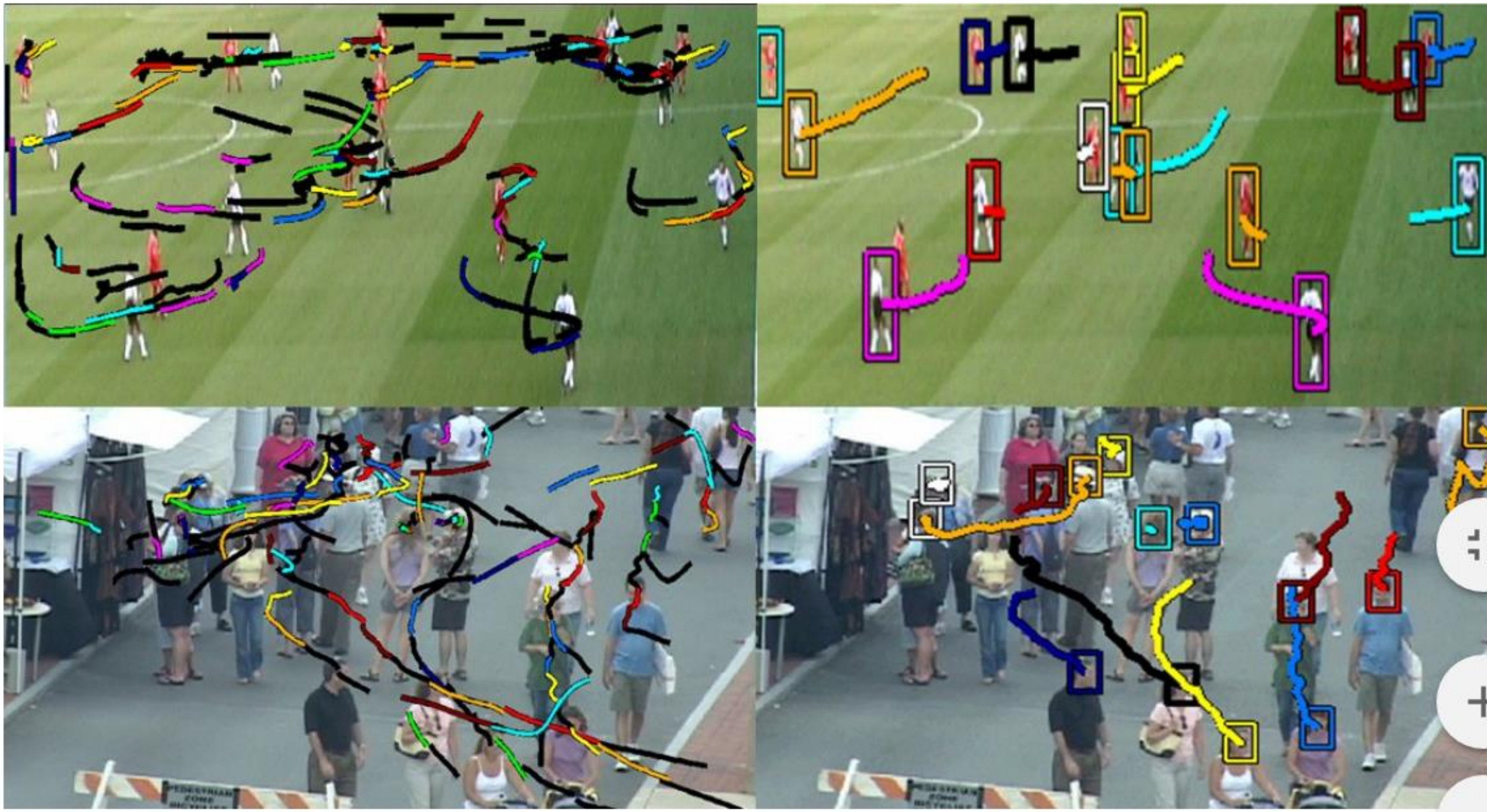


raw tracklets

estimated tracklet partition

input tracklets

hypothesized tracks (at some time)



MCMF: MIN-COST MAX-FLOW

Transform the tracking problem into a min-cost max-flow problem

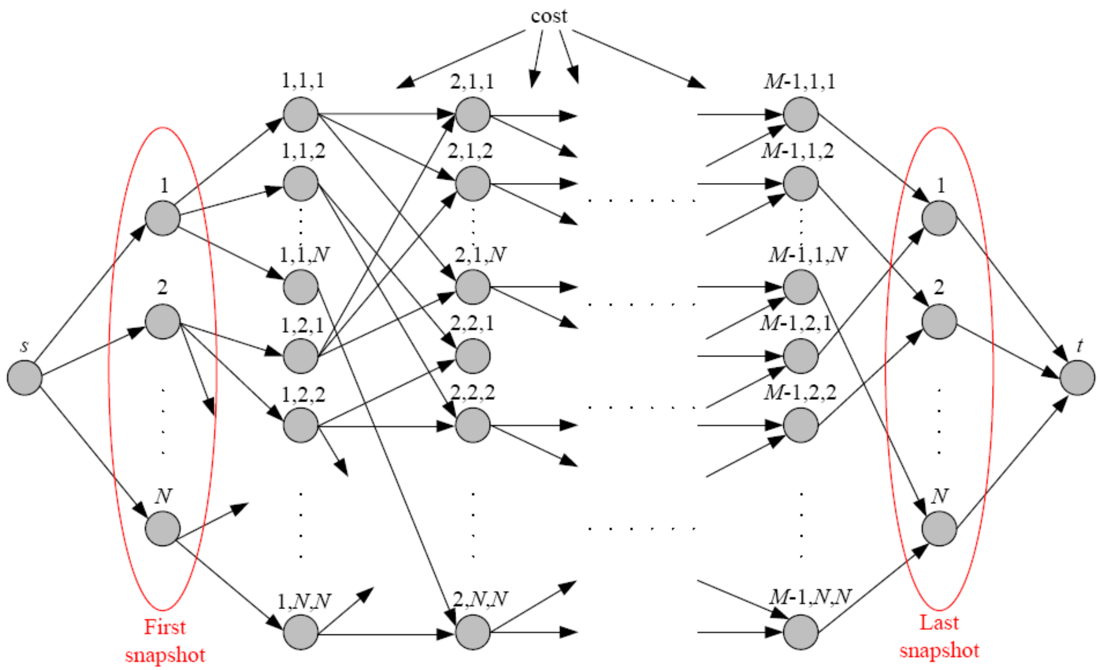
Min-cost max-flow (graph algorithm)

- Input: a weighted graph G with two special nodes (source s and destination t)
- Objective: find the maximum flow that can be sent from s to t that results in the minimum cost
- Well-known algorithms exist that work in polynomial time

All edges have capacity 1

Node id (t_i, p_i, p_j): the object moves from location p_i in timestamp t_i to location p_j in timestamp t_{i+1}

MAXIMUM FLOW: models the max number of recoverable trajectories
MIN COST: models the best frame-to-frame associations



GENERALIZED MINIMUM CLIQUE GRAPHS

ECCV 2012

GMCP-Tracker: Global Multi-object Tracking Using Generalized Minimum Clique Graphs

Amir Roshan Zamir, Afshin Dehghan, and Mubarak Shah

Not only matching on a temporal sequence

But an optimization which involves all observation in a time window

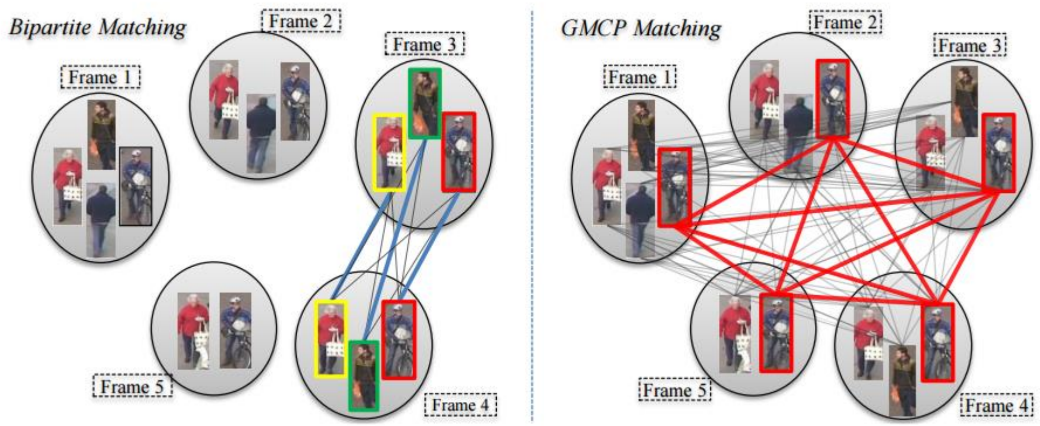
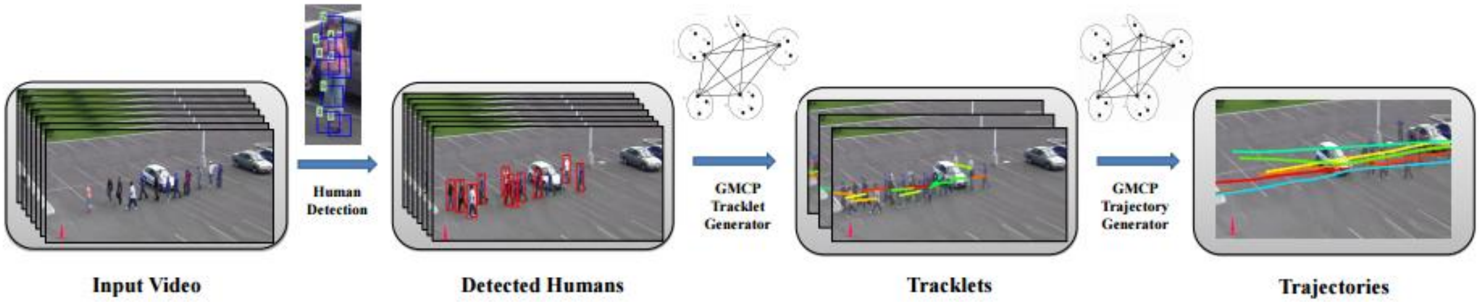


Fig. 1. Bi-partite vs. GMCP matching. Gray and colored edges represent the input graph and optimized subgraph, respectively. Bi-partite matches all objects in a limited temporal window. On the other hand, the proposed method matches one object at a time across full temporal span, while incorporating the rest of the objects implicitly.

USE OF TRACKLETS



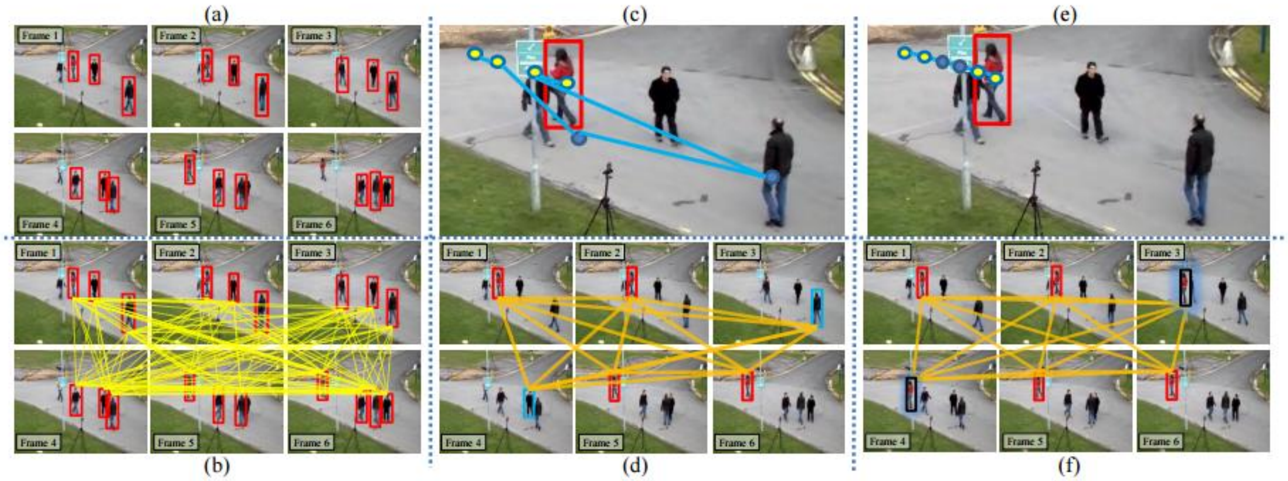


Fig. 3. Finding a tracklet for a small segment of 6 frames. The left column shows the detections in each frame along with graph G they induce. The middle column shows the feasible solution with minimal cost along with the tracklet it forms, *without* adding hypothetical nodes. The right column shows the feasible solution with minimal cost *with* hypothetical nodes added for handling occlusion, along with the tracklet it forms.

THE STATE-OF-ART TRACKING VS GMCP

[1]Benfold, B., Reid, I.: **Stable multi-target tracking** in real time surveillance video. In: CVPR. (2011)

[7] Yamaguchi, K., Berg, A., Ortiz, L., Berg, T.: who are you with and where are you going? In: CVPR. (2011)

[8] Leal-Taixe, L., et al.: Everybody needs somebody: Modeling social and grouping behavior on a linear programming multiple people tracker. (In: ICCV11Workshops)

[9] Pellegrini, S., Ess, A., van Gool, L.: Improving data association by joint modeling of pedestrian trajectories and groupings. In: ECCV. (2010) [10] Zhang, L., Li, Y., Nevatia, R.: Global data association for multi-object tracking using network flows. In: CVPR. (2008)

[11] . Brendel, W., Amer, M., Todorovic, S.: Multiobject tracking as maximumweight independent set. In: CVPR. (2011)

Table 1. Tracking results on Town Center sequence.

	MOTA	MOTP	MODP	MODA
Benfold et al. [1]	64.9	80.4	80.5	64.8
Zhang et al. [10]	65.7	71.5	71.5	66.1
Pellegrini et al. [9]	63.4	70.7	70.8	64.1
Yamaguchi et al. [7]	63.3	70.9	71.1	64.0
Leal-Taixe et al. [8]	67.3	71.5	71.6	67.6
Ours/GMCP	**75.59**	**71.93**	**72.01**	**75.71**

2012

MEASURING RESULTS IS HARD

TRACKING MEASURES

The tracking instance is correct *if the target is **detected** and **identified** and the location is correct at each frame.*

-In Single object tracking (SO-T) "detection" is the same of "**identification**" in the correct location.

-In Multiple object tracking (MO-T or MTT) a correct tracking must avoid also exchanges in the identification, so tracking is good **association**.

-In Multiple Camera tracking goodness is measured in **precision** in camera handoff, for overlapped FoVs, or in term of **re-identification** for multiple camera (and multiple targets) with not overlapped FoVs

TRACKING MEASURES

The three basic types of errors in tracking are:

- **False positive**: tracker identifies a target which is not a target.
- **False negative**: tracker misses to identify and locate the target.
- **Deviation**: the track's location deviated from the ground truth.

n_t^i **:Number of true positives in the frame i,** i.e. of correct instances

n_{fp}^i **:Number of false positives in the frame i,**

n_{fn}^i **:Number of false negatives in the frame i,**

And in MO-T:

n_{fa}^i **:Number of false associations in the frame i,**

in case of SO-T $n_t^i, n_{fn}^i, n_{fp}^i = (0, 1)$, in case of MO-T they can be more than one (!)

TRACKING MEASURES

Let's call:

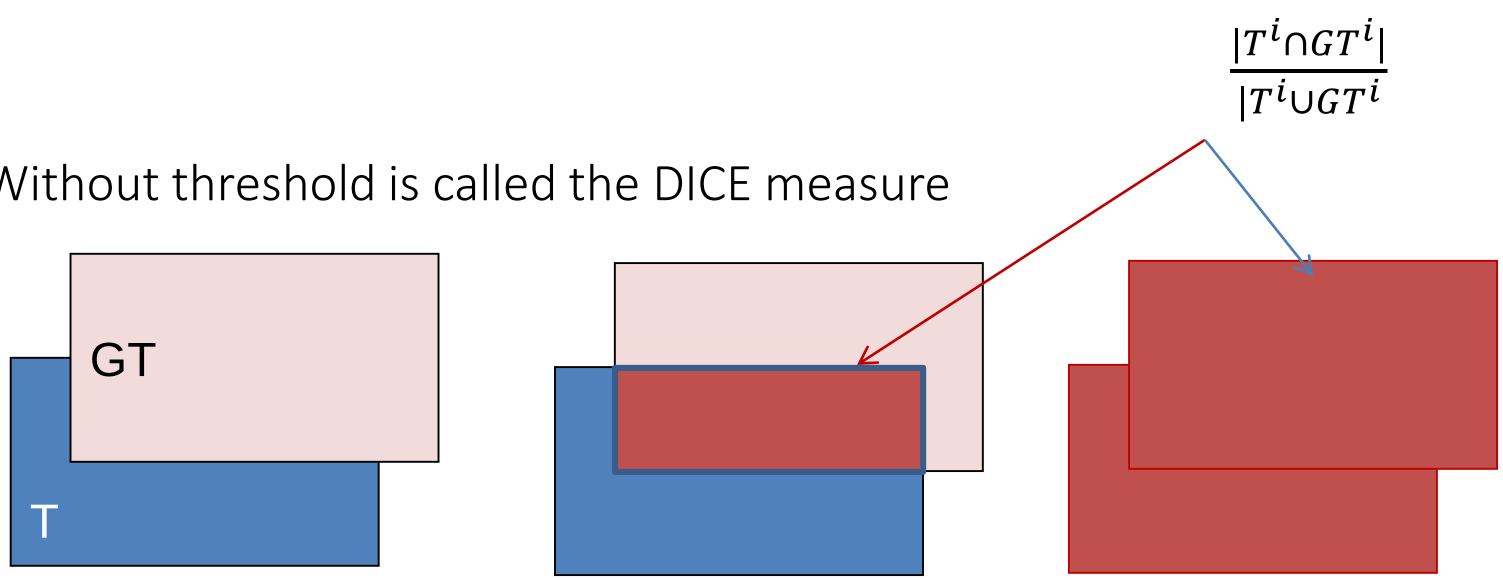
GT^i the ground truth in the frame i

T^i the Detected target in the frame i

Match degree at pixel level MD= $\frac{|T^i \cap GT^i|}{|T^i \cup GT^i|}$ intersection over union

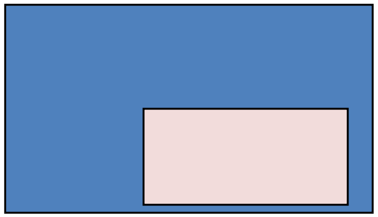
Match at pixel level if level $\frac{|T^i \cap GT^i|}{|T^i \cup GT^i|} \geq Th$ *Th=0,5 PASCAL measure [4]*

Without threshold is called the DICE measure

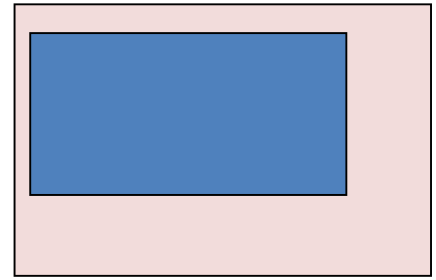


TRACKING MEASURES

The DICE measure can give the accuracy in term of precision and recall **at pixel level**



$recall = 1 \;\; precision < 1$

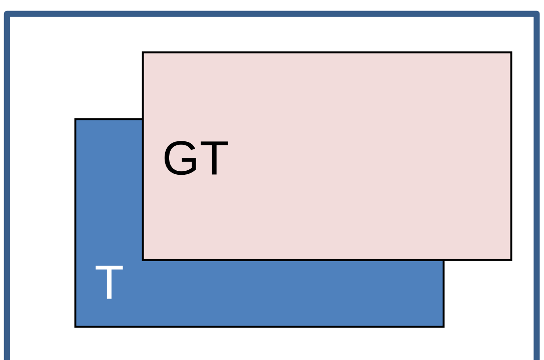


$precision = 1 \;\; recall < 1$

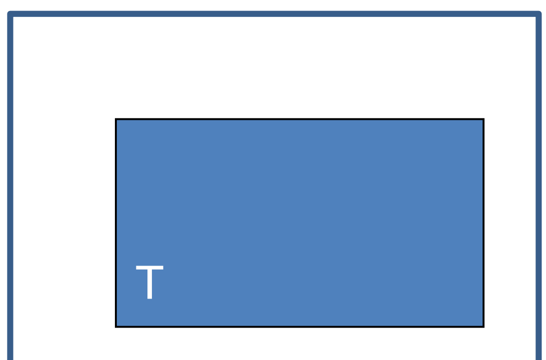
PASCAL-VOC with a given threshold works **at object level**

If (PASCAL) then $\mathbf{nt^{i}=1;}$

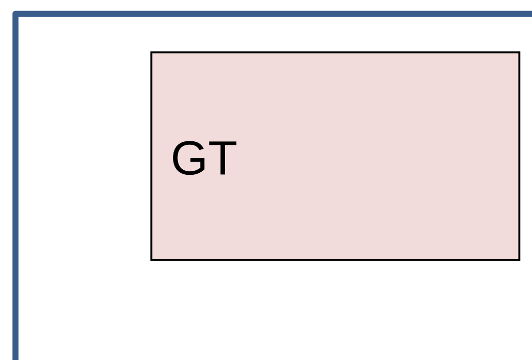
$\mathbf{n_{tp}^{\;i}=1;}$



$\mathbf{n_{fp}^{\;i}=1}$



$\mathbf{n_{fn}^{\;i}=1}$



TRACKING MEASURES

Ad object level in a sequence of frames (for i=1 Nframe):

$n_{tp} = \sum_{i=1}^{Nframe} n^i_{tp}$ $nf_p = \sum_{i=1}^{Nframe} n^{if}_{p}$ $n_{fn} = \sum_{i=1}^{Nframe} n^i_{fn}$

Precision = $(n_{tp})/(n_{tp} + n_{fp})$ Recall= $(n_{tp})/(n_{tp} + n_{fn})$

F-SCORE $\mathbf{F} = \mathbf{2}\,\frac{\boldsymbol{Precision * Recall}}{\boldsymbol{Precision + Recall}}$ (also called Correct track ratio)

At area/pixel level

$r^i = \frac{|T^i \cap GT^i|}{|T^i|}$ and $\mathrm{p}^i = \frac{|T^i \cap GT^i|}{|GT^i|}$

F1-SCORE $\mathrm{F1} = \frac{1}{Nframe} \sum_{i=1}^{Nframe} 2\,\frac{P^i * Ri}{P^i + R^i}$

TRACKING MEASURES

Similar to F-score OTA is accuracy in sequence

OTA $$\mathrm{OTA} = 1 - \frac{\sum_{i=1}^{Nframe} (n^i_{fp} + n^i_{fn})}{\sum_{i=1}^{Nframe} g^i}$$

OTA (**object track accuracy**) is generalized in MOTA for MT-tracking

gi is the number of ground truth objects in the frame i (gi=ntp+nfp) that is 1 in the frames where the object is present, to normalize OTA

And OTP (precision at pixel level) using DICE

$$OTP = \frac{1}{|Mi|} \sum_{i}^{in\ Mi} \frac{|T^i \cap GT^i|}{|T^i \cup GT^i|}$$

Mi is the frame where there is a matching

Thus OTP, OTA, and F-scores are similar.

TRACKING MEASURES

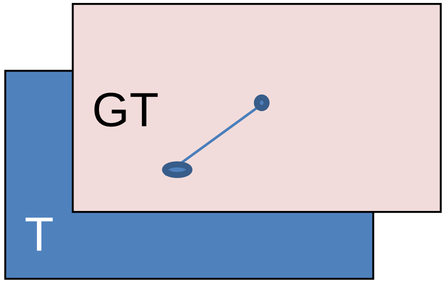
For measuring the position deviation instead

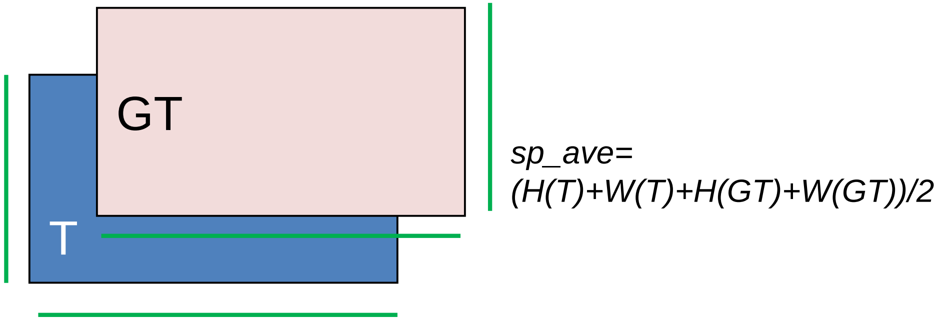
Deviation=1-$\frac{\sum_{i \in Mi} d(CTi-CGTi)}{|Mi|}$ d(x,y) distance L2 norm of the centroids

PBM Position Based Matching

PBM=$\frac{1}{Nframes}\sum_i\left(1-\frac{d1(Ti,GTi)}{sp\_ave(i)}\right)$ d1(x,y) is the L1 norm

sp_ave is the average semi-perimeter between GT and T





IN CONCLUSIONS

1. **Measures at pixel-level** or area-level, when a segmentation is available
2. **Measures at object-level** where tracking works with bounding box

- Evaluating the capacity of tracking = holding the frames: Fscore, F1score, OTA..
- Evaluate both accuracy and precision: FS-CORE varying the thershold, so if the th is lower it measures the accuracy when a lower precision is accepted
- Evaluating the capacity in position location: Deviation or BPM

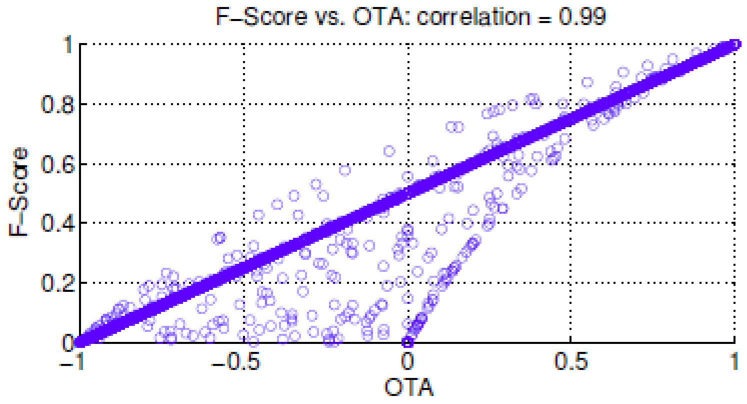
CORRELATION BETWEEN MEASURES

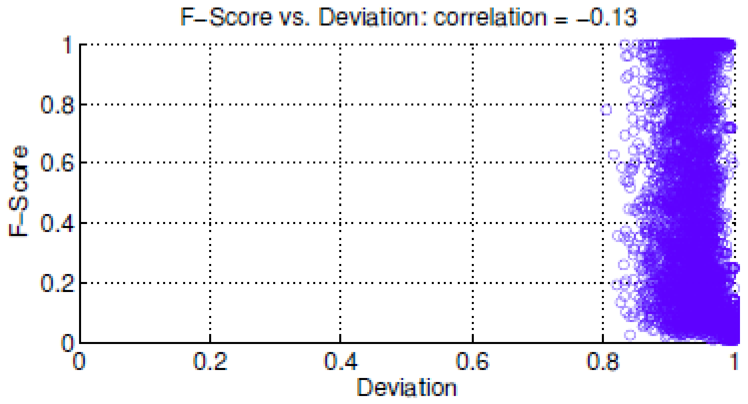
Measures are similar ; do not waste your time if measures are correlated

→ We use **F-SCORE and Deviation only** since [1]

Experiment with 19 tracker over 315 videos (5985 trials)

- F-score and OTA as a correlation of 0,99
- F-score and F1-score are correlated working at bb correlation of 0.91
- F-score and Deviation no correlation 0.13
- F-score and PBM more correlated about 0,79 (it could be useful to)





MEASURES

Name	Equation	Target	Measure
F-score [34]	$2 \cdot \frac{precision \cdot recall}{precision + recall}$	Accuracy	Thresholded precision and recall
$F1$-score [22]	$\frac{1}{N_{frames}} \sum_i 2 \cdot \frac{p^i \cdot r^i}{p^i + r^i}$	Accuracy	Precision and recall
OTA [31]	$1 - \frac{\sum_i (n^i_{fn} + n^i_{fp})}{\sum_i g^i}$	Accuracy	False positive and false negative
OTP [30]	$\frac{1}{\lvert M_s \rvert} \sum_{i \in M_s} \frac{\lvert T^i \cap GT^i \rvert}{\lvert T^i \cup GT^i \rvert}$	Accuracy	Average overlap over matched frames
ATA [22]	$\frac{1}{N_{frames}} \sum_i \frac{\lvert T^i \cap GT^i \rvert}{\lvert T^i \cup GT^i \rvert}$	Accuracy	Average overlap
$Deviation$ [38]	$1 - \frac{\sum_{i \in M_s} d(T^i, GT^i)}{\lvert M_s \rvert}$	Location	Centroid normalized distance
PBM [22]	$\frac{1}{N_{frames}} \sum_i \left[1 - \frac{Distance(i)}{T_h(i)} \right]$	Location	Centroid L1-distance

See Smeulder et al. TPAMI 2013

MULTI TARGET TRACKING

With MTT normally tracking by detection is used thus

Multiple object detection precision (MODP) is the 2D precision of the detection level

Multiple object detection accuracy(MODA) is the detection accuracy counting false positives and negatives

Multiple object tracking precision (MOTP) is the 2D location precision of the target asociation level

Multiple object tracking accuracy (MOTA) is the tracking accuracy counting false positives, negatives and identity switches too

WITH MULTIPLE TARGETS

MOTA Multiple object tracking accuracy:

$$\text{MOTA} = 1 - \frac{\sum_{i=1}^{Nframe} (n^i_{fp} + n^i_{fn} + n^i_{ids})}{\sum_{i=1}^{Nframe} g^i}$$

MOTP Multiple object tracking precision

$$\textbf{MOTP} = 1 - \frac{\sum_{i \in Mi} d(CTi - CGTi)}{|Mi|}$$

Here Mi are the number of associated tracks

MOT CHALLENGE MEASURE AND BENCHMARK

MOTChallenge 2015:
Towards a Benchmark for Multi-Target Tracking

Laura Leal-Taixé*, Anton Milan*, Ian Reid, Stefan Roth, and Konrad Schindler

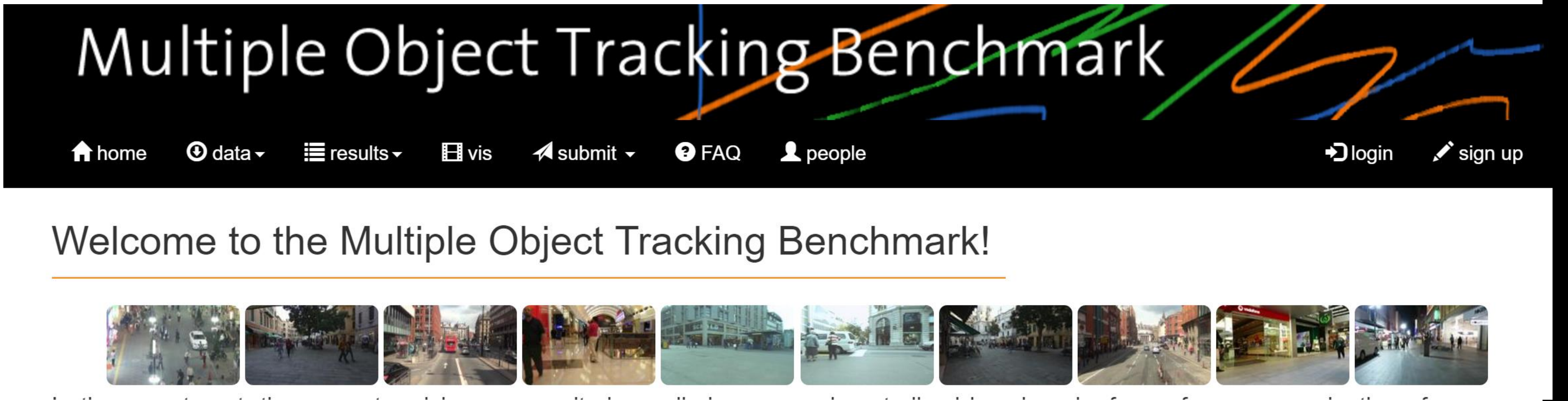
The state of the art:

Small dataset, World-wide known

- 22 sequences, half for training and half for testing, with a total of 11286 frames or 996 seconds of video.
- Camera calibration is provided for 4 to 3D real-world coordinate tracking.
- precomputed object detections, annotations, and a common evaluation method for all datasets

(only few from static cameras)

motchallenge.net



PERFORMANCE ANALYSIS

Only N.9 has a public detector!

Tracker	Avg Rank	MOTA	MOTP	FAF	MT	ML	FP	FN	ID Sw.	Frag	Hz	Detector
NOMTwSDP 1.	7.0	**55.5** ±11.2	**76.6**	1.0	39.0%	25.8%	5,594	21,322	427 (6.5)	701 (10.7)	6.4	Private
	W. Choi. Near-Online Multi-target Tracking with Aggregated Local Flow Descriptor. In ICCV, 2015.											
AMPL 2.	13.5	51.9 ±11.9	75.0	1.2	26.4%	24.8%	6,963	22,225	372 (5.8)	1,130 (17.7)	2.8	Private
	Anonymous submission											
LKDAT_CNN 3.	16.6	49.3 ±11.8	74.5	1.0	20.8%	28.4%	6,009	24,550	563 (9.4)	1,155 (19.2)	1.2	Private
	Yuan Zhang, Di Xie and Shiliang Pu (Hikvision Research Institute)											
TSML_CDE 4.	13.6	49.1 ±13.0	74.3	0.9	30.4%	26.4%	5,204	25,460	637 (10.9)	1,034 (17.7)	6.5	Private
	B. Wang, G. Wang, K. L. Chan, L. Wang. Tracklet Association by Online Target-Specific Metric Learning and Coherent Dynamics Estimation. In arXiv:1511.06654, 2015.											
justry 5.	27.9	45.2 ±17.0	74.7	2.4	**40.6%**	16.0%	14,117	**18,769**	764 (11.0)	1,413 (20.3)	2.6	Private
	Anonymous submission											
DMT 6.	23.0	44.5 ±11.8	72.9	1.4	34.7%	22.1%	8,088	25,335	684 (11.6)	1,253 (21.3)	1.2	Private
	Anonymous submission											
YTBD 7.	13.8	44.0 ±10.9	73.9	1.1	19.4%	28.7%	6,149	27,649	598 (10.9)	1,223 (22.2)	**1,156.6**	Private
	Anonymous submission											
PHD_PF 8.	28.2	42.9 ±10.3	72.3	1.6	18.0%	23.4%	9,436	24,816	809 (13.6)	1,327 (22.3)	1.0	Private
	R. Sanchez, F. Poiesi, A. Cavallaro. Under review.											
DTA 9.	28.0	41.9 ±12.5	72.3	1.6	31.9%	22.6%	9,450	25,372	856 (14.6)	1,401 (23.9)	1.2	**Public**
	Anonymous submission											

PART I: by FRANCESCO SOLERA

Multi-Target Tracking Evaluation

AVSS 2015 Best Paper Award

Towards the evaluation of reproducible robustness in tracking-by-detection

Francesco Solera   Simone Calderara   Rita Cucchiara
Department of Engineering Enzo Ferrari
University of Modena and Reggio Emilia
name.surname@unimore.it

Abstract

Conventional experiments on MTT are built upon the belief that fixing the detections to different trackers is sufficient to obtain a fair comparison. In this work we argue how the true behavior of a tracker is exposed when evaluated by varying the input detections rather than by fixing them. We propose a systematic and reproducible protocol and a MATLAB toolbox for generating synthetic data starting from ground truth detections, a proper set of metrics to understand and compare trackers peculiarities and respective visualization solutions.

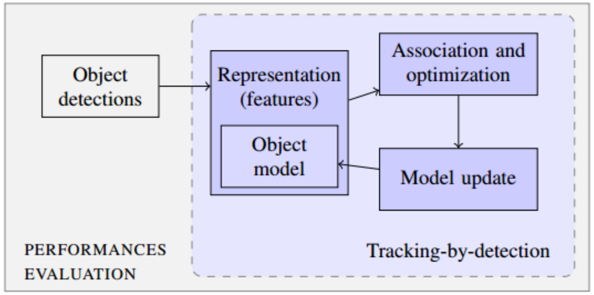
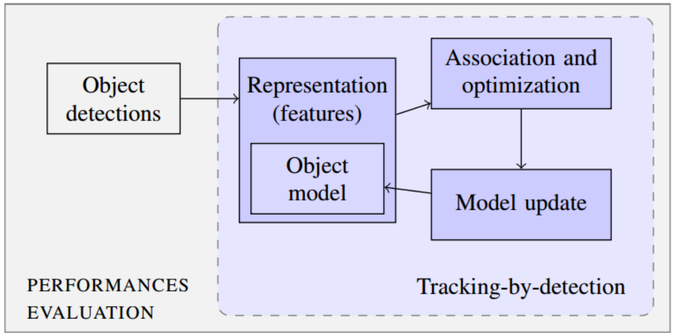


Figure 1: Tracking-by-detection overview scheme. Tracking evaluation cannot be decoupled from detections.

TRACKING BY DETECTION



Typical pipeline:

1. People detection
2. Feature extraction from BB
3. Detection-to-Identity (data) association

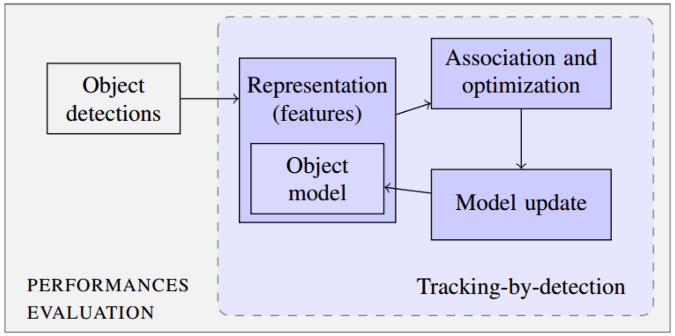
Everything comes after "people detection"!

Better (worst) detections means better (worst) tracking!

If you score higher than another tracker, who deserves the merit:

- Higher quality detections or
- Better tracking ability?

TRACKING BY DETECTION



Typical pipeline:

1. People detection
2. Feature extraction from BB
3. Detection-to-Identity (data) association

Everything comes after "people detection"!

Better (worst) detections means better (worst) tracking!

If you score higher than another tracker, who deserves the merit:

- Higher quality detections or
- Better tracking ability?

SOLUTION 1: FIX THE DETECTIONS

FIXING THE DETECTIONS IS OK… BUT NOT ENOUGH!

Multiple Object Tracking Benchmark

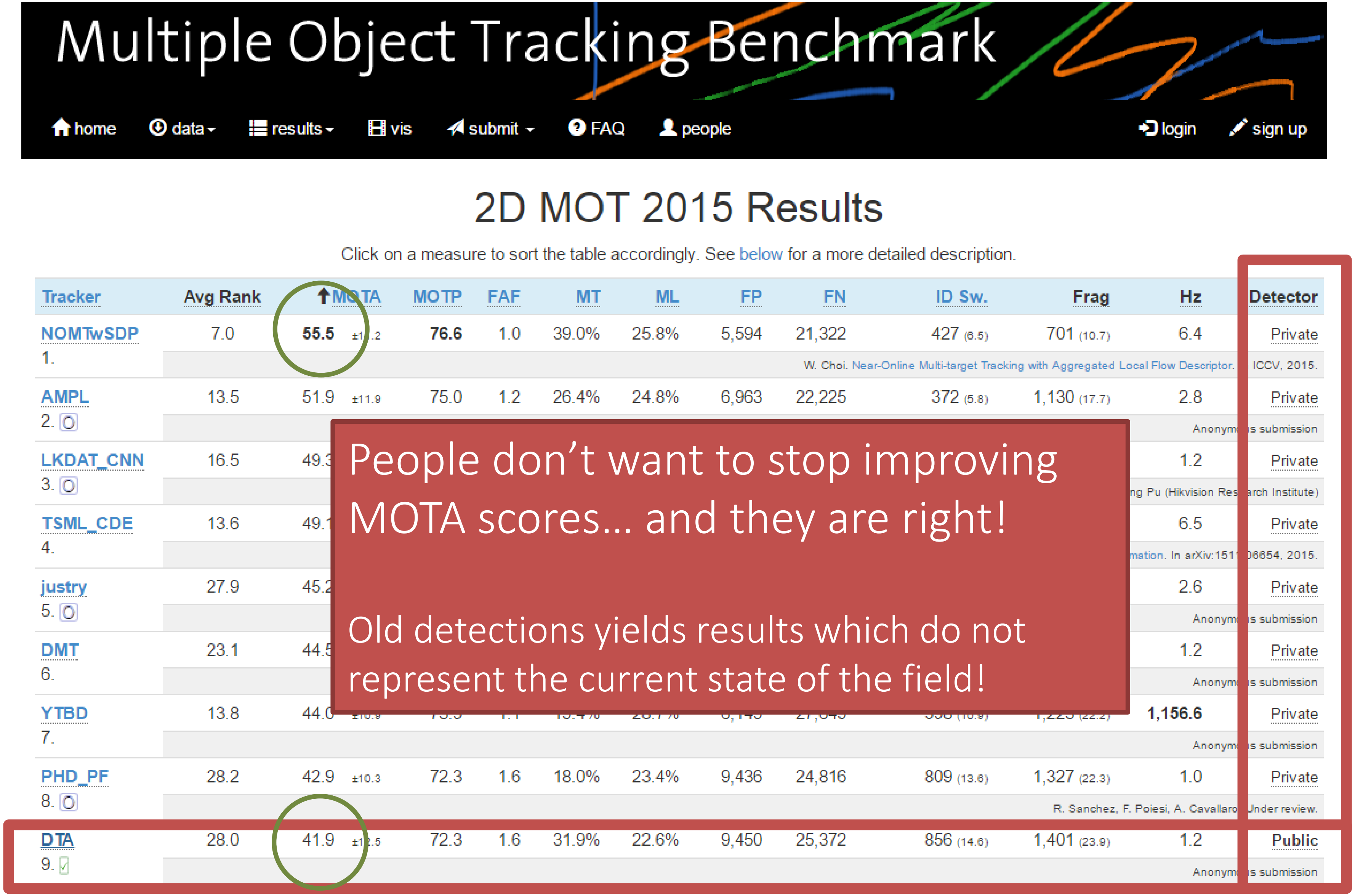
home | data | results | vis | submit | FAQ | people | login | sign up

2D MOT 2015 Results

Click on a measure to sort the table accordingly. See below for a more detailed description.

Tracker	Avg Rank	MOTA		MOTP	FAF	MT	ML	FP	FN	ID Sw.	Frag	Hz	Detector
NOMTwSDP 1.	7.0	**55.5**	±1[illegible].2	**76.6**	1.0	39.0%	25.8%	5,594	21,322	427 (6.5)	701 (10.7)	6.4	Private
										W. Choi. Near-Online Multi-target Tracking with Aggregated Local Flow Descriptor. ICCV, 2015.			
AMPL 2.	13.5	51.9	±11.9	75.0	1.2	26.4%	24.8%	6,963	22,225	372 (5.8)	1,130 (17.7)	2.8	Private
												Anonymous submission	
LKDAT_CNN 3.	16.5	49.3	±11.8	74.5	1.0	20.8%	28.4%	6,009	24,550	563 (9.4)	1,155 (19.2)	1.2	Private
										Yuan Zhang, Di Xie and Shiliang Pu (Hikvision Research Institute)			
TSML_CDE 4.	13.6	49.1	±13.0	74.3	0.9	30.4%	26.4%	5,204	25,460	637 (10.9)	1,034 (17.7)	6.5	Private
					B. Wang, G. Wang, K. L. Chan, L. Wang. Tracklet Association by Online Target-Specific Metric Learning and Coherent Dynamics Estimation. In arXiv:151[illegible]06654, 2015.								
justry 5.	27.9	45.2	±17.0	74.7	2.4	**40.6%**	16.0%	14,117	**18,769**	764 (11.0)	1,413 (20.3)	2.6	Private
												Anonymous submission	
DMT 6.	23.1	44.5	±11.8	72.9	1.4	34.7%	22.1%	8,088	25,335	684 (11.6)	1,253 (21.3)	1.2	Private
												Anonymous submission	
YTBD 7.	13.8	44.0	±10.9	73.9	1.1	19.4%	28.7%	6,149	27,649	598 (10.9)	1,223 (22.2)	**1,156.6**	Private
												Anonymous submission	
PHD_PF 8.	28.2	42.9	±10.3	72.3	1.6	18.0%	23.4%	9,436	24,816	809 (13.6)	1,327 (22.3)	1.0	Private
											R. Sanchez, F. Poiesi, A. Cavallaro. Under review.		
DTA 9.	28.0	41.9	±12.5	72.3	1.6	31.9%	22.6%	9,450	25,372	856 (14.6)	1,401 (23.9)	1.2	**Public**
												Anonymous submission	

FIXING THE DETECTIONS IS OK… BUT NOT ENOUGH!



People don’t want to stop improving MOTA scores… and they are right!

Old detections yields results which do not represent the current state of the field!

IF WE CANNOT FIX DETECTIONS…

SOLUTION 2: CHANGE THE DETECTIONS

… but in a controlled way!

IF WE CANNOT FIX DETECTIONS…

SOLUTION 2: CHANGE THE DETECTIONS

… but in a controlled way!

Detector performances are usually scored by precision P and recall R measures.

For all combinations of (P,R) in [0,1]x[0,1]:

- Starting from GT detections
- Add detections (FP), remove detections (FN)
- Move detections and change size (localization errors, yield FP+FN)
- Evaluate tracker

WE REMOVE THE DETECTOR BIAS

BY SIMULATING – IN A CONSISTENT WAY - ALL POSSIBLE DETECTORS

PROTOCOL

- ADD FP:
- add close to GT location
  - New location sampled from a gaussian distribution

$$(\bar{x}, \bar{y}) \sim \mathcal{N}((x, y), \sigma_1)$$

PROTOCOL

- ADD FP:
- add close to GT location
  - New location sampled from a gaussian distribution

$$(\bar{x}, \bar{y}) \sim \mathcal{N}((x, y), \sigma_1)$$

PROTOCOL

- ALTER FP:
  - Modify the BB size
  - Scale factor [0.5,1.5] with uniform **probability**

$$\mathcal{N}((w,h),\sigma_2)$$

- ADD FP:
- add close to GT location
  - New location sampled from a gaussian distribution

$$(\bar{x},\bar{y}) \sim \mathcal{N}((x,y),\sigma_1)$$

PROTOCOL

- ALTER FP:
  - Modify the BB size
  - Scale factor [0.5,1.5] with uniform **probability**

$$\mathcal{N}((w,h),\sigma_2)$$

- ADD FP:
- add close to GT location
  - New location sampled from a gaussian distribution

$$(\bar{x},\bar{y}) \sim \mathcal{N}((x,y),\sigma_1)$$

PROTOCOL

- ALTER FP:
  - Modify the BB size
  - Scale factor [0.5,1.5] with uniform **probability**

$$\mathcal{N}((w,h),\sigma_2)$$

- ADD FP:
- add close to GT location
  - New location sampled from a gaussian distribution

$$(\bar{x},\bar{y}) \sim \mathcal{N}((x,y),\sigma_1)$$

- REMOVE TP:
  - Randomly with uniform probability
  - Create FN

PROTOCOL

- ALTER FP:
  - Modify the BB size
  - Scale factor [0.5,1.5] with uniform **probability**

$$\mathcal{N}((w,h),\sigma_2)$$

- ADD FP:
- add close to GT location
  - New location sampled from a gaussian distribution

$$(\bar{x},\bar{y}) \sim \mathcal{N}((x,y),\sigma_1)$$

- REMOVE TP:
  - Randomly with uniform probability
  - Create FN

PROTOCOL

- ALTER FP:
  - Modify the BB size
  - Scale factor [0.5,1.5] with uniform **probability**

$$\mathcal{N}((w,h),\sigma_2)$$

- ADD FP:
- add close to GT location
  - New location sampled from a gaussian distribution

$$(\bar{x},\bar{y}) \sim \mathcal{N}((x,y),\sigma_1)$$

- REMOVE TP:
  - Randomly with uniform probability
  - Create FN

- RESIZE TP:
  - Sample new size from a Gaussian

PROTOCOL

- ALTER FP:
  - Modify the BB size
  - Scale factor [0.5,1.5] with uniform **probability**

$$\mathcal{N}((w,h),\sigma_2)$$

- ADD FP:
- add close to GT location
  - New location sampled from a gaussian distribution

$$(\bar{x},\bar{y}) \sim \mathcal{N}((x,y),\sigma_1)$$

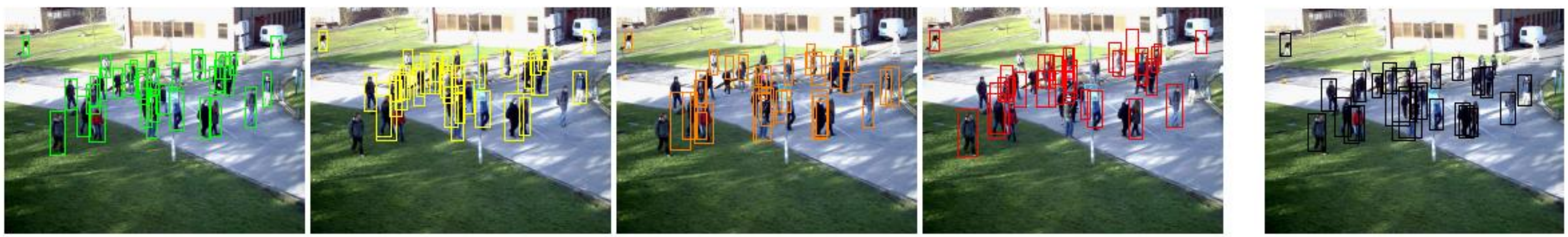
- REMOVE TP:
  - Randomly with uniform probability
  - Create FN

- RESIZE TP:
  - Sample new size from a Gaussian

PROTOCOL (CONTINUED)

To account for randomness:

- compute 5 instances for each (P,R) pair and
- report mean and variance



(a) P and R = 1 (b) P and R = 0.8 (c) P and R = 0.6 (d) P and R = 0.4 (e) MOT Challenge

SCENE COMPLEXITIES

Tracker must deal efficiently with occlusion

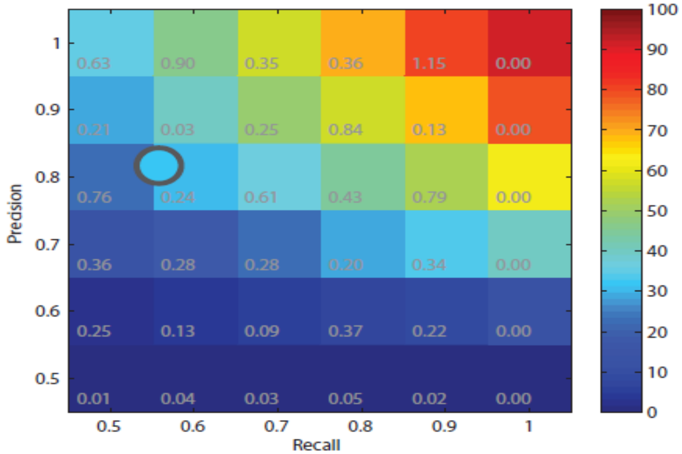
Occlusions:

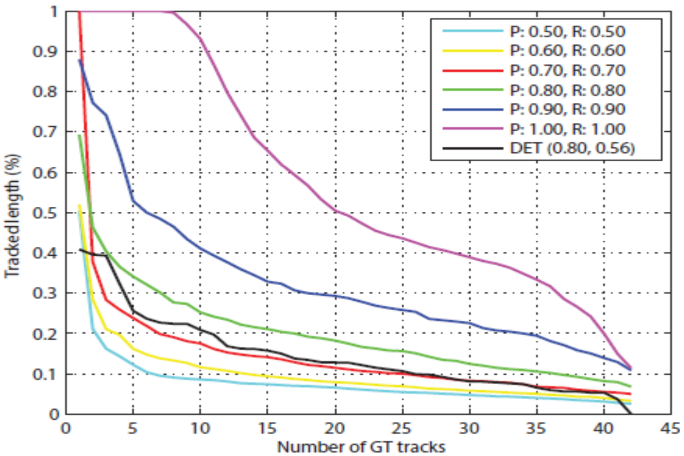
- depend on the scene (world occlusion) or
- on other targets
- but do not depend on the detector!

Generate occlusion from GT data under 2 parameters:

1. The percentage of occluded targets **N**
2. The percentage of occlusion w.r.t. the trajectory lenght **L**

VISUALIZATION TOOLS





MOTA matrix : MOTA values varying the 2 dataset parameters

TL Plots: different curve on matrix diagonal

- TL is defined as the % of the trajectory correctly tracked w.r.t. its lenght

A CASE STUDY PETS S2L2

2 trackers trk1 and trk2 10 years one from another

Well known sequence PETS S2L2

Which one is the best?

NEW TRENDS IN MTT

IMPROVING «OLD» APPROACHES

MHT

multiple hypothesis tracker (MHT) [Reid IEEE TAC79]

+ apperance models [JRehg et al ICCV15].

JPDA

joint probabilistic data association (JPDA) [Fortmann et al IEEE CDC80]

+ Optimization [Milan et al ICCV2015]

GNN

K shortest path optimization [Fua et al IEEE TPAMI11]

Appearance constraints [Fua et al ICCV11]

DEEP LEARNING FOR TRACKING?

CNN difficulties for

number parameters and thus data

constrained number of output

RNN loop is better

inserting **concept of memory**

mapping input in arbitrary ouput sequence as long as the sequence alignment and the input and output dimensions are known in advance.

CVPR2016

Online Multi-target Tracking using Recurrent Neural Networks

Anton Milan[1] Seyed Hamid Rezatofighi[1] Anthony Dick[1]
Konrad Schindler[2] Ian Reid[1]

[1]School of Computer Science, The University of Adelaide, Australia
[2]Photogrammetry and Remote Sensing Group, ETH Zürich

HOW IT IS WORK..

RNNs work in a sequential manner, where a prediction is made at each time step, given the previous state and possibly an additional input.

The core of an RNN is its hidden state h e Rn of size n that acts as the main control mechanism for predicting the output, one step at a time. In general,

RNNs may have multiple layers l = 1; :::;L.

We will denote hlt as the hidden state at time t on layer l.

h0 can be thought of as the input layer, holding the input vector, while hL the final representation to produce the output yt

Online Multi-target Tracking using Recurrent Neural Networks

Anton Milan[1] Seyed Hamid Rezatofighi[1] Anthony Dick[1]
Konrad Schindler[2] Ian Reid[1]

[1]School of Computer Science, The University of Adelaide, Australia
[2]Photogrammetry and Remote Sensing Group, ETH Zürich

PUTTING TOGETHER

Markov assumption+ Bayesian filtering

$$p(x_t|z_{1:t}) \propto p(z_t|x_t) \int p(x_t|x_{t-1}) p(x_{t-1}|z_{1:t-1}) dx_{t-1},$$

- Data (observations –models) association is not straigthforward if candidates , observations and states are multiple
- Time varying number of targets
  - spam new targets enteringn
  - Remove exit targets which disappeares indefinetily
- Problems:
- new target or false alarms?
- Exiting targets or miss detection?
- D?4 , x,y,h,w

$$x_t \in \mathbb{R}^{N \cdot D}$$

$$z_t \in \mathbb{R}^{M \cdot D}$$

Online Multi-target Tracking using Recurrent Neural Networks

Anton Milan[1] Seyed Hamid Rezatofighi[1] Anthony Dick[1]
Konrad Schindler[2] Ian Reid[1]

[1]School of Computer Science, The University of Adelaide, Australia
[2]Photogrammetry and Remote Sensing Group, ETH Zürich

A matrix of assignment probability

The assignment probability matrix $\mathcal{A} \in [0,1]^{N\times(M+1)}$ represents for each target (row) the distribution of assigning individual measurements to that target, *i.e.* $\mathcal{A}_{ij} = p(i \text{ assigned to } j)$ and $\forall i : \sum_j \mathcal{A}_{ij} = 1$. Note that an extra column in $\mathcal{A}$ is needed to incorporate the case that a measurement is missing. Finally, $\mathcal{E} \in [0,1]^N$ is an indicator vector that represents the existence probability of a target and is necessary to deal with an unknown and time-varying number of targets. We will use $(\sim)$ to explicitly denote the ground truth variables.

Online Multi-target Tracking using Recurrent Neural Networks

Anton Milan[1] Seyed Hamid Rezatofighi[1] Anthony Dick[1]
Konrad Schindler[2] Ian Reid[1]

[1]School of Computer Science, The University of Adelaide, Australia
[2]Photogrammetry and Remote Sensing Group, ETH Zürich

RNN hidden state

State, observation

Data association

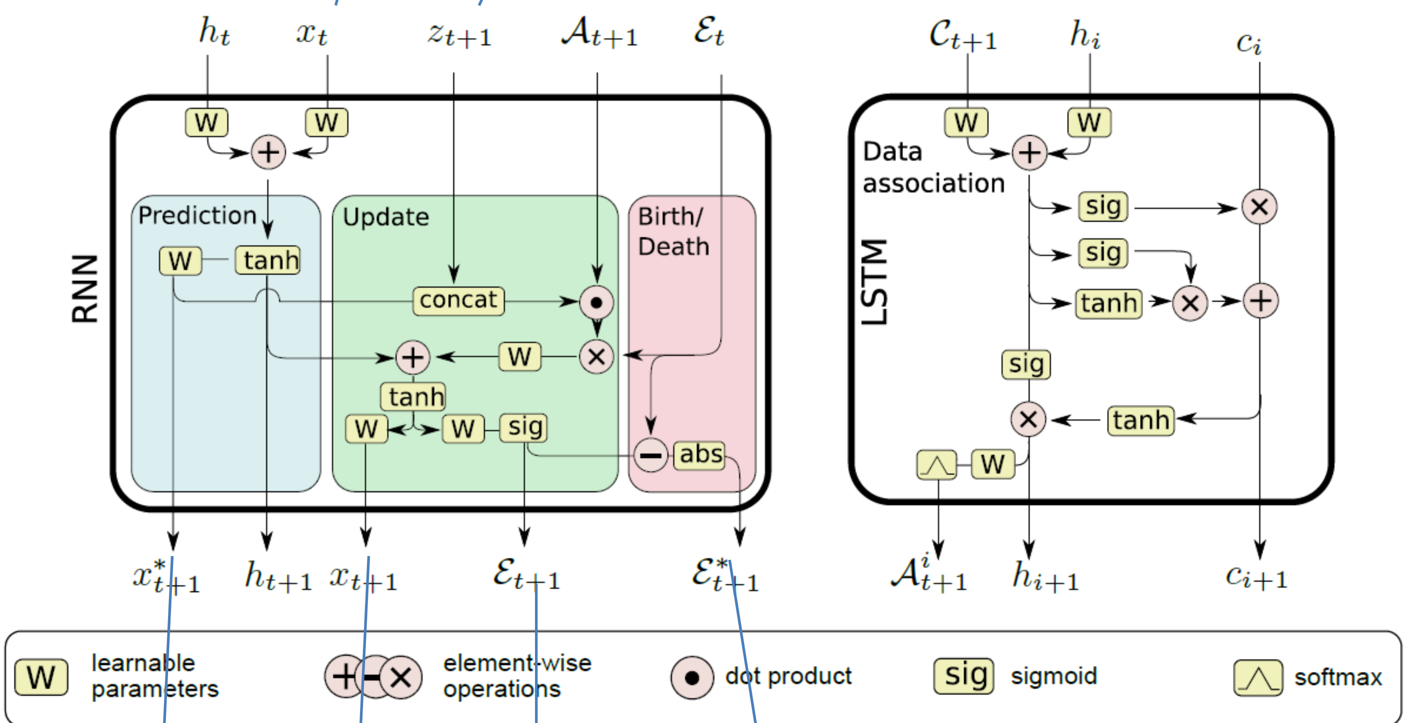


Fig. 2. Left: An RNN-based architecture for state prediction, state update, and target existence probability estimation. Right: An LSTM-based model for data association.

Predicted * (updated) state for each target

Online Multi-target Tracking using Recurrent Neural Networks

Anton Milan[1] Seyed Hamid Rezatofighi[1] Anthony Dick[1] Konrad Schindler[2] Ian Reid[1]

[1]School of Computer Science, The University of Adelaide, Australia
[2]Photogrammetry and Remote Sensing Group, ETH Zürich

Predicted (updated) probability
of existence of a real trajectory

LOSS

Predicted values and GT values

$$\mathcal{L}(x^*, x, \mathcal{E}, \widetilde{x}, \widetilde{\mathcal{E}}) = \underbrace{\frac{\lambda}{ND} \sum \|x^* - \widetilde{x}\|^2}_{\text{prediction}} + \underbrace{\frac{\kappa}{ND} \|x - \widetilde{x}\|^2}_{\text{update}} + \underbrace{\nu \mathcal{L}_{\mathcal{E}} + \xi \mathcal{E}^*}_{\text{birth/death + reg.}},$$

The loss take into account all the errors averaged over all targets on all frames

Intuitively, we aim to learn a network that predicts trajectories that are close to the ground truth tracks.

→we minimise the mean squared error (MSE) between state predictions and state update and the ground truth.

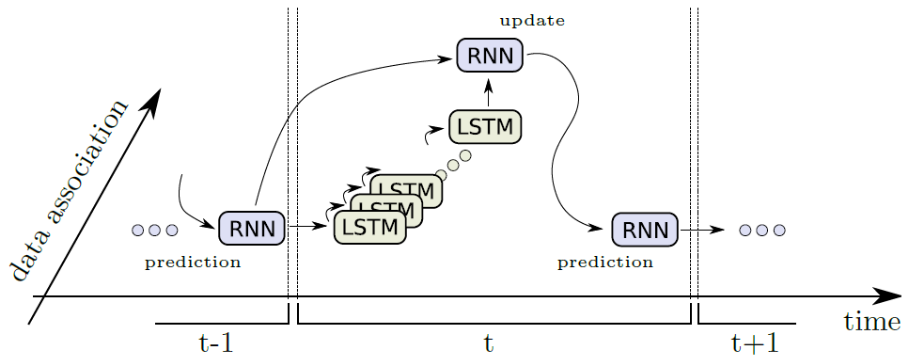
Online Multi-target Tracking using Recurrent Neural Networks

Anton Milan[1] Seyed Hamid Rezatofighi[1] Anthony Dick[1]
Konrad Schindler[2] Ian Reid[1]

[1]School of Computer Science, The University of Adelaide, Australia
[2]Photogrammetry and Remote Sensing Group, ETH Zürich

LONG SHORT-TERM MEMORY FOR DA

LSTM-based architecture that is able to learn to solve this task entirely from training data. We believe that the LSTM's non-linear transformations and its strong memory component, **can solve the discrete combinatorial problem of association and is able to replicate the linear assignment problem**



Loss. To measure the misassignment cost, we employ the widely used negative log-likelihood loss

$$\mathcal{L}(\mathcal{A}^i, \tilde{a}) = -\log(\mathcal{A}_{i\tilde{a}}), \qquad (9)$$

where $\tilde{a}$ is the correct assignment and $\mathcal{A}_{ij}$ is the target i to measurement j

Online Multi-target Tracking using Recurrent Neural Networks

Anton Milan[1] Seyed Hamid Rezatofighi[1] Anthony Dick[1] Konrad Schindler[2] Ian Reid[1]

[1]School of Computer Science, The University of Adelaide, Australia
[2]Photogrammetry and Remote Sensing Group, ETH Zürich

Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. 9(8) (November 1997)

IMPLEMENTATION

1) <u>Data augmentation</u>

- by randomly perturbing real data;
- by sampling from a simple generative trajectory model learned from real data;
- 3) by generating physically motivated 3D-world projections.

2) RNN 1 layer with 300 hidden units

LSTM 2 layers with 500 hidden units

Training 30 H CPU

Simulated data

22 video of MOT Challenge

COMPARISON WITH ONLINE METHODS

Table 1. Tracking results on the MOTChallenge training dataset. *Denotes offline post-processing.

Method	Rcll	Prcn	MT	ML	FP	FN	IDs	FM	MOTA	MOTP
Kalman-HA	28.5	79.0	32	334	3,031	28,520	685	837	19.2	69.9
Kalman-HA2*	28.3	83.4	39	354	2,245	28,626	105	342	22.4	69.4
$JPDA_m$*	30.6	81.7	38	348	2,728	27,707	109	380	23.5	69.0
RNN_HA	37.8	75.2	50	267	4,984	24,832	518	963	24.0	68.7
RNN_LSTM	37.1	73.5	50	260	5,327	25,094	572	983	22.3	69.0

Table 2. Tracking results on the MOTChallenge test dataset. *Denotes an offline (or delayed) method.

Method	MOTA	MOTP	FAR	MT%	ML%	FP	FN	IDs	Frag.	FPS
MDP [48]	30.3%	71.3%	1.7	13.0	38.4	9,717	32,422	680	1,500	1.1
$JPDA_m$* [13]	23.8%	68.2%	1.1	5.0	58.1	6,373	40,084	365	869	32.6
TC_ODAL [49]	15.1%	70.5%	2.2	3.2	55.8	12,970	38,538	637	1,716	1.7
RNN_LSTM	19.0%	71.0%	2.0	5.5	45.6	11,578	36,706	1,490	2,081	165.2

Fig. 7. Our RNN tracking results on selected MOTChallenge sequences including ADL-Rundle-3 (first row), TUD-Crossing (second row) and PETS S2.L2 (bottom).

PART II: FRANCESCO SOLERA

Our approach to MTT

ICCV 2015

Learning to Divide and Conquer for Online Multi-Target Tracking

Francesco Solera Simone Calderara Rita Cucchiara

Department of Engineering
University of Modena and Reggio Emilia
name.surname@unimore.it

Abstract

Online Multiple Target Tracking (MTT) is often addressed within the tracking-by-detection *paradigm. Detections are previously extracted independently in each frame and then objects trajectories are built by maximizing specifically designed coherence functions. Nevertheless, ambiguities arise in presence of occlusions or detection errors. In this paper we claim that the ambiguities in tracking could be solved by a selective use of the features, by working with more reliable features if possible and exploiting a deeper representation of the target only if necessary. To this end, we propose an online divide and conquer tracker for static camera scenes, which partitions the assignment problem in local subproblems and solves them by selectively choosing and combining the best*

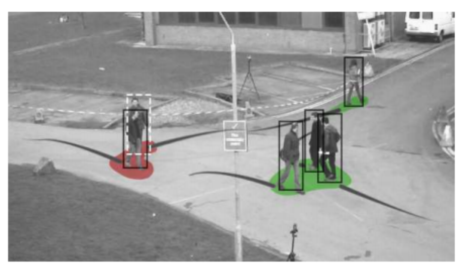


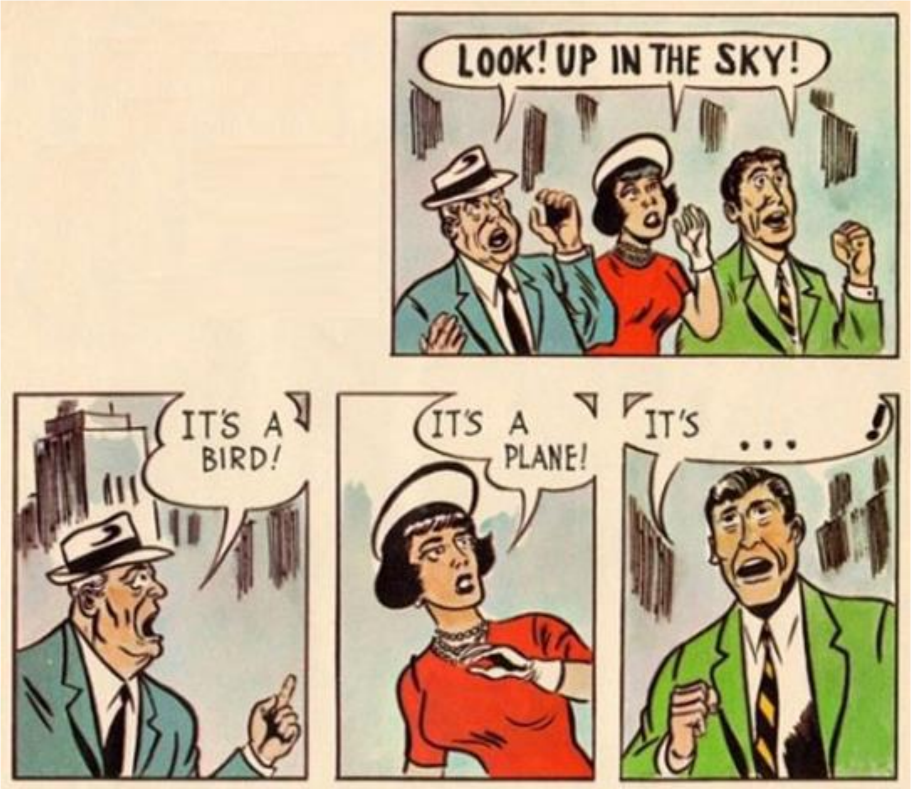
Figure 1: The scene is partitioned in local zones. Green zones is where the same number of tracks and detections are present. Red zones, where miss and false detections (white dashed contours) are discovered and solving the associations may call for complex appearance or motion features.

DRIVING QUESTION: HOW DO WE *HUMANS* TRACK?

Every time we blink, change focus or simply drive in a car, we have to complete a multi target tracking process. How do we do it?

DRIVING QUESTION: HOW DO WE HUMANS TRACK?

Every time we blink, change focus or simply drive in a car, we have to complete a multi target tracking process. How do we do it?



DRIVING QUESTION: HOW DO WE HUMANS TRACK?

Every time we blink, change focus or simply drive in a car, we have to complete a multi target tracking process. How do we do it?

EVOLUTION RULE OF THUMBS: SIMPLE IS BETTER

- Our evolution has thought us to prefer spatial information over surface features (patterns, colors, …) or motion
- It is faster, it is less prone to errors in “feature extraction” step and more reliable
- Position is always meaningful, while other features benefit changes from scene to scene
- You can always refer to more complex features in case of need

OUR IDEA

Human Visual Tracking

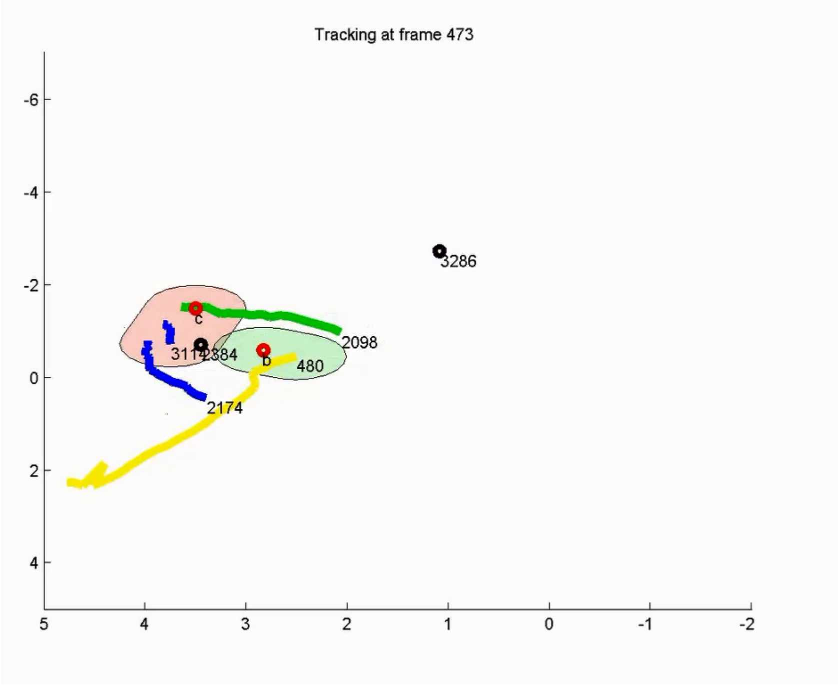
- From **neuroscience** : the ways of where and what
- From **perceptual psychology** : the object file theory

(Kahnemann, Treisman, Gibbs 1995)

- Whenever possible we prefer **distance** only
- Motion prediction and appearance is a plus when useful and safe

Thus? Learning to Divide and Conquer in MTT

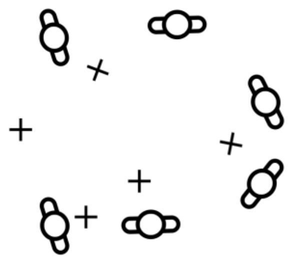
1. Split the crowd in **influence zones** (latent knowledge)
2. Decide whether those **zones are ambiguous** (also latent)
3. Solve unambiguous associations with distance only
4. Employ different level features in ambiguous cases (ask for shapes, color.. edges.. motion)



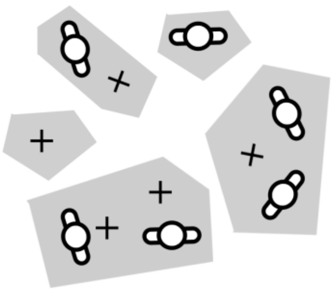
WHAT DOES IT MEAN TO **DIVIDE** THE TRACKING?

At each frame-by-frame association,
split detections and tracks in locally compact clusters.

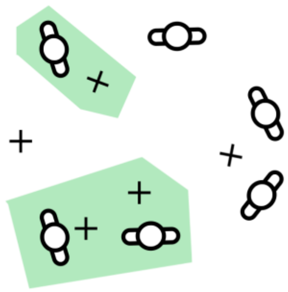
The key idea is that some of these will be really easy to solve!
So easy to solve that spatial information will be enough.



TARGETS and DETECTIONS (+)



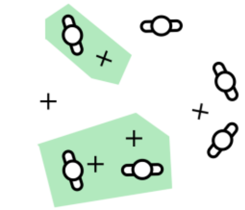
CLUSTERING STEP



FIND EASY TO SOLVE ASSOCIATIONS

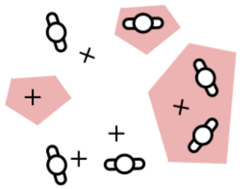
WHAT DOES IT MEAN TO **CONQUER** THE TRACKING?

We define a zone **simple** if it contains an equal number of targets and detections. Associations in simple clusters are solved by using spatial information only.



ASSOCIATE ON
DISTANCE FEATURES
ONLY

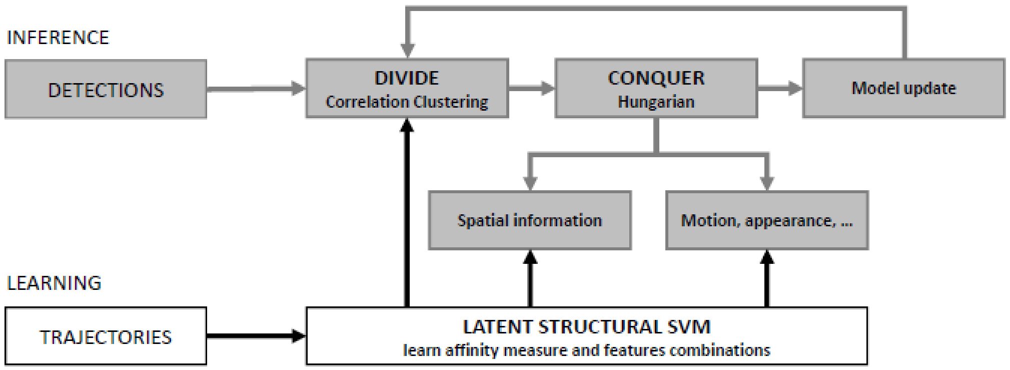
More **complex** (but also more unstable) features, e.g. appearance or motion, are invoked when ambiguity and uncertainty arise in the unassociated zones.



ASSOCIATE ON
DISTANCE AND MORE
COMPLEX FEATURES

WHAT DO WE **LEARN**?

The clustering is cast as latent structured variable, while the association is the output to be predicted. Our method, simultaneously learns in a Latent Structural SVM framework:



- the CLUSTERING AFFINITY MEASURE, needed to split the targets and detections in smaller local association sub-problems;
- the ASSOCIATION COST FUNCTIONS of simple and complex clusters by finding the best weighted combination of simple and complex features respectively.

WHAT IS **GOOD** ABOUT THIS METHOD?

1. Complex features such as appearance or motion **may cause the tracker to drift**. Our method use them only when strictly necessary. In many cases, spatial information turns out to be sufficient.
2. Simple associations are solved independently, so the matching in this local sub-problems can be **computed in parallel**.
3. We **don't fix the clustering scheme**, but learn the affinity measure from examples, since locality may be scene dependent.
4. We also **learn to combine the features** at best to complete the data association step, as different sequences may provide different challenges.
5. Our method is an **extensible framework** – any number of complex features can be added!
6. Overall, the method is **online and fast**. This is thanks to both the smaller sub-problems and the reduced number of calls to complex features extraction.

WHAT IS **BAD** ABOUT THIS METHOD?

- Only sees one new frame at a time (less robust than flow/clique methods)
- Need to re-train for different scenarios
- No moving cameras (is this bad?)

WHAT IS **BAD** ABOUT THIS METHOD?

- Only sees one new frame at a time (less robust than flow/clique methods)
- Need to re-train for different scenarios
- No moving cameras (is this bad?)

LDCT (our)

Sequence	MOTA	MOTP	FAF	GT	MT
TUD-Crossing	67.7	82.9	0.5	13	69.2 %
PETS09-S2L2	47.4	70.8	2.3	42	14.3 %
AVG-TownCentre	31.7	72.2	4.2	226	15.9 %
ADL-Rundle-3	25.2	73.4	0.7	44	4.5 %
KITTI-16	53.0	79.0	0.4	17	11.8 %
Venice-1	33.5	68.4	1.3	17	0.0 %

RNN_LSTM

Sequence	MOTA	MOTP	FAF	GT	MT
TUD-Crossing	57.2	71.7	0.4	13	30.8 %
PETS09-S2L2	38.3	71.6	2.3	42	9.5 %
AVG-TownCentre	13.4	68.8	2.7	226	3.5 %
ADL-Rundle-3	23.7	72.0	3.5	44	6.8 %
KITTI-16	26.3	68.5	1.4	17	0.0 %
Venice-1	12.7	71.7	1.5	17	0.0 %

MOTChallenge static camera test sequences…

PART III: Francesco SOLERA

Multi Camera Tracking

with some help from social groups

Trans. On CSVT 2016, collaboration with Duke University

Tracking Social Groups Within and Across Cameras

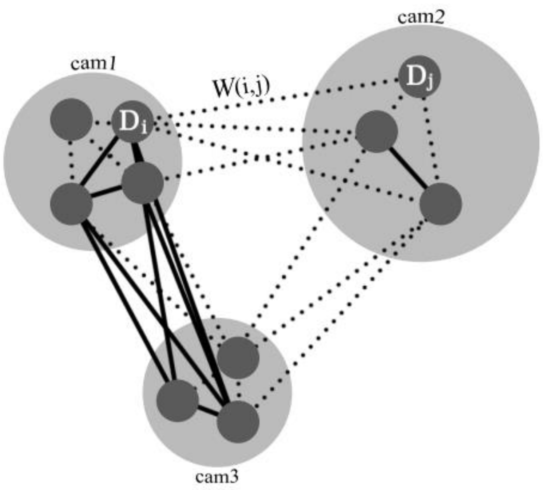
Francesco Solera, Simone Calderara, Ergys Ristani, Carlo Tomasi, Rita Cucchiara

***Abstract*—Groups are considered by modern sociological crowd theories the atomic entities where social processes arise and develop. In computer vision, group analysis has gained momentum only recently due to the complexities of the group detection task in real life scenarios. In this context, the conventional people tracking problem can be re-instantiated considering groups playing a central role in the process. Thus, we propose a method for solving the group tracking problem seamlessly on single and multiple disjoint cameras. Our formulation follows the tracking by detection paradigm where groups are the atomic entities and are linked along time forming long and consistent trajectories. To this aim, we formulate the problem as a supervised clustering problem where a Structural SVM classifier is used to learn a proper similarity measure among such group entities. Multi-Camera group tracking is handled inside the framework by adopting an orthogonal feature encoding allowing the classifier to learn differently inter and intra cameras features weights. Experiments were carried out on a novel annotated data set of**

Fig. 1: An example of groups detected in the four different cameras of the proposed data set `DukeChapel-Groups`.

HOW DO WE GO FROM SINGLE TO MULTI-CAMERA?

- Same way we go:
  1. from detections to tracklets and
  2. from tracklets to trajectories

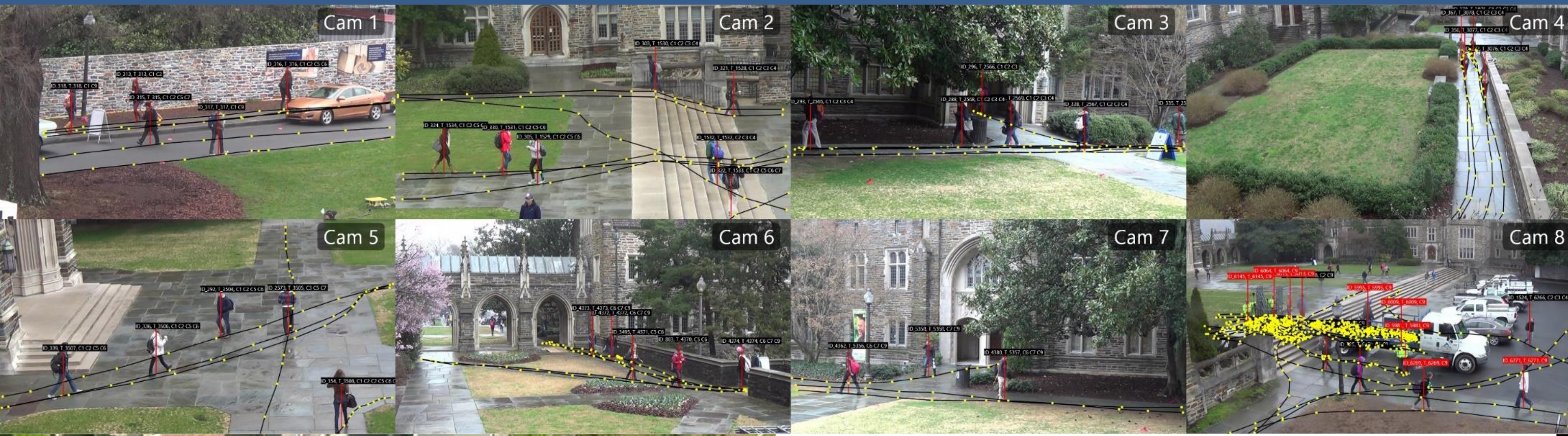


(a) Group clustering at T_k

We want to cluster together trajectories belonging to the same person.

HOW DO WE GO FROM SINGLE TO MULTI-CAMERA?

DUKE DATASET



Different camera placements:

1. which features are best to track in a specific camera?
2. which features are best to associate between two specific cameras?

FEATURE IMPORTANCE IS STRONGLY INFLUENCED BY CAMERA SETTING

HOW DO WE GO FROM SINGLE TO MULTI-CAMERA?

DUKE DATASET

Above all:
people placement inside social groups change from camera to camera

Different camera placements:

1. which features are best to track in **a specific** camera?
2. which features are best to associate between **two specific** cameras?

FEATURE IMPORTANCE IS STRONGLY INFLUENCED BY CAMERA SETTING

CAN WE EXPLOIT IT INSTEAD OF SUFFERING FROM IT?

If we can recognize the group to which a pedestrian belongs to:

- we can **stop tracking singletons** and
- **start tracking groups**! (at least, until they split)

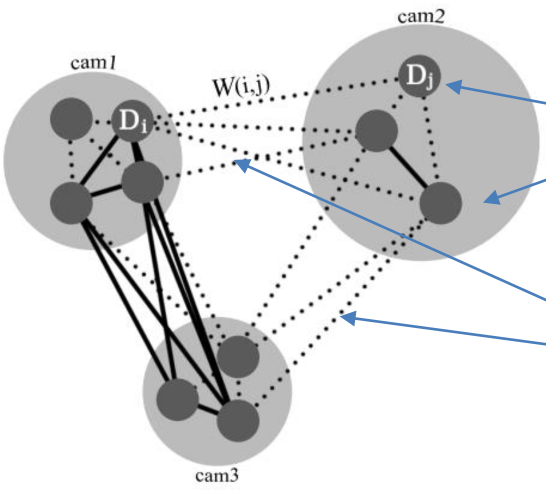
CAN WE EXPLOIT IT INSTEAD OF SUFFERING FROM IT?

If we can recognize the group to which a pedestrian belongs to:

- we can **stop tracking singletons** and
- **start tracking groups**! (at least, until they split)

PEOPLE DETECTOR → GROUP DETECTOR

PAMI 2015



some of this detections will be pedestrians others will be groups

weights on edges are learnt based on cameras involved in association:

- similar viewpoint -> appearance important
- different viewpoint -> motion/time reasoning is better

SOME EXAMPLES…

PEOPLE TRACKING

GROUP TRACKING

FINAL CONCLUSIONS

1) **There is no conclusion to tracking problem** (at least for NOW)

   - it is hard

   - it comprises different sub-problems

2) Many approaches for MTT

   GNN, JPDA, MHT, DL

   a large area of converging research

3) People detection and re-identification is always trendy

4) Many domain specific contexts that are interesting (look at the egocentric ones, automotive…)

5) A VERY VERY HOT RESEARCH AREA

http://imagelab.ing.unimo.it

Centro del
Tecnopolo di Modena
Rete ad Alta Tecnologia
Regione Emilia Romagna

Prof. Costantino Grana, *PhD*
Associate Professor
Dipartimento di Ingegneria "Enzo Ferrari", Modena, Italy

Ing. Roberto Vezzani, *PhD*
Assistant Professor
Dipartimento di Ingegneria "Enzo Ferrari", Modena, Italy

Ing. Simone Calderara, *PhD*
Assistant Professor
Dipartimento di Ingegneria "Enzo Ferrari", Modena, Italy

Ing. Giuseppe Serra, *PhD*
Post-graduated Grant
Dipartimento di Ingegneria "Enzo Ferrari", Modena, Italy

Dr. Marco Manfredi, *PhD*
Post-Doc Research Grant
Dipartimento di Ingegneria "Enzo Ferrari", Modena, Italy

Dr. Martino Lombardi
PhD Student
Dipartimento di Ingegneria "Enzo Ferrari", Modena, Italy

Dr. Francesco Solera
PhD Student
Dipartimento di Ingegneria "Enzo Ferrari", Modena, Italy

Dr. Lorenzo Baraldi
PhD Student
Dipartimento di Ingegneria "Enzo Ferrari", Modena, Italy

Dr. Stefano Alletto
PhD Student
Dipartimento di Ingegneria "Enzo Ferrari", Modena, Italy

Dr. Francesco Paci
PhD Student
Dipartimento DEI, Unibo, Modena, Italy

Dr. Fabrizio Balducci
PhD Student
Dipartimento di Ingegneria "Enzo Ferrari", Modena, Italy

Dr. Guido Borghi
PhD Student
Dipartimento di Ingegneria "Enzo Ferrari", Modena, Italy

Dr. Patrizia Varini
PhD Student
Dipartimento di Ingegneria "Enzo Ferrari", Modena, Italy

Dr. Andrea Corbelli
Research Grant
Dipartimento di Ingegneria "Enzo Ferrari", Modena, Italy

Dr. Andrea Palazzi
Research Grant
Dipartimento di Ingegneria "Enzo Ferrari", Modena, Italy

Ing. Paolo Santinelli
Research Fellow
Dipartimento di Ingegneria "Enzo Ferrari", Modena, Italy

Ing. Augusto Pieracci
Research Fellow
Dipartimento di Ingegneria "Enzo Ferrari", Modena, Italy