Human Behavior understanding in cars and around

Prof. Ing. Rita Cucchiara

Imagelab,
Dipartimento di Ingegneria «Enzo Ferrari»,
Università di Modena e Reggio Emilia,

Director of the
Research Center in ICT, Softech-ICT
Modena,Italy

Invited talk
Venezia , 4/11/2016
@ECLT Ca'Foscari

UNIMORE Presentation

- UNIMORE Università di Modena e Reggio Emilia, Italy
- funded in 1175, in 2 cities:
- Modena (11 Departments, 14.100 students)
- Reggio Emilia (3 Departments, 5.400 students)

- Dipartimento di Ingegneria «Enzo Ferrari», Modena
- 6.000 students, c.a.100 faculties, 7 curricula
- Mechanical Engineering, Vehicle Engineering, Material engineering
- Computer Engineering, Electronic Engineering,
- Civil and Environmental Engineering
  - 2 International Phd Curricula (High Mechanics, **ICT**)
  - 2 High Technology Network Centres: Intermech and Softech-ICT
  - 2 Master in ICT: Vision Learning and Multimedia Technology (MUMET 2017), Cybersecurity Academy

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

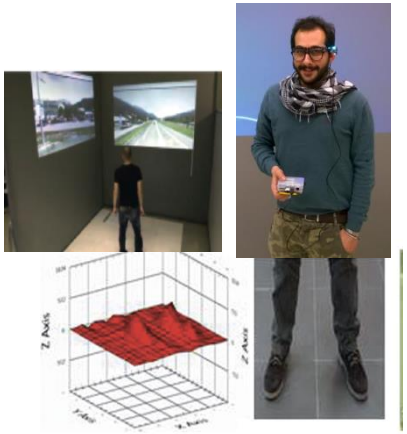
Imagelab : research since 1998

www.imagelab.unimore.it

- Pattern recognition and Image processing
- Medical Imaging
- Digitalized Document analysis

- Multimedia
- Multimedia big data annotation
- Video captioning

- 2D, 3D, wearable Computer vision
- Augmented experiences in culture and museums
- Experience with Wearable devices, floors and IoT



- Computer vision for Behavior analysis
- Children and people behavior analysis
- Surveillance (in crowd)
- Automotive driver behavior understanding

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Computer Vision and Human Behavior Understanding

in cars and around

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

..with Computer Vision

- **Computer Vision** is the scientific discipline studying **how to perceive and understand the world through visual data by computers.**

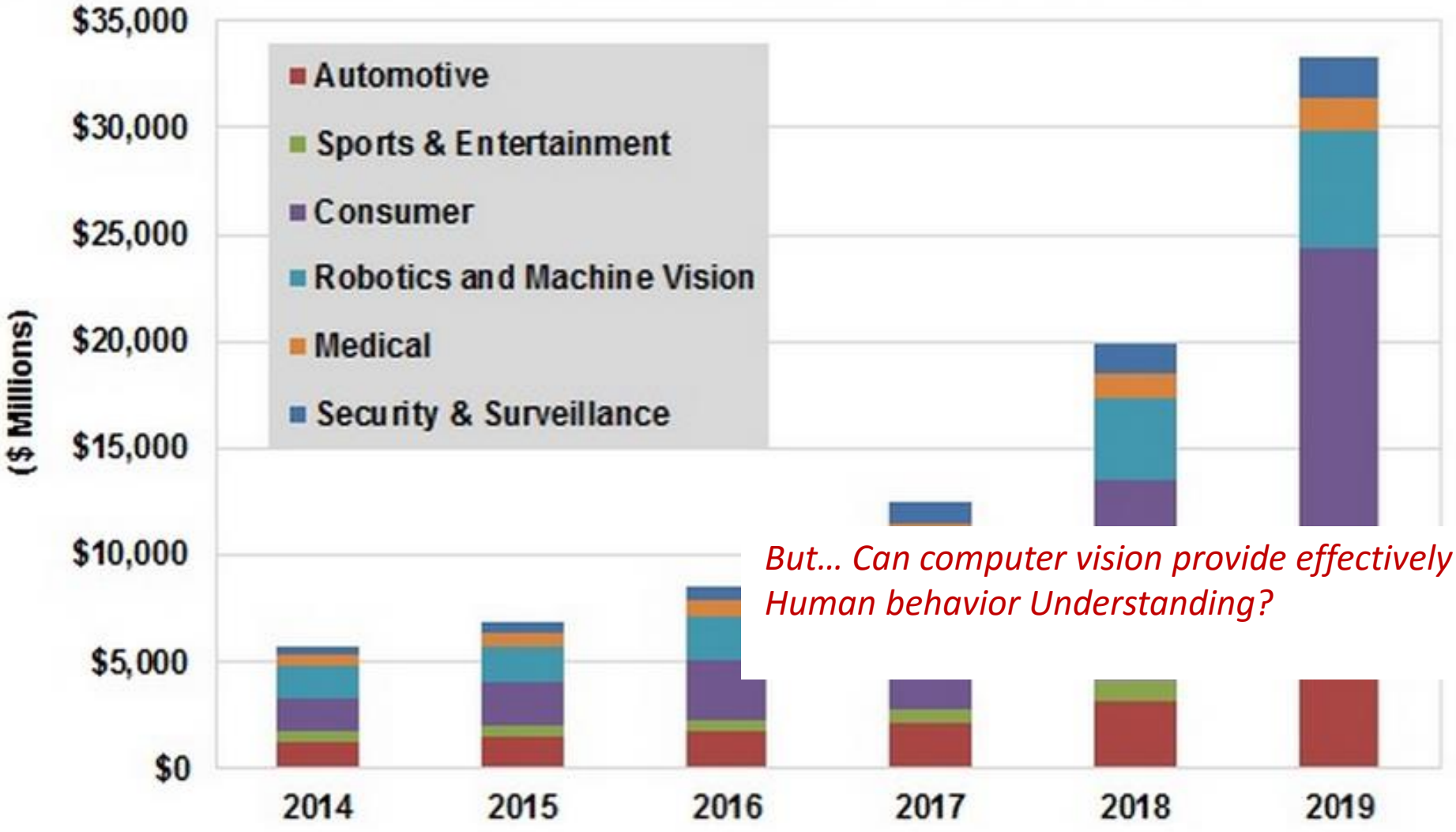
- **Pattern Recognition** is the ensemble of theories, models , techniques **to recognize patterns in unknown or unordered data, generally images and multimedia content.**

- **Machine Learning** is the science of **getting computers to act without being explicitly programmed**

Computer Vision & Machine Learning *represent now the most advanced frontier of Artificial Intelligence studies*

Tractica

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



But… Can computer vision provide effectively Human behavior Understanding?

Source: Tractica

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Human behaviors

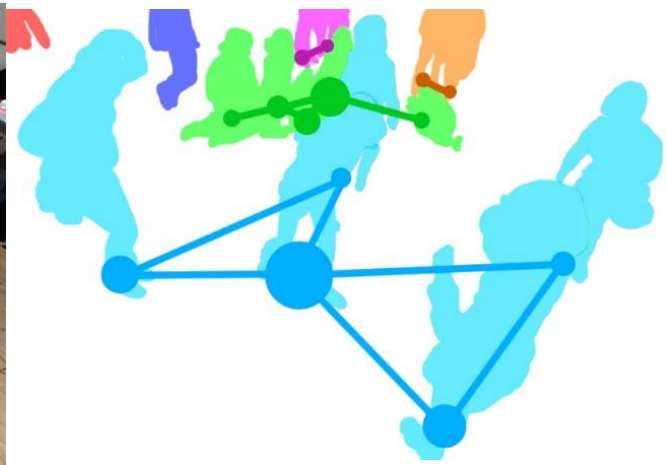
- What he is doing?
- What are they doing?

MIUR-FAR project
DriveAttention
2016-2018

- Single and collective behavior
- Collaborative or not collaborative behaviors

PRIN project
“2106-2018

project “VAEX 2015-2018



MIUR project “the educating city” cluster smart city 2015-2018

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Understanding Human behavior by video

- Why:

- To support sociologists' and psychologists' work

  (e.g. education, social interaction..)

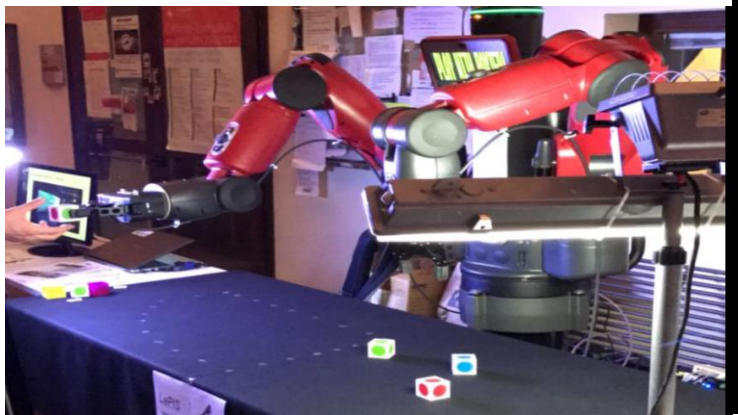
  to understand humans

- To support computers' work

  in on-line or off-line knowledge extraction about humans

  for a huge number of applications, services and systems

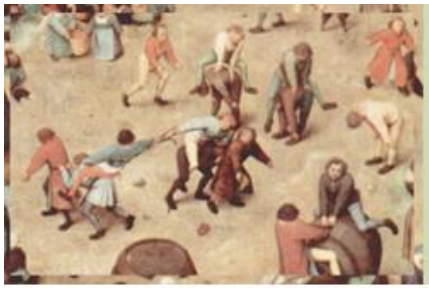
  (surveillance, HCI, automotive, augmented experiences..)



UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

HBU: a long story

- 1997-2000 MIT Alex Pentland: PFINDER projects and understanding interactions
- 2006- datasets for action analysis (Weizmann ICCV2005), action understanding now is popular*
- 2010- 7 workshops on **HBU** (from 2010: IAPR, AMI, IROS, ACM MM, ECCV, ACMMM2016)
- **2011- Chalearn** workshops 2011- 2016; CVPR 2016 challenge "Looking at People"
- Many many datasets…



7TH INTERNATIONAL WORKSHOP ON HUMAN BEHAVIOR UNDERSTANDING

behavior analysis & multimedia for children

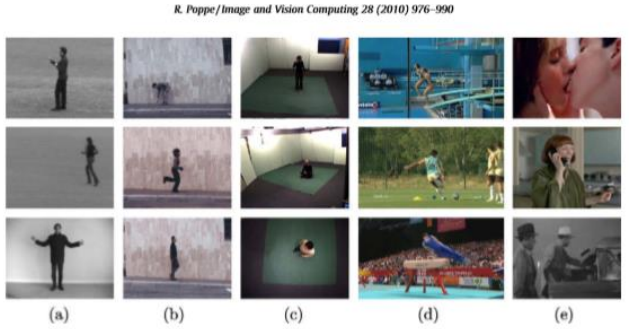


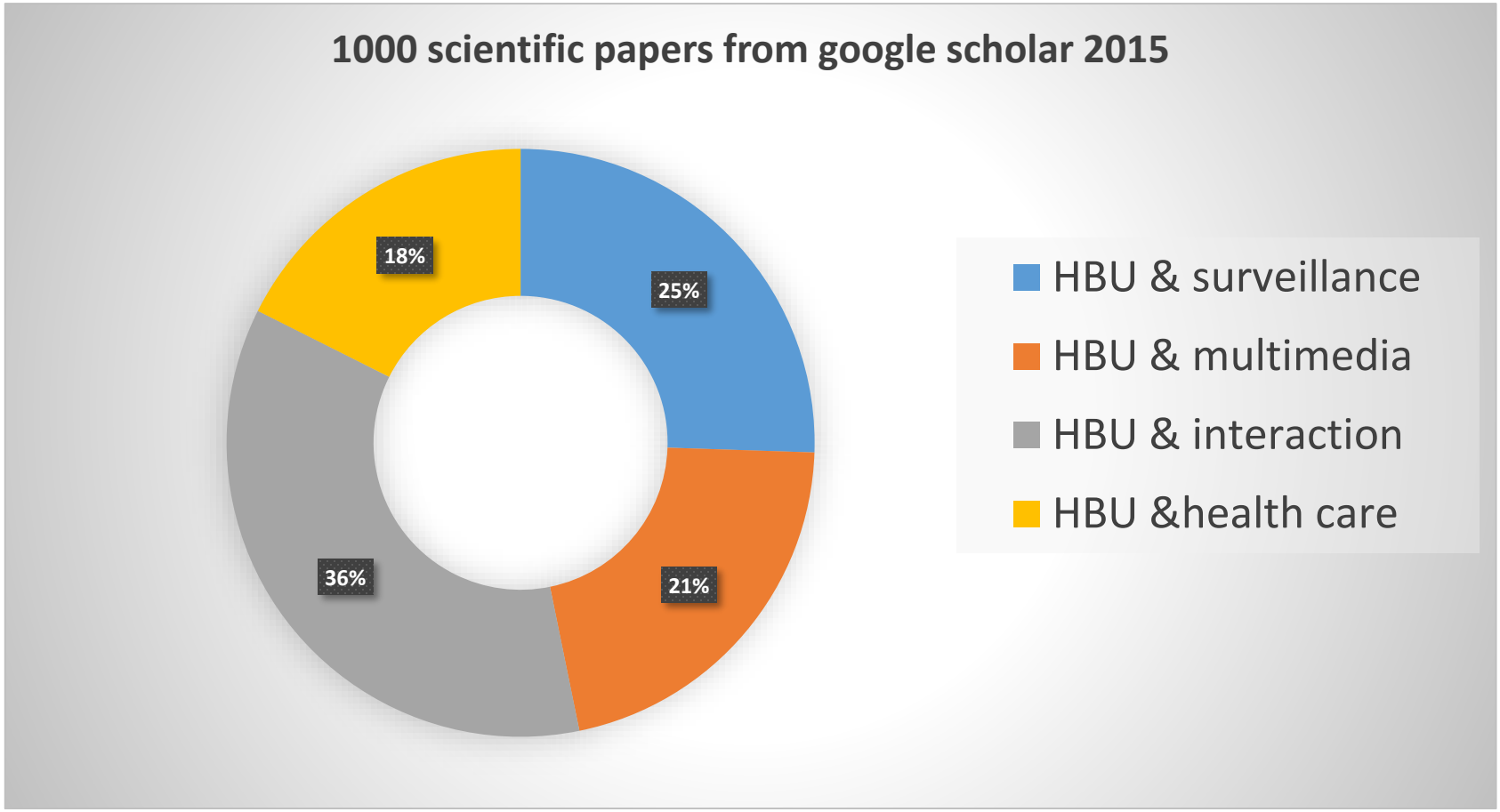
Fig. 1. Example frames of (a) KTH dataset, (b) Weizmann dataset, (c) Inria XMAS dataset, (d) UCF sports action dataset and (e) Hollywood human action dataset.

- One for all: * R. Poppe "A survey on vision based action recognition" Image and vision computing 2010

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Research in HBU

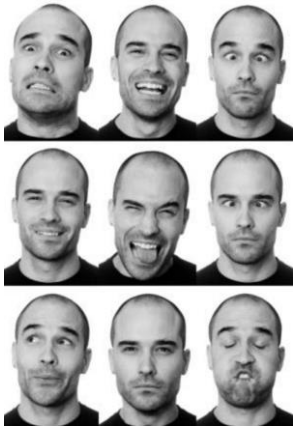
- HBU by vision:
- More than 1000 papers from 2010 to 2015
- In 2016? Surely more



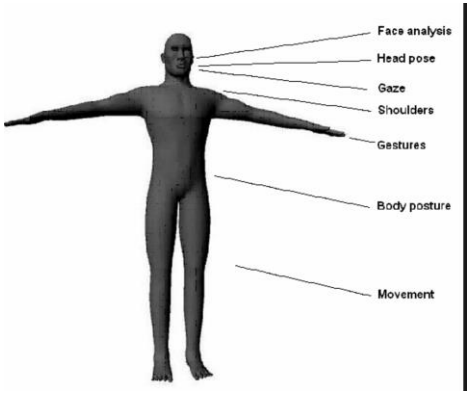
UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

HBU, what knowledge can be extracted by video?

Movements
- Body movements
- Gestures
- Poses
- Gaze and Expression



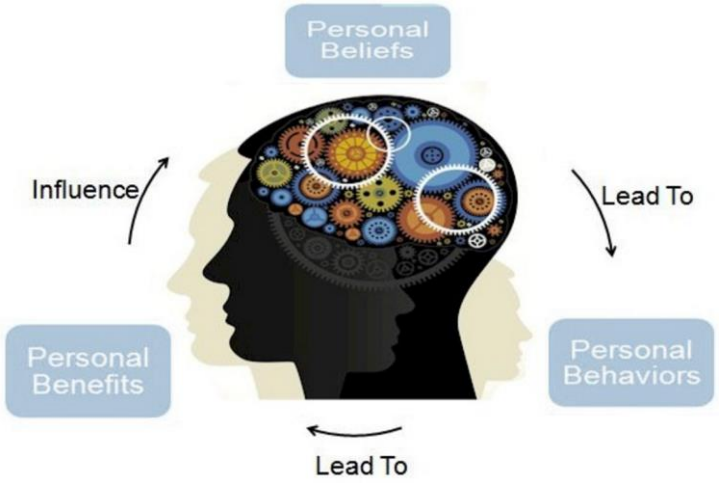
Actions



Activities

Bheavior?

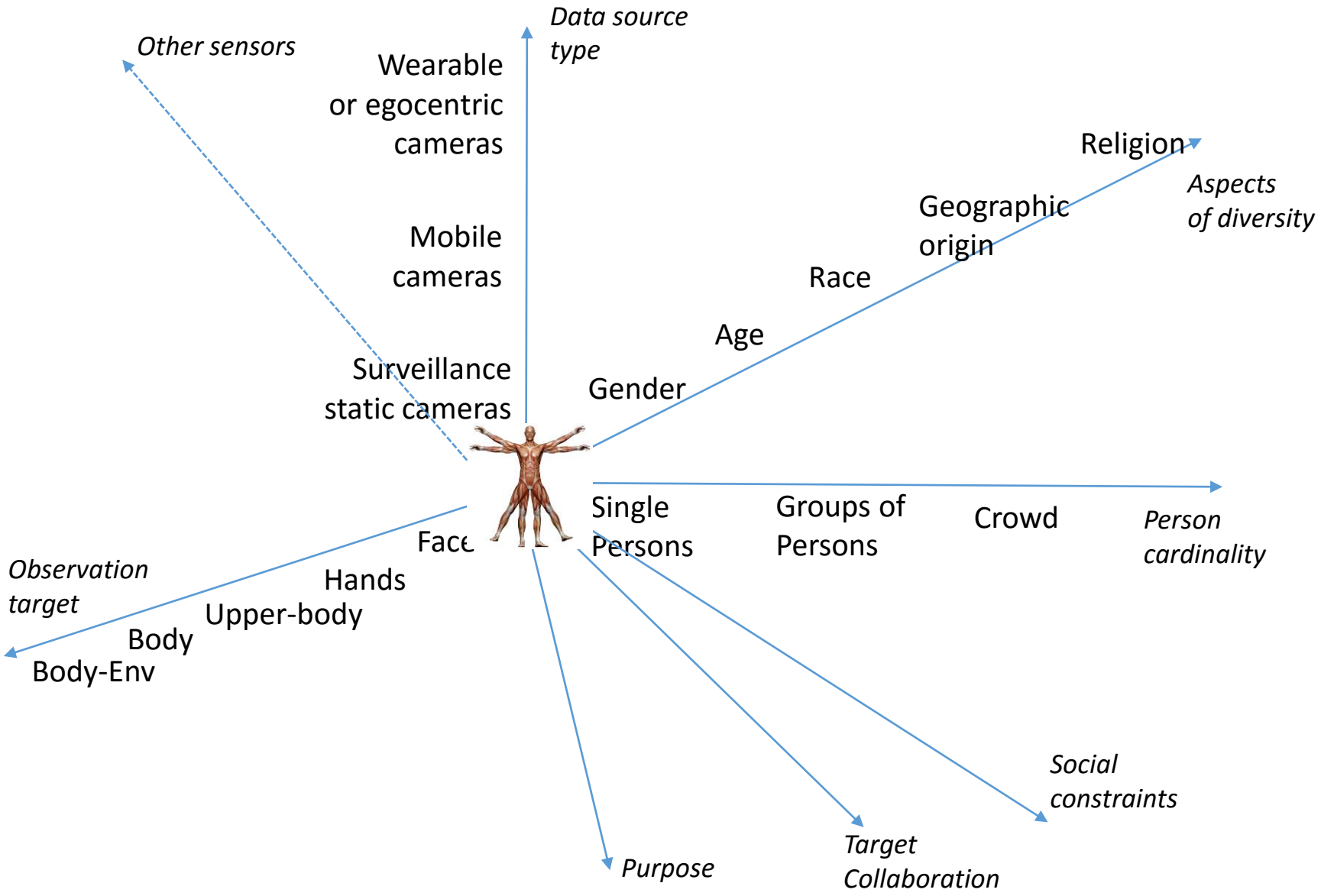
- Behavior
- $\{movements, actions, activities\} + \{environment, objects, people\} + \{purposes, beliefs, habits\}$



- Self-behavior, Personal behavior, Social behavior, …

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

HBU by Vision, a multidimensional space



the levels of details vs the level of sociality

- Levels of details

Human parts

Human Full body

Human(s) In the environemnt

Humans In crowd

Lev3 Social Behavior

Lev1 self Behavior expression, gesture, pose

Lev2 whole person Behavior Action and interaction

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Lev 1: Expressions, poses and gestures

Understanding humans for New natural Human Computer interaction systems *

What are they doing?

Deaf Sign language

Picture from Benjamin Lewis UCLA

Maja Pantic, Alex Pentland, Anton Nijholt and Thomas Huang uman Computing and Machine Understanding of Human Behavior: A Survey ICMI 2006

S. Rautaray, A Agrawal Vision based hand gesture recognition for human computer interaction: a survey Artificial Intelligence Review January 2015,

….

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

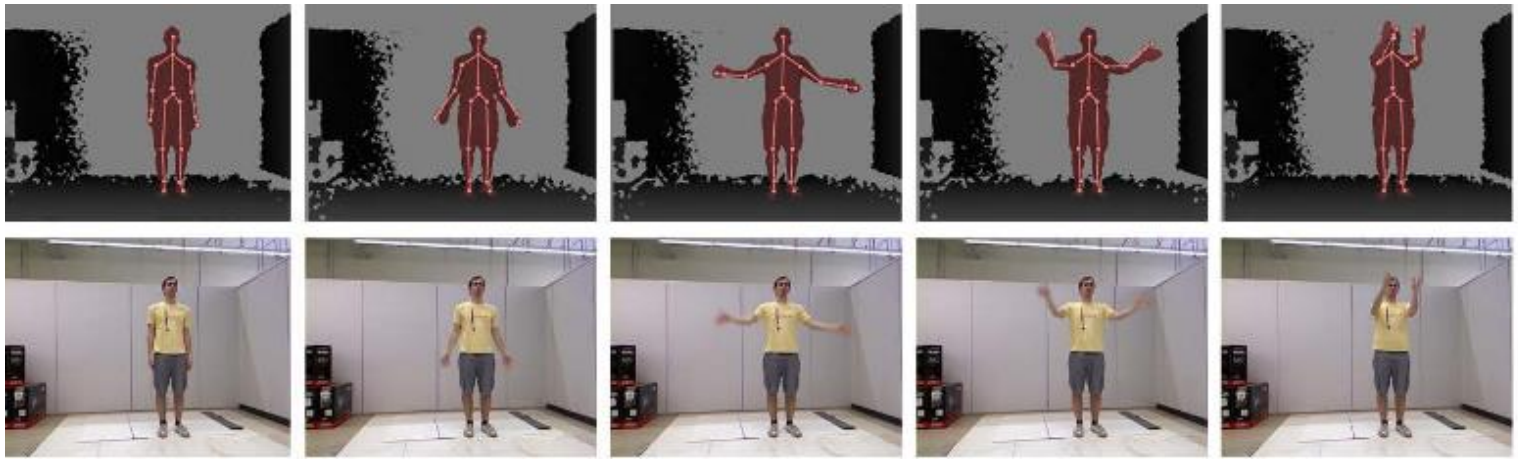
Lev 1: Expressions, poses and gestures

- Human attention in the car

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Lev2: Body actions

Fig. 7. Sample input frame of the Weizmann dataset



R.Vezzani,D.Baltieri, R.Cucchiara HMM Based Action Recognition with Projection Histogram Features ICPRW2010 supported by EU THIS Project

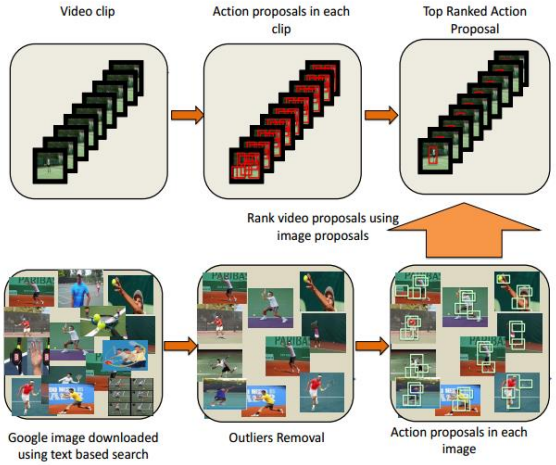
G.Borghi, R. Vezzani, R.Cucchiara; "Fast gesture recognition with Multiple StreamDiscrete HMMs on 3D Skeletons"*Proceedings of the 23rd International Conference on Pattern Recognition*, Cancun, Dec 4-8, 2016, 2016

Lev 2 Action again, new trends

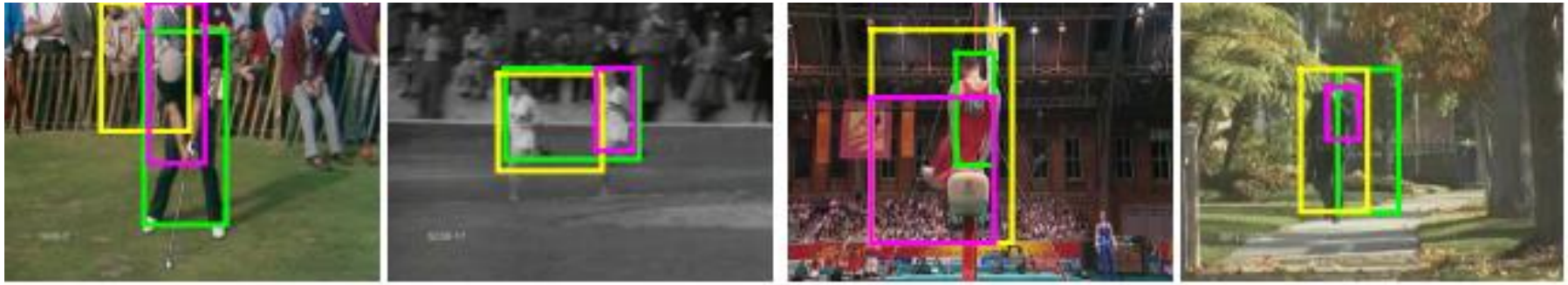
- CVPR2016

What if we do not have multiple videos of the same action? — Video Action Localization Using Web Images

Waqas Sultani, Mubarak Shah



- Graph representation on DL features
- Graph based optimization
- Probabilistic Hough Matching for proposals
- Optimization in superpixel
- Image and video proposal



Lev 2: Action Again, new trends

- CVPR2016

VLAD^3: Encoding Dynamics of Deep Features for Action Recognition

Yingwei Li† Weixin Li† Vijay Mahadevan§ Nuno Vasconcelos†

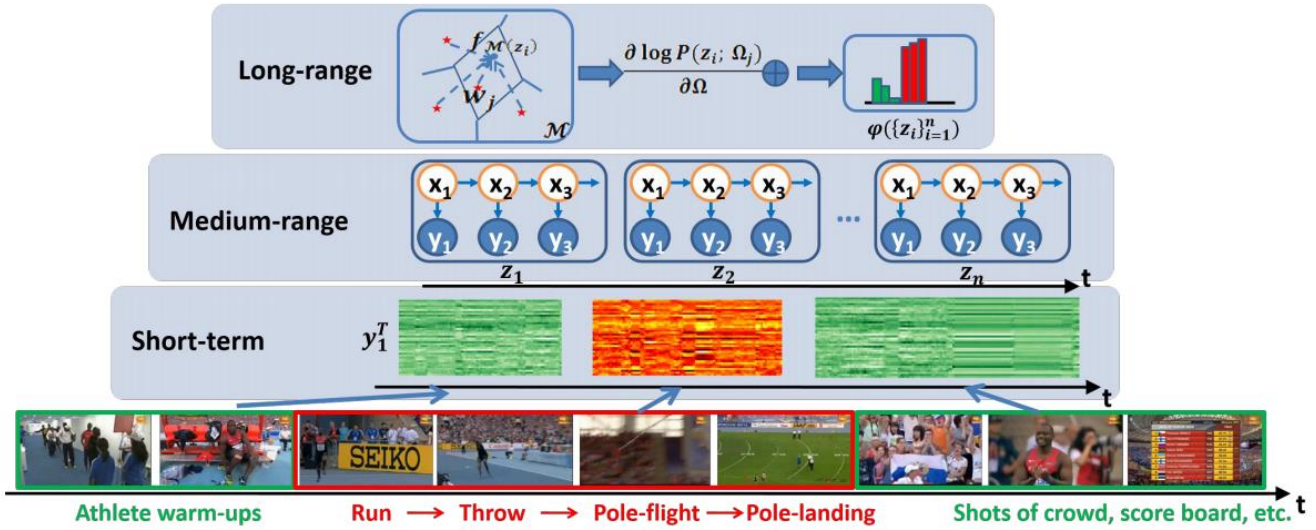
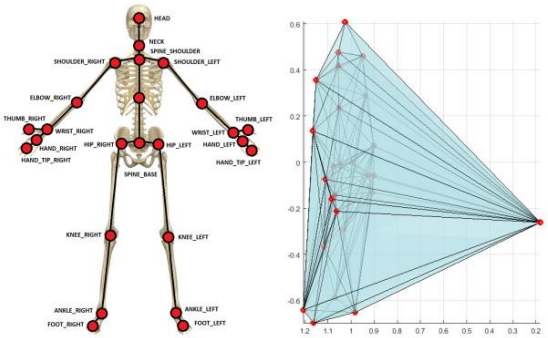


Figure 1: The VLAD^3 is inspired by the hierarchical structure of video dynamics. A **short-term** stage captures short-term appearance and motion patterns with deep features. A **medium-range** stage models the dynamics of segments of deep features, using an LDS. Finally, a **long-range** stage computes and pools a VLAD descriptor, derived from the LDS.

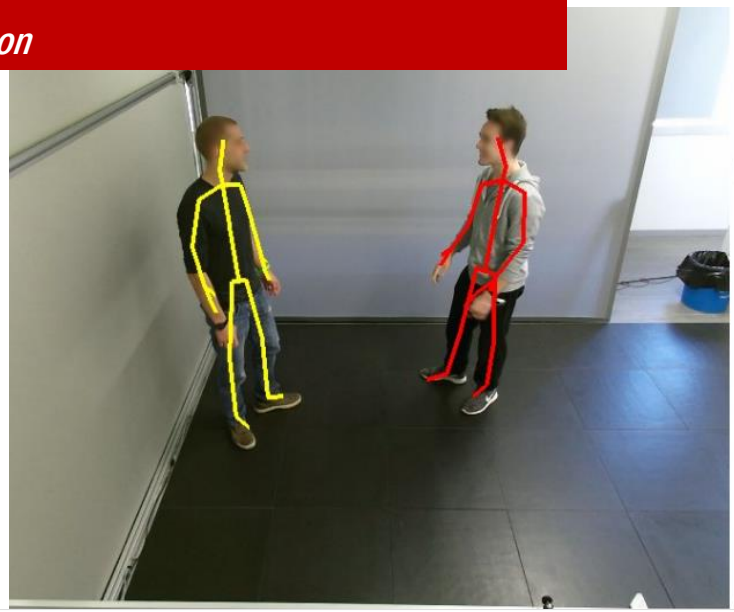
UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

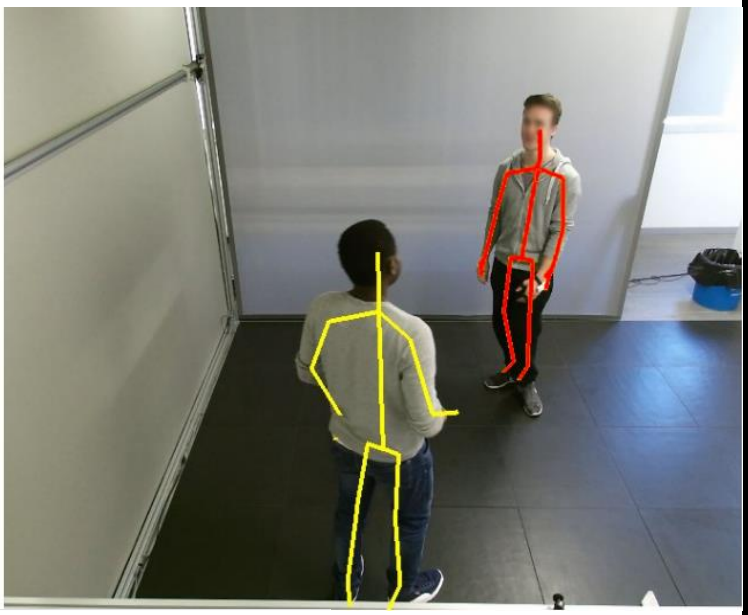
Lev 2 Action, Interaction, to behavior

Thanks to Andrea Palazzi and Simone Calderara @Imagelab



Measuring human interaction





Palazzi,A.; Calderara,S.; Bicocchi,N; Vezzali,L.; di Bernardo,G.; Zambonelli,F; Cucchiara, R.["Spotting prejudice with nonverbal behaviours"] *ACM International Joint Conference on Pervasive and Ubiquitous Computing*, Heidelberg, Settembre 2016,

IAT (Implicit Association Test)

Thanks to Andrea Palazzi and Simone Calderara @Imagelab

- Simple but effective CV features

Volume distance

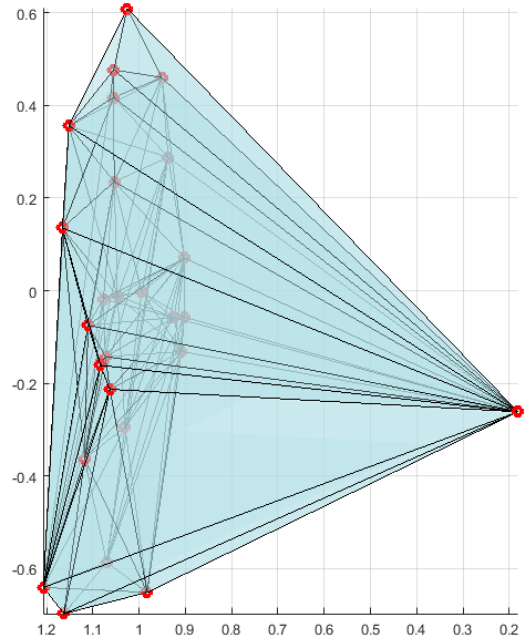
$$F_{vol}(f) = Vol(DT(P(f)))$$

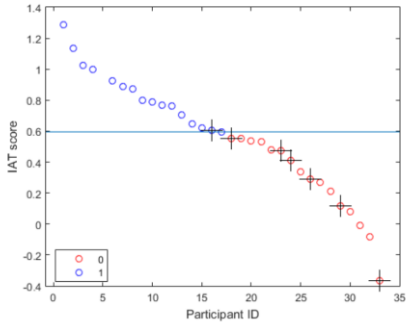
$DT(P(f))$ = Delaunay Triangolation

Mutual distance

$$D_m(f) = dist(C^p, C^c)$$

$$C^p(f) = \frac{1}{m}\sum_{i=1}^{m} joint_i^p(f) \qquad C^c(f) = \frac{1}{m}\sum_{i=1}^{m} joint_i^c(f)$$

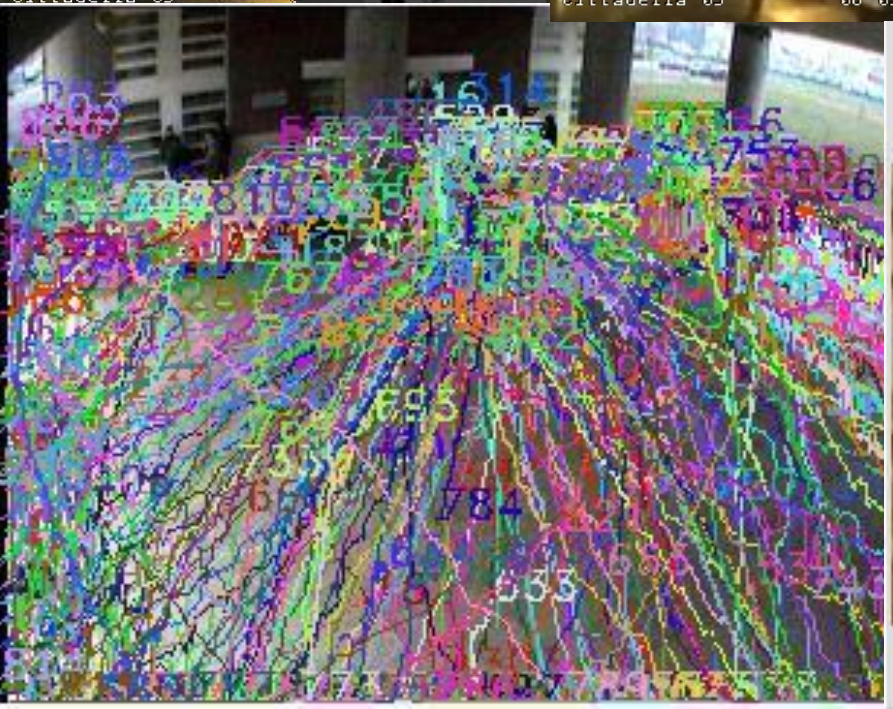




Lev. 3 social behavior: people in the environment

- What are they doing?

Real-time surveillance



We need

- Detection
- Tracking (single , MTT, MTT in Multiple cameras..)
- Understanding movements, actions, activity, behaviors and social relation

Alex Prager Crowd #8 (city Hall) 2013

A LONG STORY, AGAIN

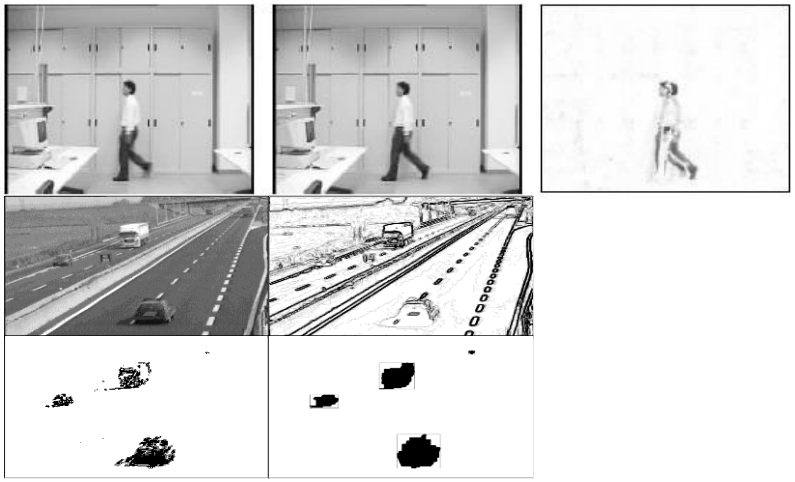
Segmentation/detection in surveillance

?

Segmentation by motion

Detection with learning and classifiers

Differential Motion



Background suppression

R. Cucchiara, C. Grana, M. Piccardi, A. Prati "Detecting Moving Objects, Ghosts and Shadows in Video Streams", IEEE *Trans on PAMI,* 2003

HOG, part based, CNNs People detectors

End-To-End People Detection in Crowded Scenes
Russell Stewart, Mykhaylo Andriluka, Andrew Y. Ng CVPR2016

Tracking (few) people

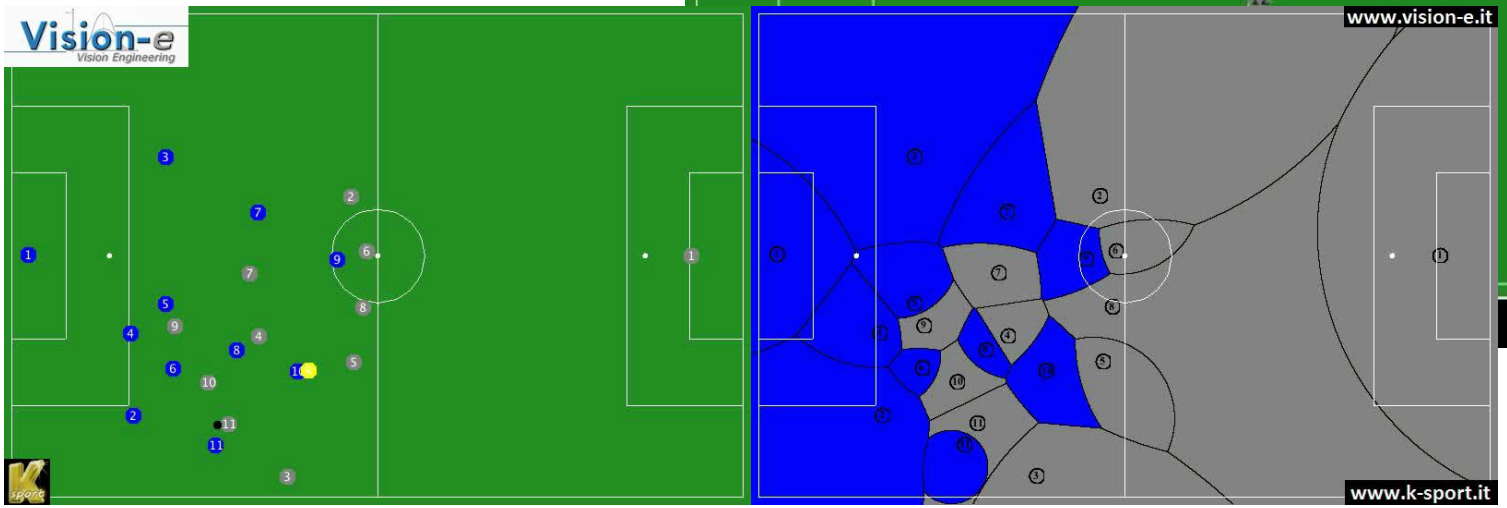
- Tracking few people in a constrained environment: «solved problem»

S. Calderara, R. Cucchiara, A. Prati, "Bayesian-competitive Consistent Labeling for People Surveillance"in IEEE Trans. On PAMI , 2008

Detection and tracking in commercial systems

Tracking by detection: using people detection for initialize ROI-based tracking (eg particle filter)

In semi-constrained world
Tracking is possible



UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

..and tracking single (people) target

- **Is tracking a solved problem?**

- We tried to answer this questions in an "**experimental evaluation**"
- Even in case of single target tracking*

- - a very large dataset
- of 14 categories of challenges

315 video ALOV++
http://www.alov300.org
http://imagelab.ing.unimo.it/dsm

- - a large set of performance measures

MOTA; OTA; Deviaton….
F-Measure
SURVIVAL CURVES..

- - a large experimentation
- (with code available over 3 clusters in 3 labs)

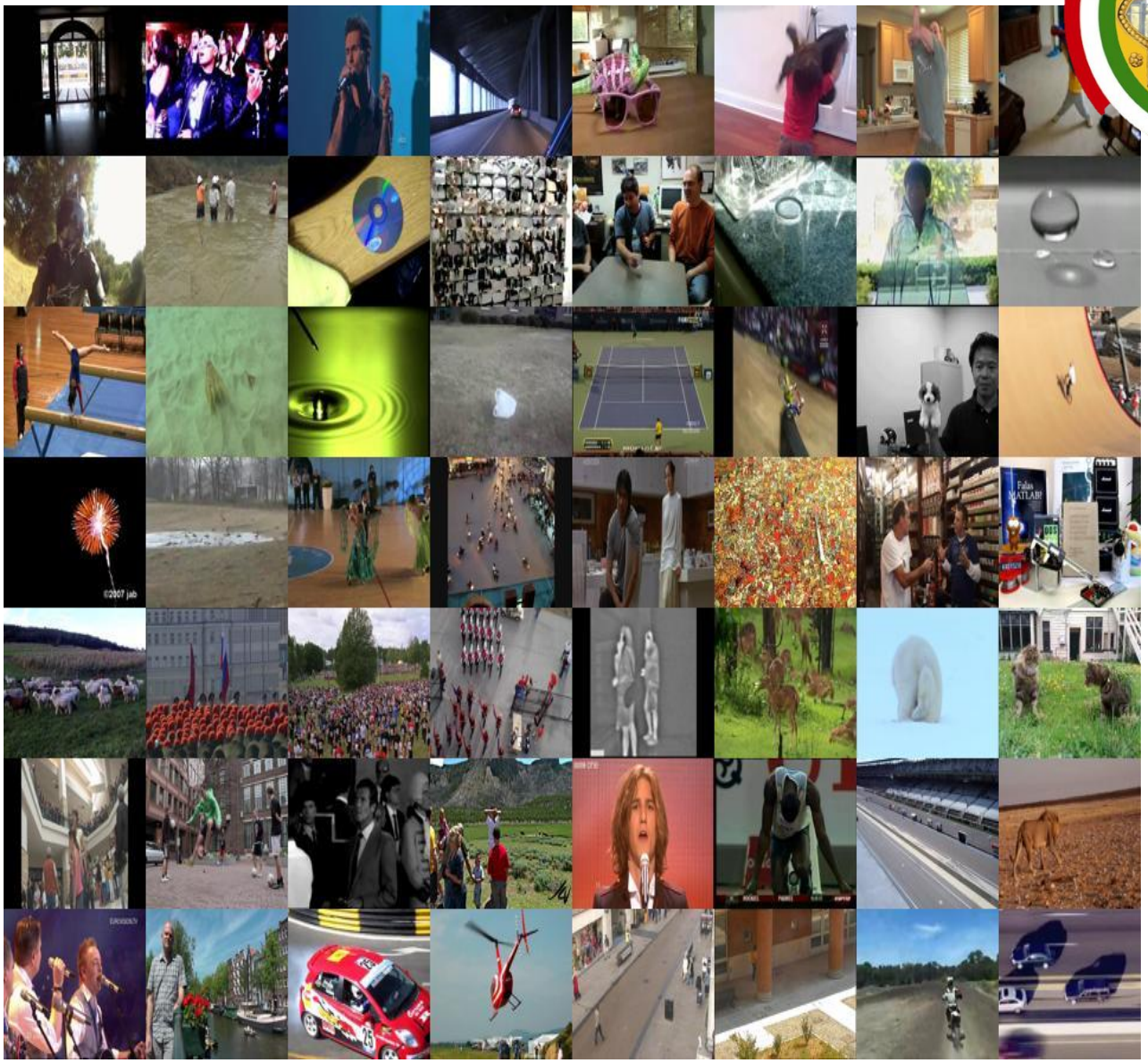
19 trackers
BASELINES
STATE OF THE ART

* *D.Chu, A.Smeulders, S.Calderara, R.Cucchiara, A. Dehghan, M.Shah* **Visual Tracking: an Experimental Survey** *Transactions on PAMI 2013*

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

14 tracking challenges in 313 videos

- 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

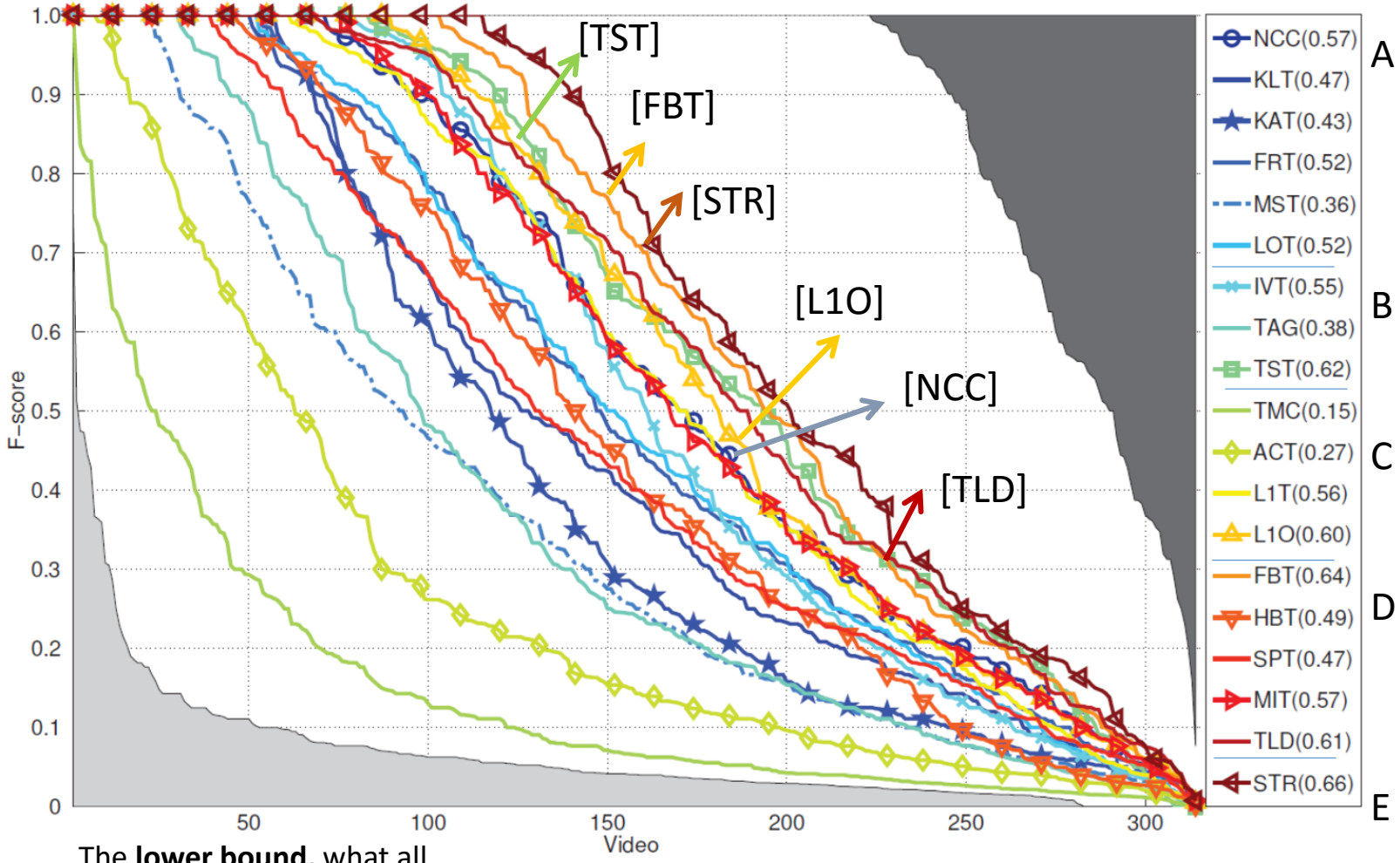


UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

A comprehensive view Survival curve

About the 30%, correctly tracked only

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

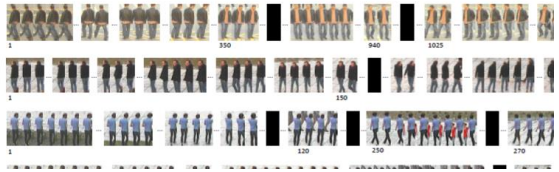


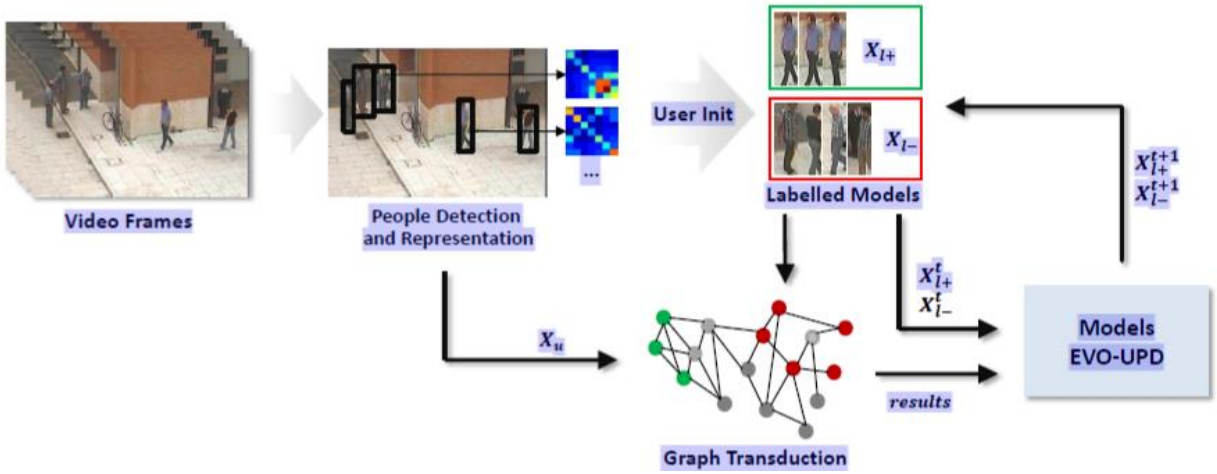
The **lower bound,** what all trackers can do 7%

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

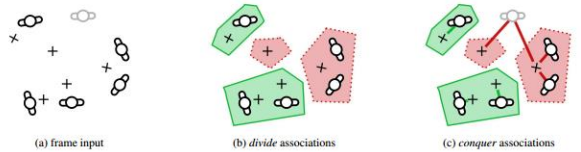
Multiple Target Tracking

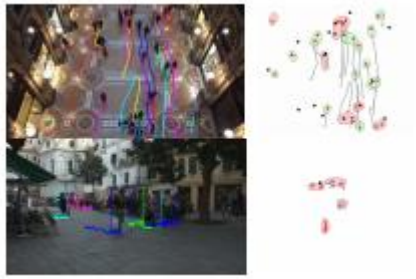
Tracking by detection with transductive learning *





Tracking by detection with structural SVM **





*D. Coppi, S.Calderara, R. Cucchiara “Transductive People Tracking in Unconstrained Surveillance” Transactions on CSVT 2015

** Francesco, Solera; Simone, Calderara; Rita, Cucchiara "Learning to Divide and Conquer for Online Multi-Target Tracking" Proceedings of ICCV 2015

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Understanding anomalous behavior

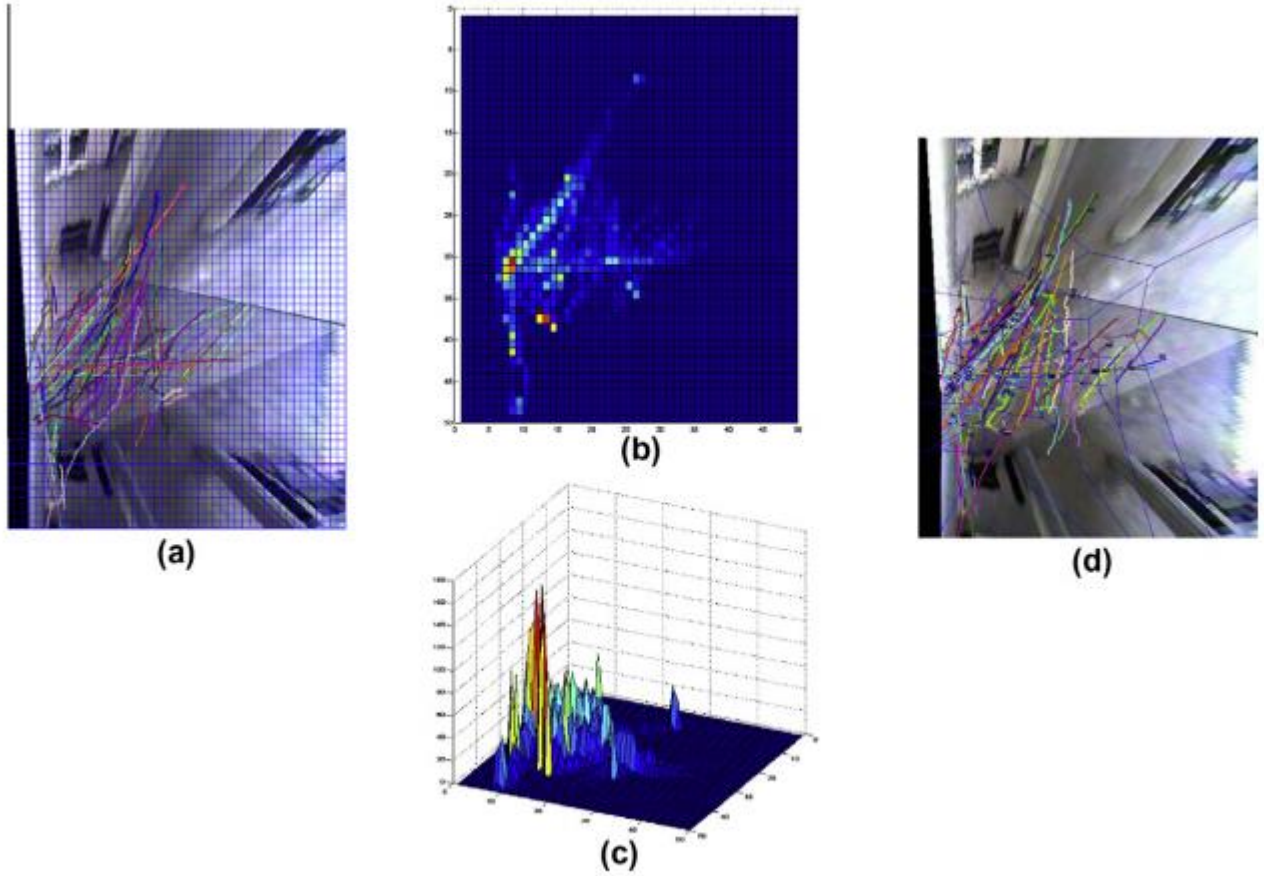
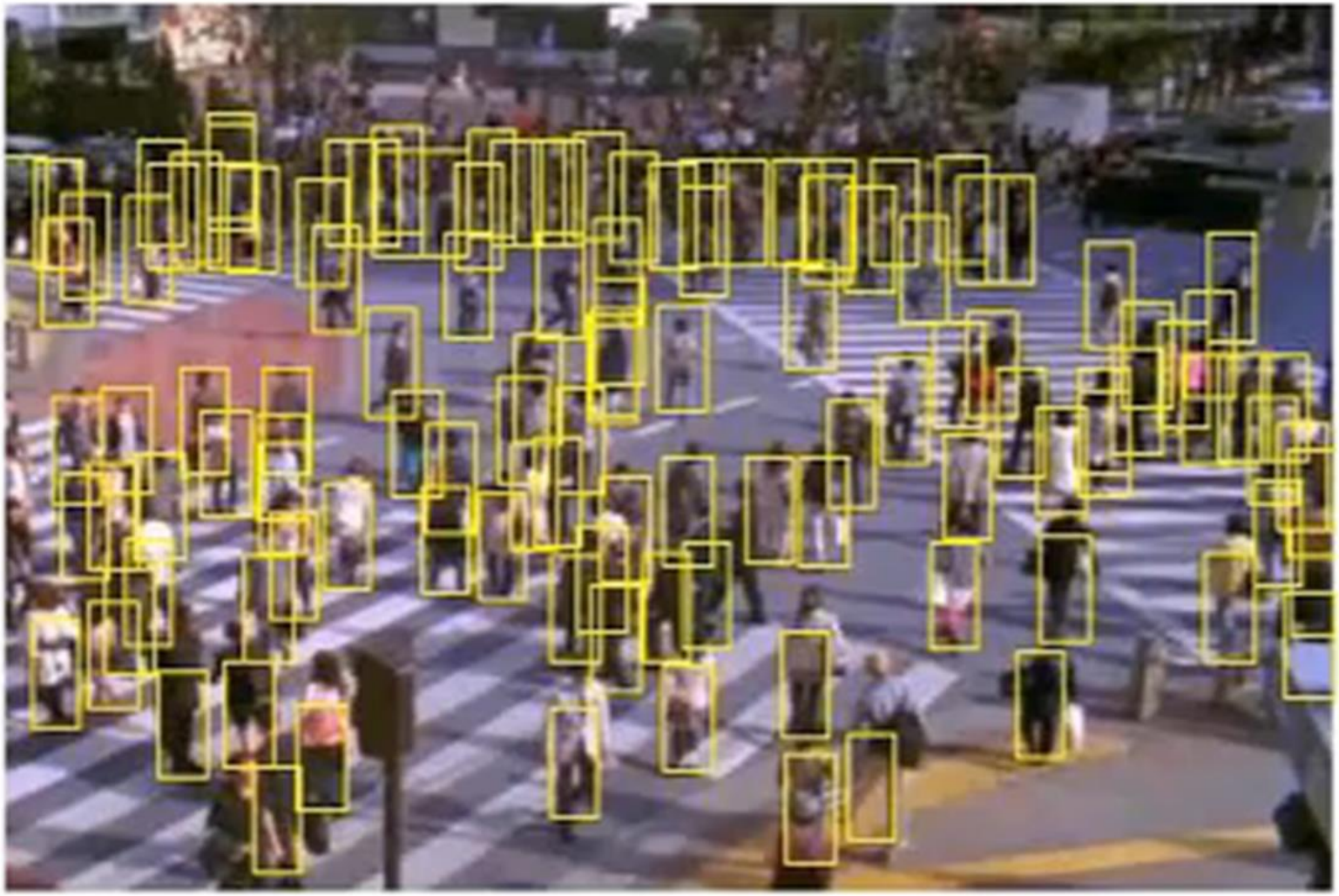


Fig. 2. Irregular partitioning of the image area through Voronoi diagrams: (a) Reports the first regular division of the image (50 × 50 = 2500 cells in this example); (b) shows the top view of the 2D histogram, while (c) shows a side view; and (d) shows the resulting Voronoi diagram with 50 cells.

S. Calderara, U. Heinemann, A. Prati, R. Cucchiara, N. Tishby,"Detecting Anomalies in People's Trajectories using Spectral Graph Analysis" Computer Vision and Image Understanding, vol. 115, n. 8, pp. 1099-1111, 2011

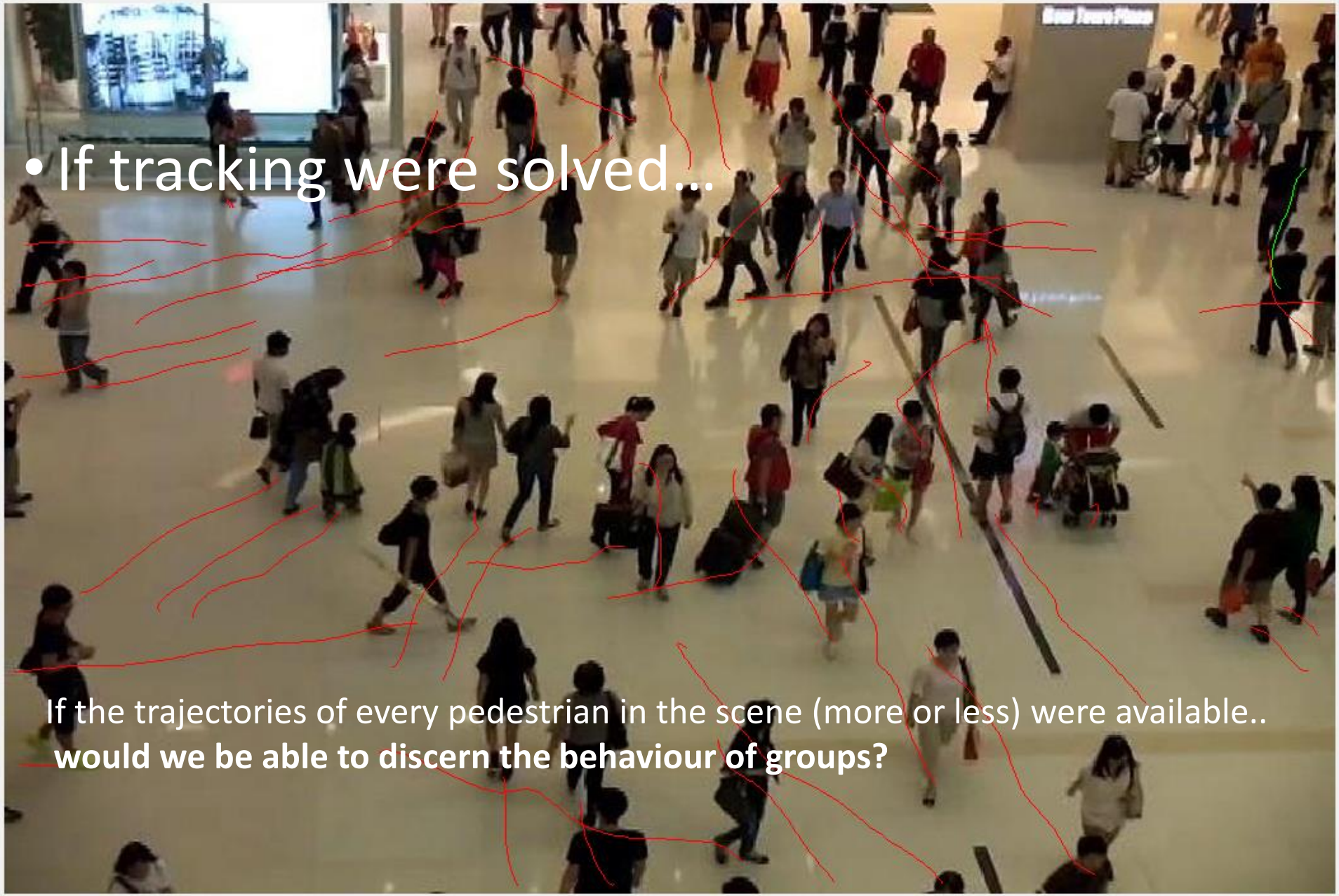
UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

HBU Around…



UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Groups of People



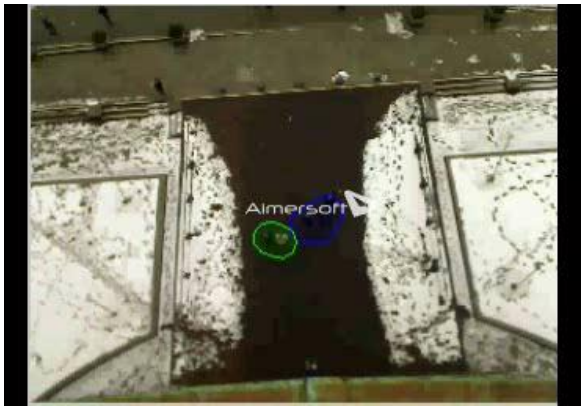
- If tracking were solved…

If the trajectories of every pedestrian in the scene (more or less) were available..
would we be able to discern the behaviour of groups?

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Some Results

Features: Proxemics and Granger causality
Structure function: pair-wise correlation clustering
Group detection: Structured SVM [groups]

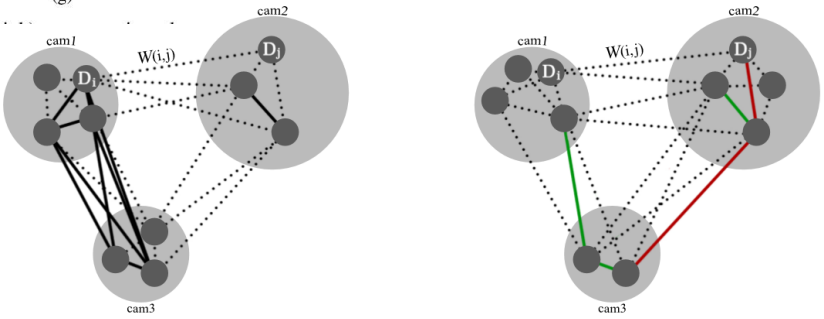


F.Solera, S.Calderara, R.Cucchiara"Socially Constrained Structural Learning for Groups Detection in Crowd *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, vol. 99, pp. 1 -14 , 2015

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Results in Multicamera MTT

- UNIMRE and Duke University
- Duke dataset



(a) Group clustering at T_k (b) Group clustering at T_{k+1}

Fig. 2: The problem of tracking groups is cast as correlation clustering. In (a) all the detected groups D_i observed in the time window T_k are taken into account, all the pairwise correlation $W(i, j)$ are computed (dashed lines) and a solution to tracking is found (solid lines). In (b), since time windows overlap in time, T_{k+1} will include group associations that were already solved in T_k. The new clustering is thus constrained by the previous solution forcing some observations to join (green lines) and others to remain separated (red lines), inducing consistent results across different time windows.

Tracking Social Groups Within and Across Cameras

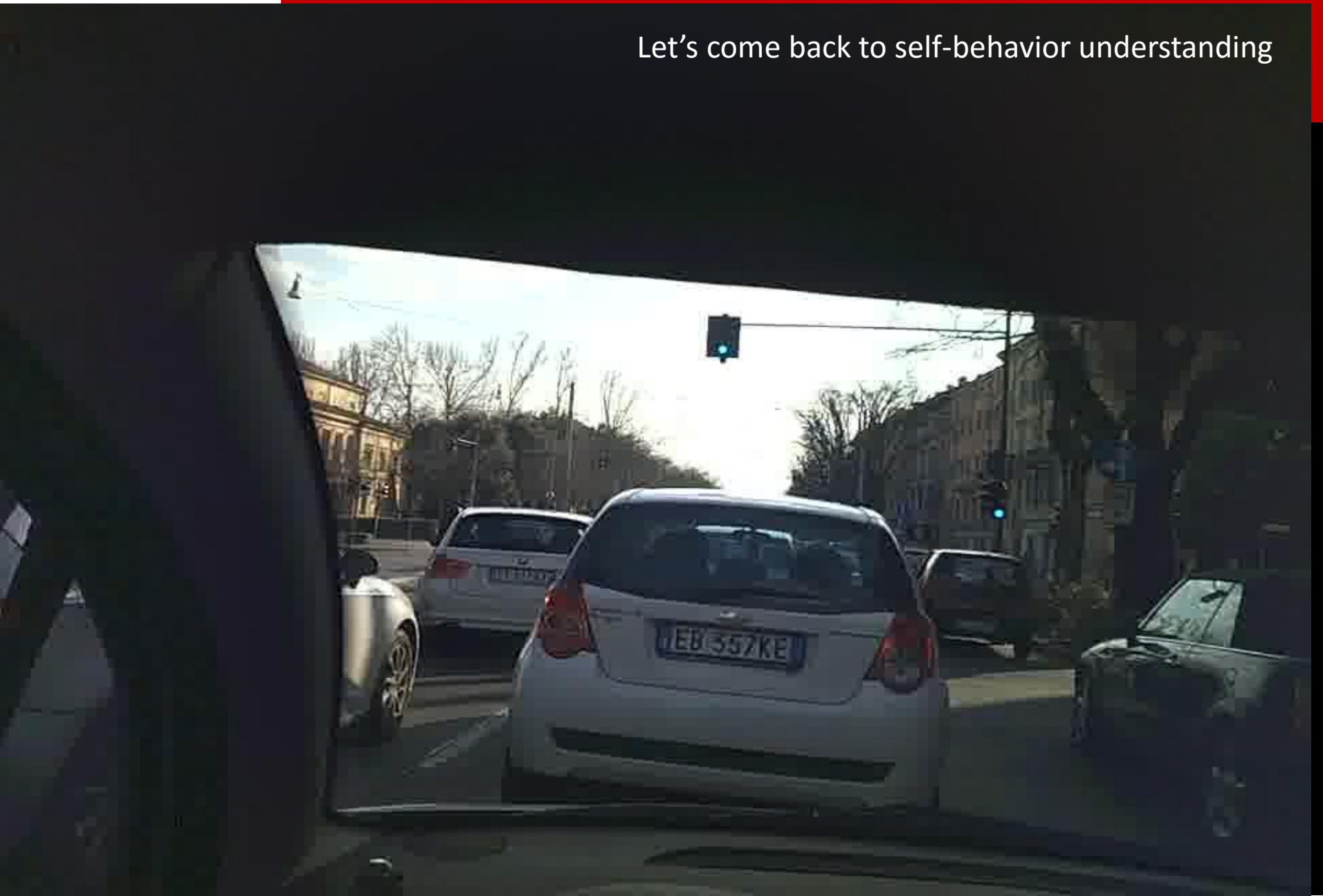
Francesco Solera, Simone Calderara, Ergys Ristani, Carlo Tomasi, Rita Cucchiara

IEEE Trans. On Circuit and Systems for Video technolgoy 2016

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

HBU in cars

Let’s come back to self-behavior understanding



UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

HBU

- Understanding the human pose/gaze
- Understanding the human attention/distraction
- Learn the human driving behavior

…use both car and human intelligences together.

Look outside

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

HBU

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

HBU –DR(eye)ve

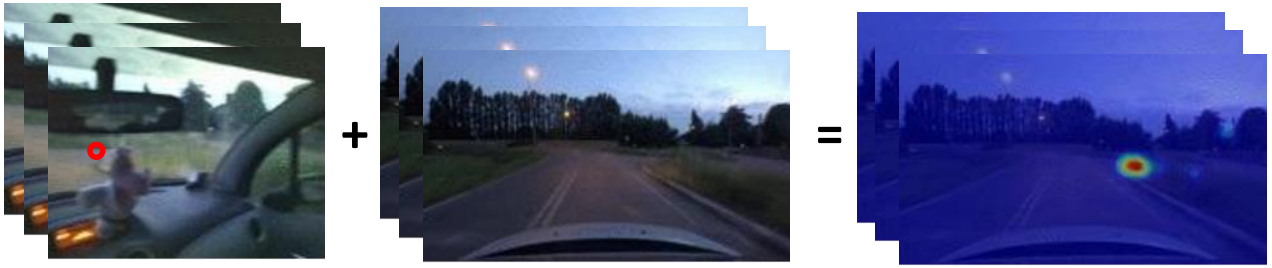
Stefano Alletto∗ , Andrea Palazzi∗ , Francesco Solera∗ , Simone Calderara and Rita Cucchiara **DR(eye)VE** a Dataset for Attention-Based Tasks with Applications to Autonomous and Assisted Driving CVPRW2016

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

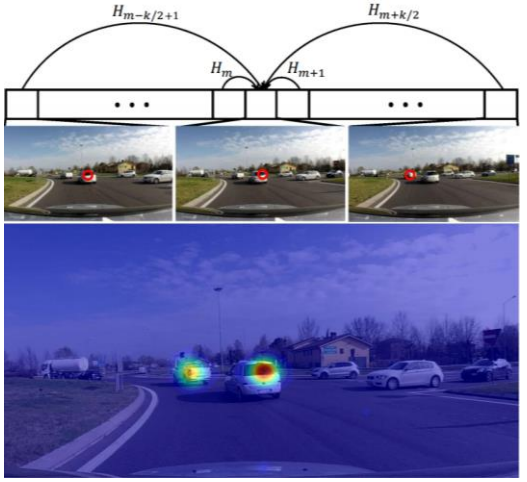
DR(eye)VE dataset: overview

Acquisition rig:

- Car mounted camera: Garmin VirbX 1080p/25fps, embedded GPS
- Eye tracker POV: SMI ETG HD camera 720p/30fps



Gaze position projected on the video of the roof-mounted camera

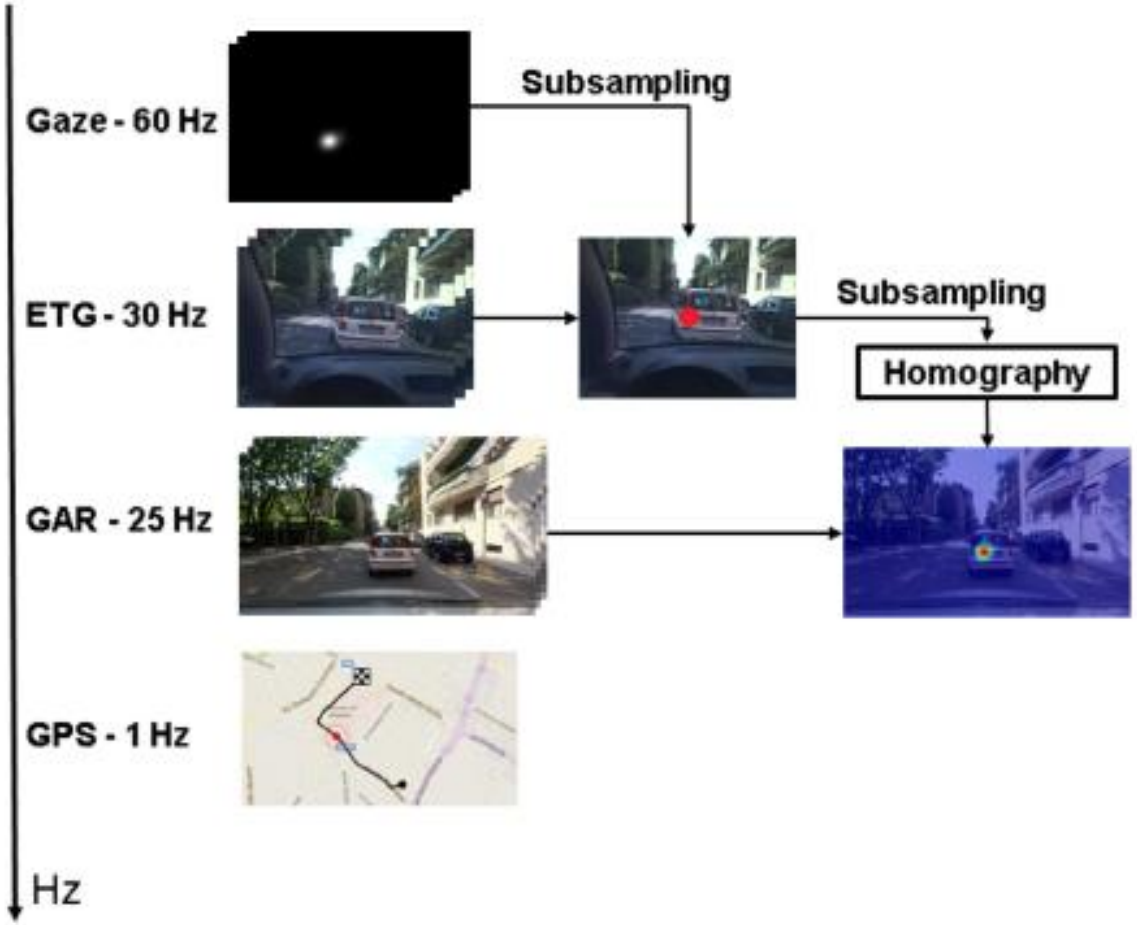


Ground Truth:

- Attentional map integrated over 25 frames (1 sec)
- Speed/GPS and driving course information

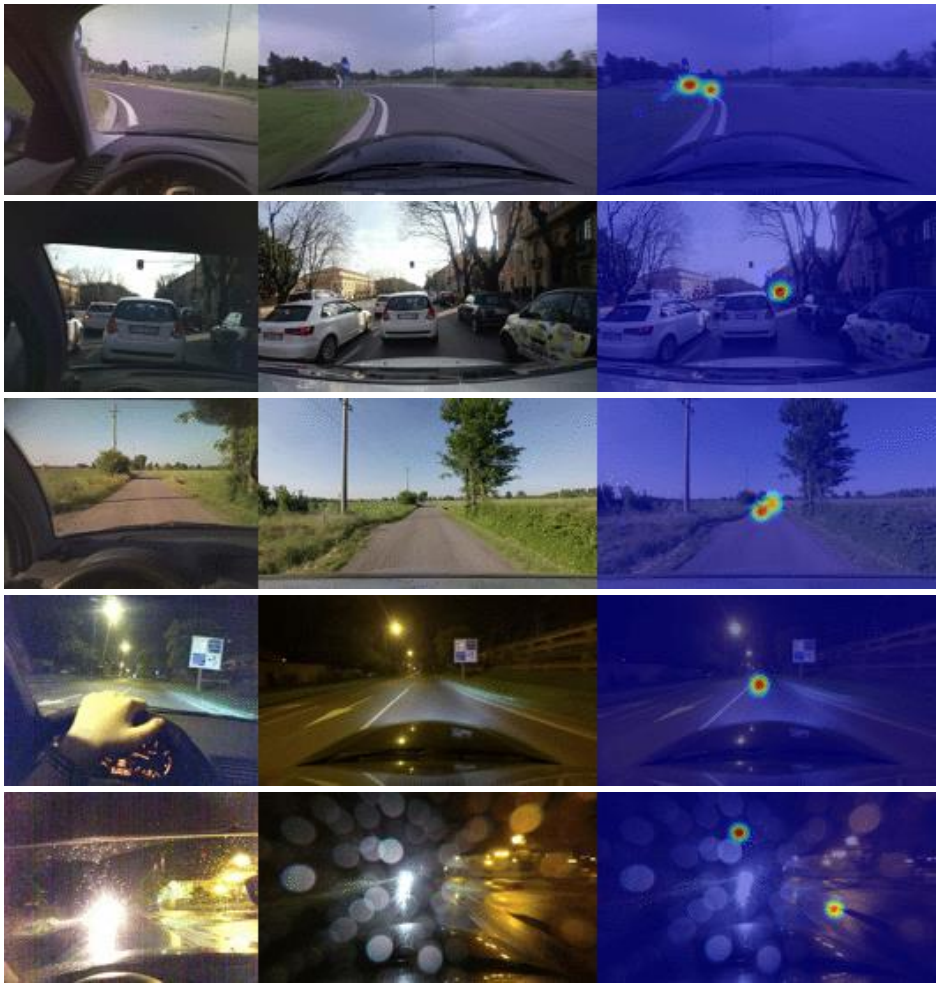
- Image registration and synchronization

- synchronization



UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

DR(EYE)ve dataset



- 8 different drivers
- 3 different landscapes
  {Highway, Countryside, Downtown}
- 3 different weather's conditions:
  {Sunny, Cloudy, Rainy}
- 3 different light's conditions:
  {Morning, Evening, Night}

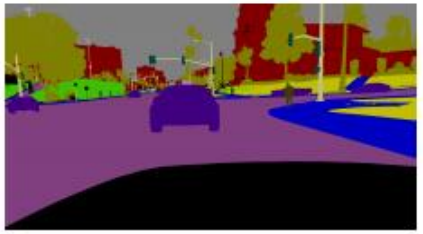
74 videos of 5 minutes each!

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Other datasets

Dataset	Frames	Drivers	Scenarios	Annotations	Real-world	Public
Pugeault *et al.* [17]	158,668	n.d.	Countryside, Highway Downtown	9 classes in Environment Road, Junction, Attributes	Yes	No
Simon *et al.* [19]	40	30	Downtown	Gaze Maps	No	No
Underwood *et al.* [23]	120	77	Urban Motorway	n.d.	No	No
Fridman *et al.* [6]	1,860,761	50	Highway	6 Gaze Location Classes	Yes	No
Dr(eye)ve	555,000	8	Countryside, Highway Downtown	Gaze Maps	Yes	Yes

Cityscapes [2]: 25000 frames depicting street scenes. Each frame is annotated for both pixel-level and instance-level segmentation (20000 coarse, 5000 fine).



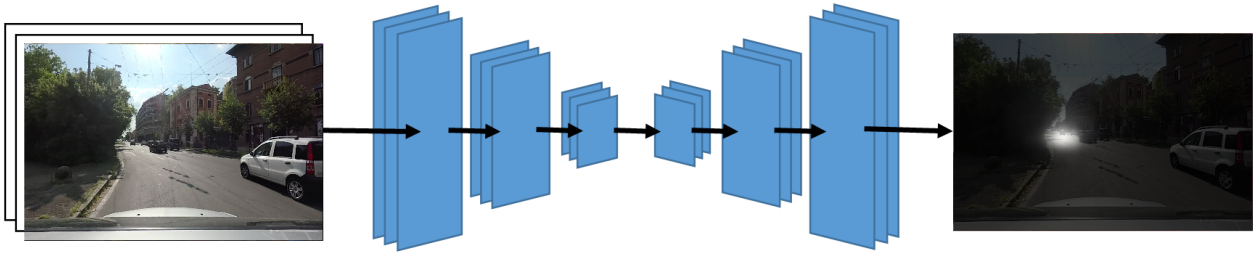
	frames	differentiation	real world	benchmark	addressed tasks
Dr(eye)ve	500000	landscapes, weather, daytime	yes	no	visual saliency
Cityscapes	25000	cities	yes	yes	semantic segmentation, instance level segmentation
Kitti	depends on the benchmark	landscapes	yes	yes	stereo, optical flow, visual odometry, object detection, tracking
PfD	25000	weather, daytime	no	yes	semantic segmentation

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

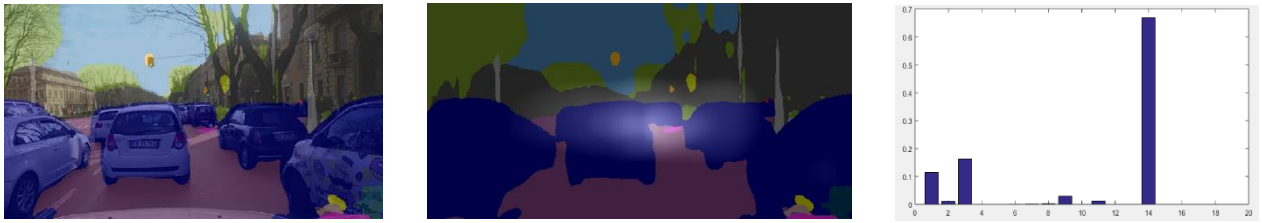
Dr(Eye)Ve project @Imagelab

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Dr(Eye)Ve project HBU



Good driving habits model: **where should we attend?**

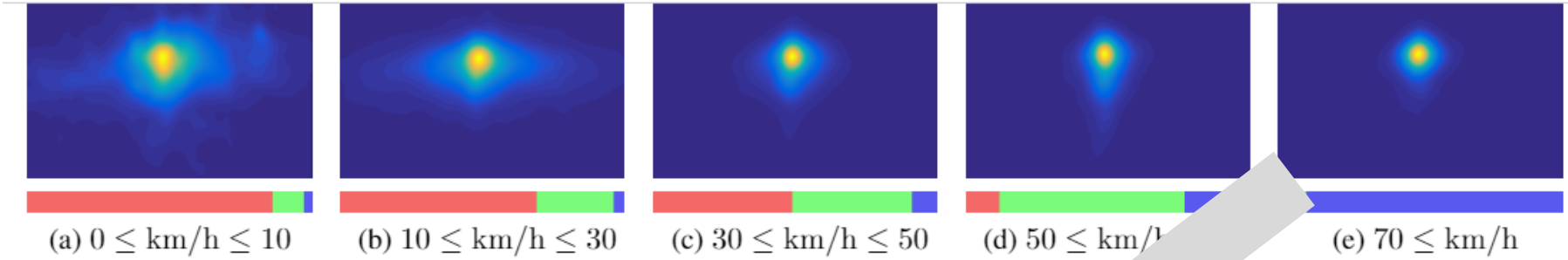


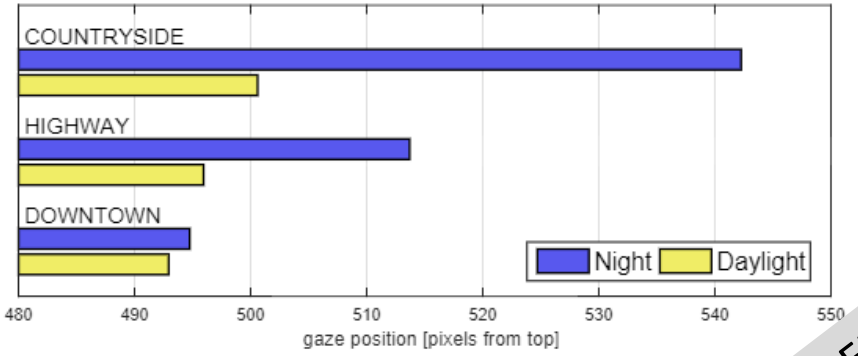
Semantic segmentation: **what are we actually looking at?**

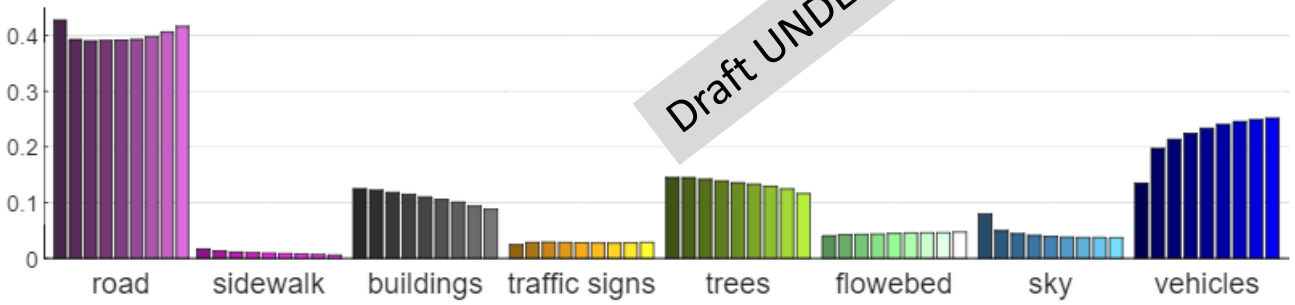
Look for us on http://imagelab.ing.unimore.it/dreyeve

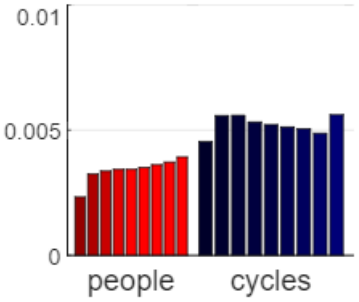
UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Some results: Attentive Behavior, measured





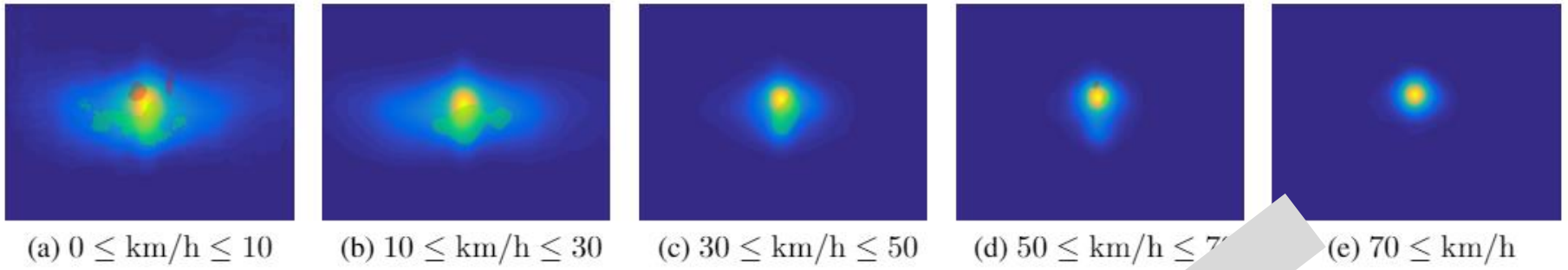




* Under submission

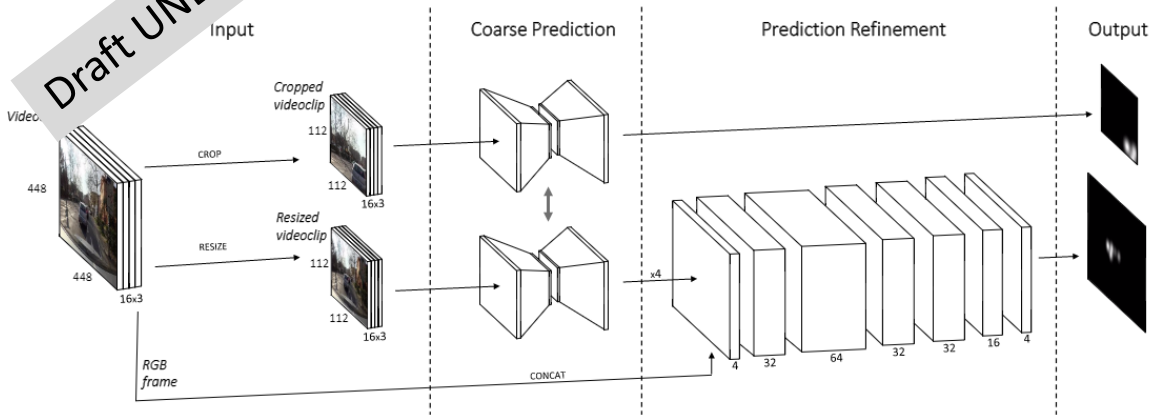
UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Some results: Attentive Behavior, predicted



(a) $0 \leq \text{km/h} \leq 10$ (b) $10 \leq \text{km/h} \leq 30$ (c) $30 \leq \text{km/h} \leq 50$ (d) $50 \leq \text{km/h} \leq 7$ (e) $70 \leq \text{km/h}$

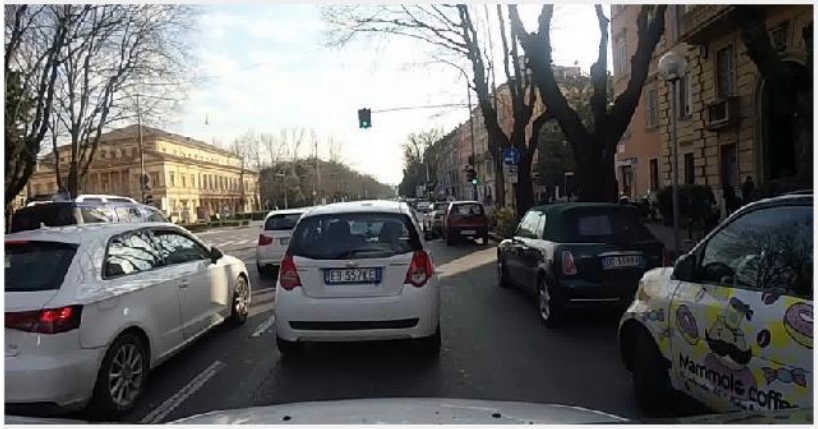
Fig. 9. Architecture of the coarse prediction module. The first part of the network perfo ature encoding of the input videoclip. The input videoclip is a tensor of size 3x16x112x112 that undergoes a sequence of conv3D ar yers that gradually squeeze it to size 512x7x7. All conv3D have kernel size (3,3,3) and ReLU activation units; all pool3D have pool s except the first one that has pool size (1,2,2). In order to obtain a saliency map with the same spatial size of the input frame, the feat entation is decoded through a series of intertwined layers of batch normalization, conv2D and x2 upsampling on the spatial dimensi v2D have kernel size (3,3) and are followed by leaky ReLU activations with $\alpha = .001$. As a result, the output of the network is a ten x112x112, *i.e.* the predicted grayscale saliency map.



Dr(eye)ve learned where the drivers see,
and what the drivers pay attention on…

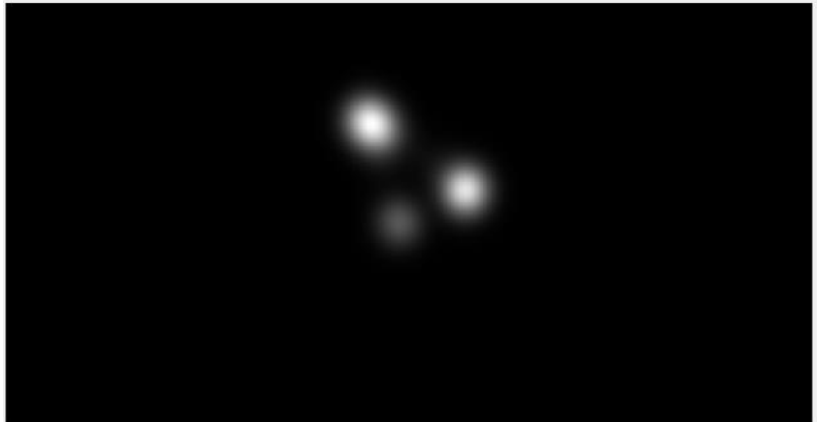
it is learning an intelligent visual behavior!

roof mounted camera



semantic segmentation

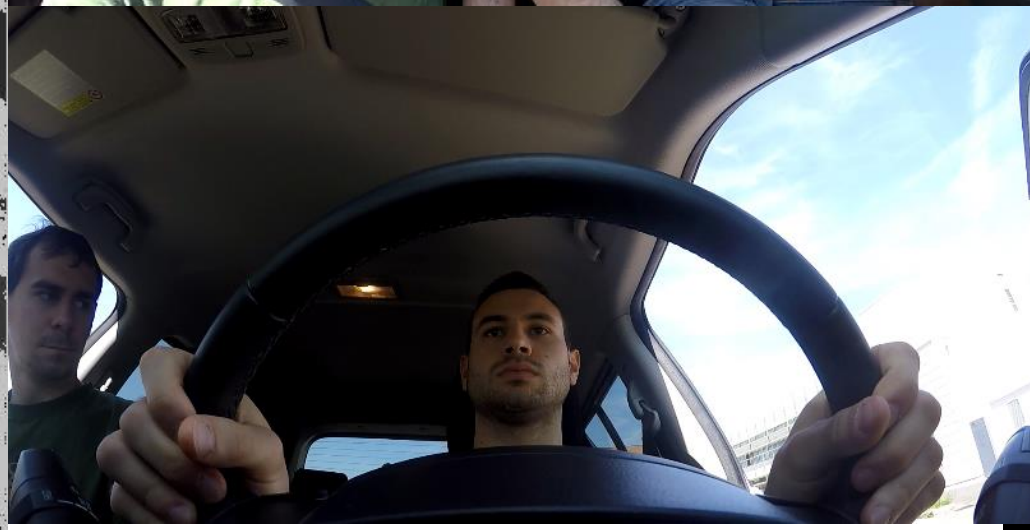
attention model prediction



overlay

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

HBU IN THE CAR by depth images



Look outside

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Human (un) attentive behavior→ Head Pose estimation

- Head Pose estimation
- Large literature on methods for Head Pose estimation**
- Approaches:
- Feature-based,
  - Nose, eyes
  - Landmark
- Appearance-based
  - Pixels classifiers
  - CNNS*
- 3D model registration ***
- Optimization based

**E. Murphy-Chutorian and M. M. Trivedi. Head pose estimation in computer vision: A survey. IEEE Trans. Pattern Anal.Mach. Intell., 31(4):607–626, Apr. 2009.

*S. S. Mukherjee and N. M. Robertson. Deep head pose:Gaze-direction estimation in multimodal video. IEEE Transactions on Multimedia, 17(11):2094–2107, 2015.

***C. Papazov, T. K. Marks, and M. Jones. Real-time 3d headpose and facial landmark estimation from depth images using triangular surface patch features. CVPR2015

POSEidon

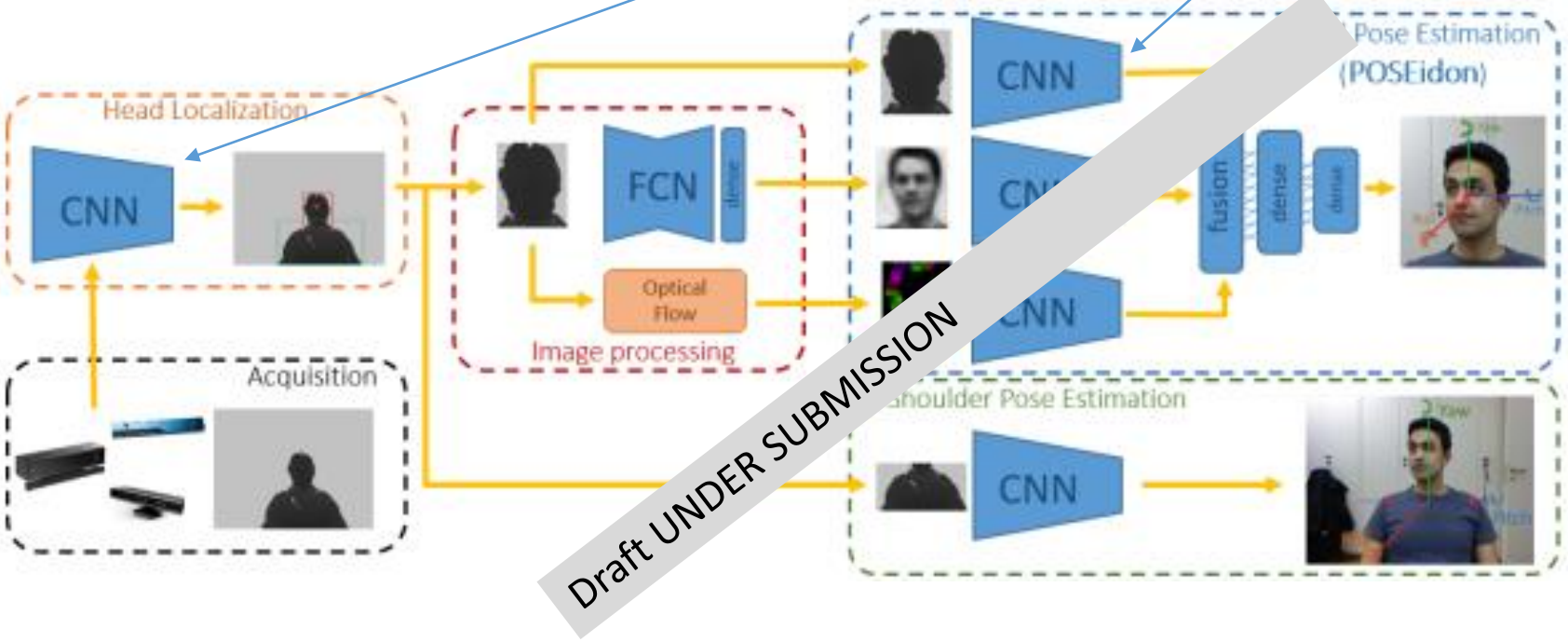
- A new approach:*

only row depth images

POSEidon a trident CNN

Order	Layer Type	# f/n	Kernel
1-2	2D Convolution (x2)	30	5x5
3	2D Convolution	30	4x4
4	2D Convolution	30	3x3
5	2D Convolution	120	3x3
6-7	2D Convolution (x2)	256	3x3
8-9	Dense (x2)	256	-
10	Dense	2	-

Order	Layer Type	# f/n	Kernel
1-2	2D Convolution (x2)	32	5x5
3	2D Convolution	32	4x4
4	2D Convolution	32	3x3
5	2D Convolution	128	3x3
6	Dense	128	-
7	Dense	84	-
8	Dense	3	-



Venturelli, Marco; Borghi, Guido; Vezzani, Roberto; Cucchiara, Rita "Deep Head Pose Estimation from Depth Data for In-car Automotive Applications" *Proceedings of the 2nd International Workshop on Understanding Human Activities through 3D Sensors*, Cancun (Mexico), Dec 4 , 2016, 2016

- next submission (Arxiv in 2 weeks)

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Preliminary results

- Preliminary comparative results

Method	Year	Data	Location	Pitch	Roll	Yaw	an	Acc.
Fanelli [16]*	2011	Depth	14.0	8.5 ± 9.9	7.9 ± 8.3	8.9 ± 13	± 10.4	0.790%
Yang [51]	2012	RGB + Depth	3.97 ± 2.18	9.1 ± 7.4	7.4 ± 4.9	8.9	8.5 ± 6.9	-
Padeleris [33]	2012	Depth	13.8	6.6	6.7		8.1	76.0
Rekik [38]	2013	RGB + Depth	5.1	4.3	5.2	.1	4.9	-
Baltrusaitis [2]	2012	RGB + Depth	7.6	5.1	11	6.3	7.6	-
Ahn [1]*	2014	RGB	-	3.4 ± 2.9	2	2.8 ± 2.4	2.9 ± 2.6	-
Martin [28]*	2014	Depth	5.8	2.5		3.6	2.9	-
Saeed [39]	2015	RGB + Depth	-	5.0 ±	± 4.6	3.9 ± 4.2	4.4 ± 4.9	-
Papazov [35]	2015	Depth	8.4	2.5	3.8 ± 16.0	3.0 ± 9.6	4.0 ± 11.0	-
Drouard [13]	2015	RGB	-	.8	4.7 ± 4.6	4.9 ± 4.1	5.2 ± 4.5	-
Meyer [30]	2015	RGB	-	2.4	2.1	2.1	2.2	0.946
Liu [25]	2016	RGB		6.0 ± 5.8	5.7 ± 7.3	6.1 ± 5.2	5.9 ± 6.1	-
POSEidon	2016	Depth	5	**1.6 ± 1.7**	**1.8 ± 1.8**	**1.7 ± 1.5**	**1.7 ± 1.7**	**0.950**

Draft UNDER SUBMISSION

Table 4. Results on *Biwi Dataset*. In this case, no h localization is performed for head pose estimation task. Location is expressed in millimeters. The last column report the accuracy, established as the number of angle prediction below a certain threshold (10°).

Understanding the human pose by depth only (with DL)

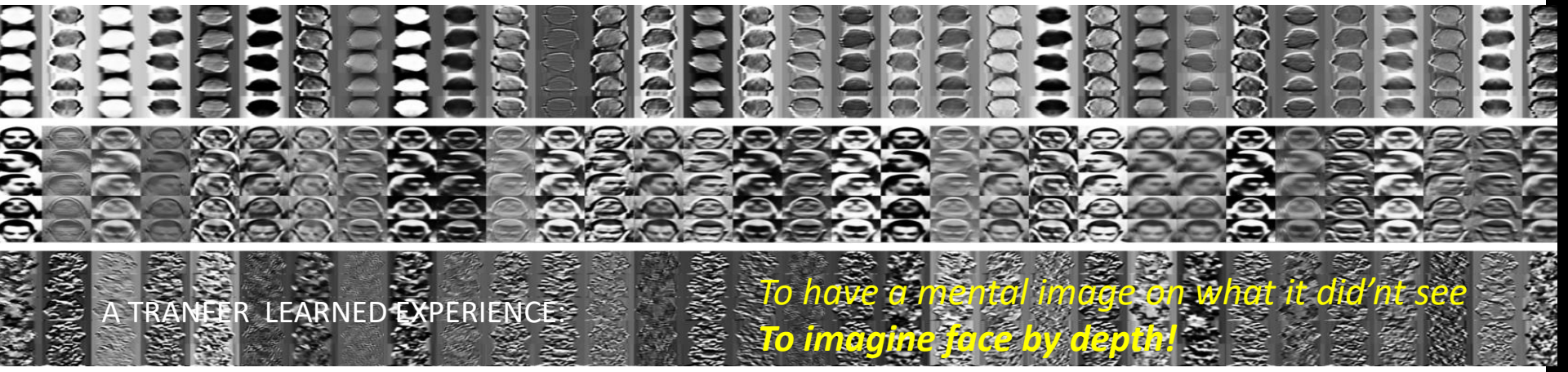
UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Depth-to-face an imrpessive side effect

DEPTH | RGB from DEPTH | Original unseenRGB

Learned by the POSEidon Net @Imagelab

POSEidon learned something more…



A TRANEER LEARNED EXPERIENCE

To have a mental image on what it did'nt see

To imagine face by depth!

Depth to face

- A new CNN architecture (thanks to Guido Borghi and Marco Venturelli, Rob Vezzani)
- fuses the key aspects of autoencoder and fully connected deep networks
- The loss function works on centred images with a multivariate Gaussian on prior mask (parameters 3.5, 2,5)
- And Adadelta optimizer

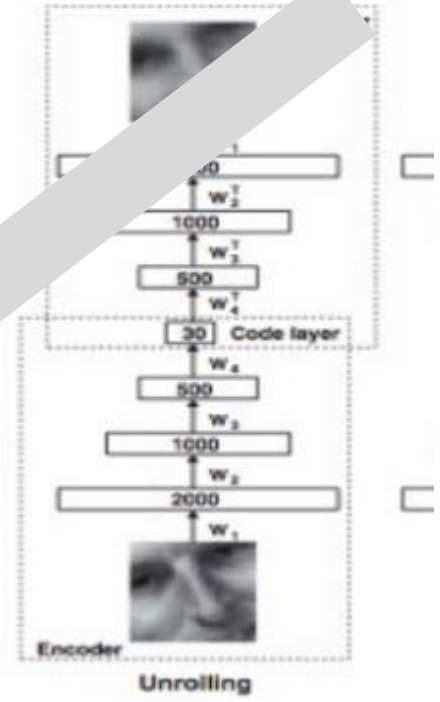
$$L = \frac{1}{N}\sum^{R}\sum^{C}\left[\frac{1}{ch}\sum^{ch}(y_{ik} - \bar{y}_{ik})\right]w_{ij}$$

$$w = (2\pi)^{-\frac{k}{2}}|\Sigma|^{-\frac{1}{2}}e^{(x-\mu)^T\Sigma^{-1}(x-\mu)}$$

$$\mu, \Sigma = \left[\begin{pmatrix}\frac{R}{2}\\ \frac{C}{2}\end{pmatrix}, \begin{pmatrix}\left(\frac{R}{\alpha}\right)^2 & \frac{R \cdot C}{\alpha\beta}\\ \frac{R \cdot C}{\alpha\beta} & \left(\frac{C}{\beta}\right)^2\end{pmatrix}\right]$$

Order	Layer Type	#f/n	Kernel
1-2	2D Convolution (x2)	30	5x5
3	2D Convolution	60	4x4
4	2D Convolution	60	3x3
5-6	2D Convolution (x2)	120	3x3
7	2D Convolution	256	1x1
8	2D Convolution	256	1x1
9-10	2D Convolution (x2)	120	3x3
11	2D Convolution	60	3x3
12	2D Convolution	60	4x4
13-14	2D Convolution (x2)	30	5x5
15	Dense	$r \times c \times ch$	-

Table 2. Architecture for depth-to-gray reconstruction n



Draft UNDER SUBMISSION

M. D. Zeiler. Adadelta: an adaptive learning rate method. arXiv preprint ar :1212.5701, 2012.
Pierre Baldi Autoencoders, Unsupervised Learning, and Deep Architectures JMLR W 2012

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

3D Pandora dataset @Imagelab

2D Pandora dataset @Imagelab

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Learned by the Poseydon Net @Imagelab

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Very very preliminary results

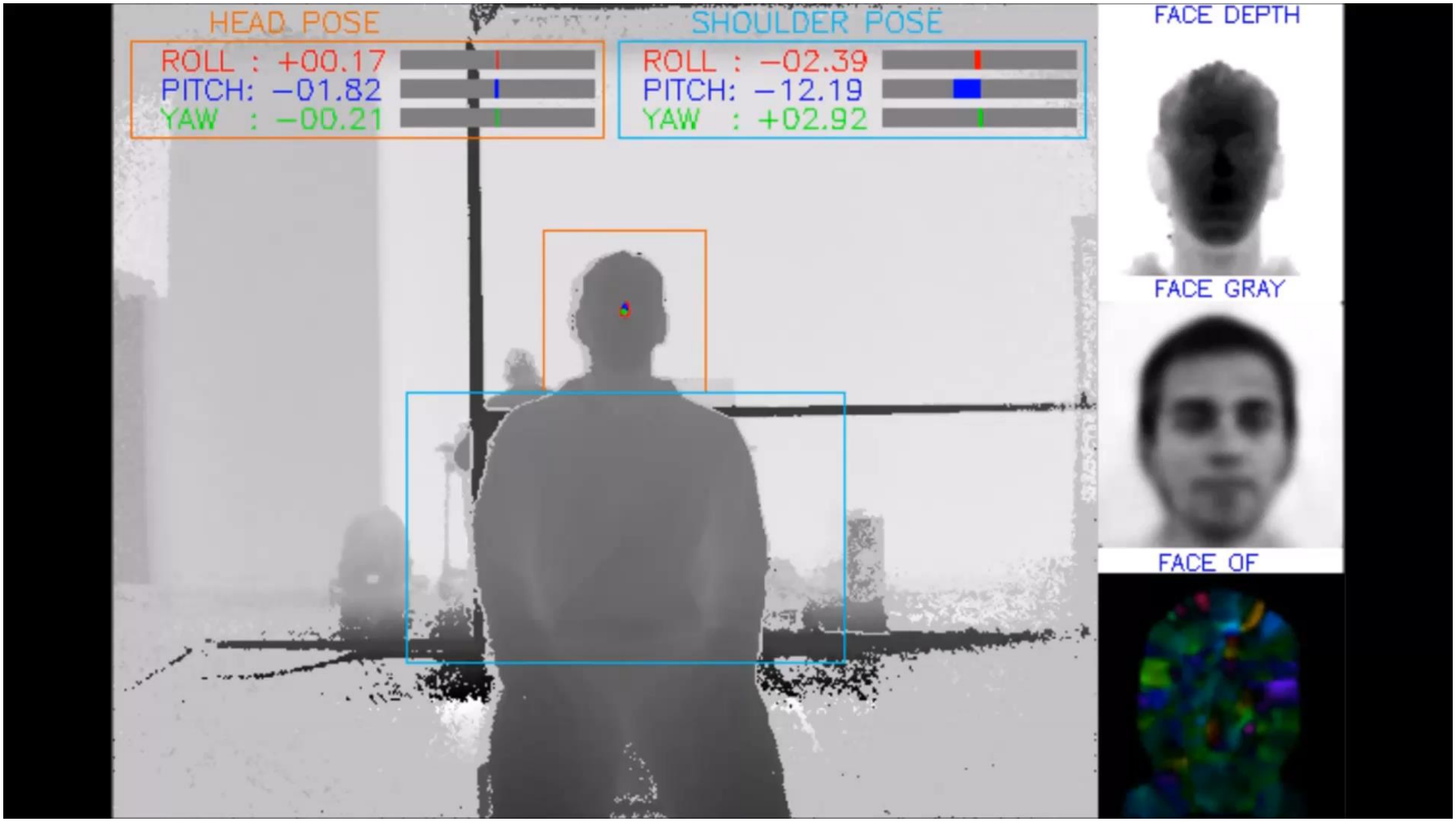
- Results on PANDORA Dataset

Method	Fusion	Head			Accuracy
		Pitch	Roll	Yaw	
depth (no crop)	-	8.0 ± 7.0	6.2 ± 5.3	11.7 [illegible]	0.553
depth	-	6.5 ± 6.6	5.4 ± 5.1	[illegible] 11.8	0.646
gray	-	6.8 ± 7.0	5.7 ± 5.7	[illegible] ± 14.6	0.647
rgb	-	7.1 ± 6.6	5.6 ± [illegible]	9.0 ± 10.9	0.639
optical flow	-	7.7 ± 7.5	5 [illegible]	10.0 ± 12.5	0.609
depth + gray	concat	5.6 ± 5.0	4 [illegible] ± 5.0	9.8 ± 13.4	0.698
depth + OF	concat	6.0 ± 6.1	4.5 ± 4.8	9.2 ±11.5	0.690
POSEidon	conv	5.7 ± 5.6	4.9 ± 5.1	9.0 ±11.9	0.715
POSEidon	concat	6.3 ± 6.1	5.0 ± 5.0	10.6 ±14.2	0.657
POSEidon	mul+concat	**5.6 ± 5.6**	**4.9 ± 5.2**	**9.1± 11.9**	0.712

Draft UNDER SUBMISSION

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

A demo



UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Concluding. Where we are now..

- Human behavior understanding (by vision)
  - A lot of stuff in computer vision and pattern recognition
  - Geometry, graphs and statistical data analysis is unavoidable
  - Features are mostly CNN-based
  - Detection is not enough
  - Spatio temporal coherence is needed (often tracking)
  - New forms of input data are useful (sensors 3D…)
  - A lot of learning … the importance of datasets
  - A strong knowledge of the context with experts (drivers, automotive industries, security persons, psychologists)
  - **We are becoming truly multidisciplinar and our systems truly intelligent.**

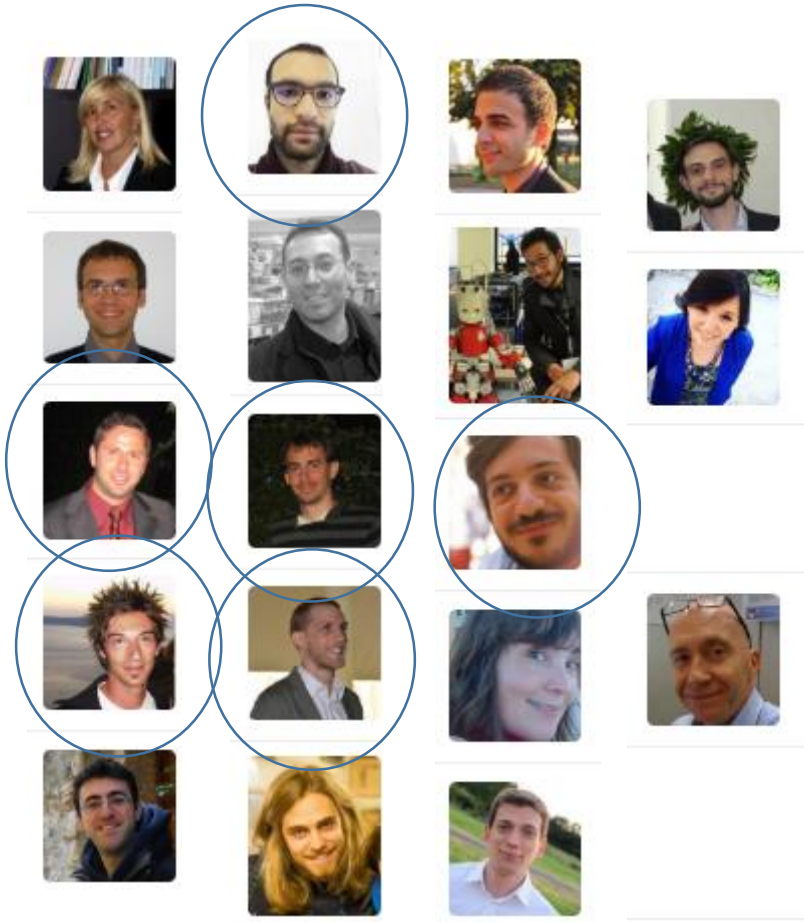
Thanks to ImageLab

http://imagelab.ing.unimo.it

RETE ALTA TECNOLOGIA
EMILIA-ROMAGNA
HIGH TECHNOLOGY NETWORK

softech-ict

Inter-dipartimental Research Center in ICT
Tecnopolo di Modena
Emila Romagna High Technology Network



Thanks to:
Rita Cucchiara, Costantino Grana, **Roberto Vezzani**, **Simone Calderara**,[Giuseppe Serra],
Stefano Aletto, Fabrizio Balducci, **Guido Borghi**, **Andrea Palazzi**, Federico Bolelli,
[Marco Manfredi], Francesco Paci, **Francesco Solera**, Patrizia Varini, Lorenzo Baraldi,
Andrea Corbelli, Marcella Cornia, Augusto Pieracci, Paolo Santinelli, Silvia Calio and **Marco Venturelli**