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 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
Imagelab : research since 1998

- 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

www.imagelab.unimore.it
Computer Vision and Human Behavior Understanding in cars and around
• **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.
But... Can computer vision provide effectively Human behavior Understanding?
Human behaviors

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

- Single and collective behavior
- Collaborative or not collaborative behaviors

MIUR-FAR project
DriveAttention
2016-2018

PRIN project
“2106-2018

project “VAEX 2015-2018

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

• One for all: * R. Poppe “A survey on vision based action recognition” Image and vision computing 2010
Research in HBU

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

1000 scientific papers from google scholar 2015

- HBU & surveillance: 25%
- HBU & multimedia: 21%
- HBU & interaction: 36%
- HBU & health care: 18%
Movements
- Body movements
- Gestures
- Poses
- Gaze and Expression

Actions

Activities

Behavior?

HBU, what knowledge can be extracted by video?
• Behavior

• \{\textit{movements, actions, activities}\} + \{\textit{environment, objects, people}\} + \{\textit{purposes, beliefs, habits}\}

• Self-behavior, Personal behavior, Social behavior, ...
HBU by Vision, a multidimensional space

Data source type
- Wearable or egocentric cameras
- Mobile cameras
- Surveillance static cameras

Aspects of diversity
- Gender
- Age
- Race
- Geographical origin
- Religion

Person cardinality
- Single Persons
- Groups of Persons
- Crowd

Social constraints

Purpose

Target Collaboration

Observation target
- Body
- Upper-body
- Hands
- Face

Other sensors

Body-Env

Collaboration

Target

UNIMORE
the levels of details vs the level of sociality

- **Levels of details**

  - Lev1 self: Behavior, expression, gesture, pose
  - Lev2 whole: Person, action, interaction
  - Lev3 social: Behavior, in the environment, in crowd

- Human parts
- Human full body
- Human(s) in the environment
- Humans in crowd

---

**Human**

**Levels of details vs the level of sociality**

**Human parts**

**Levels of details**

- Lev1 self: Behavior, expression, gesture, pose
- Lev2 whole: Person, action, interaction
- Lev3 social: Behavior, in the environment, in crowd

---

**Human full body**

**Levels of details**

- Lev1 self: Behavior, expression, gesture, pose
- Lev2 whole: Person, action, interaction
- Lev3 social: Behavior, in the environment, in crowd

---

**Human(s) in the environment**

**Levels of details**

- Lev1 self: Behavior, expression, gesture, pose
- Lev2 whole: Person, action, interaction
- Lev3 social: Behavior, in the environment, in crowd

---

**Humans in crowd**

**Levels of details**

- Lev1 self: Behavior, expression, gesture, pose
- Lev2 whole: Person, action, interaction
- Lev3 social: Behavior, in the environment, in crowd
Understanding humans for New natural Human Computer interaction systems *

What are they doing?

Deaf Sign language

Maja Pantic, Alex Pentland, Anton Nijholt and Thomas Huang


S. Rautaray, A Agrawal Vision based hand gesture recognition for human computer interaction: a survey Artificial Intelligence Review January 2015,
Lev 1: Expressions, poses and gestures

• Human attention in the car
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 Stream Discrete HMMs on 3D Skeletons" Proceedings of the 23rd International Conference on Pattern Recognition, Cancun, Dec 4-8, 2016, 2016
CVPR2016

Lev 2 Action again, new trends

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
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.
• Simple but effective CV features

**Volume distance**

\[ F_{vol}(f) = Vol(DT(P(f))) \]

**Delaunay Triangulation**

\[ DT(P(f)) = \text{Delaunay Triangulation} \]

**Mutual distance**

\[ D_m(f) = dist(C^p, C^c) \]

\[ C^p(f) = \frac{1}{m} \sum_{i=1}^{m} \text{joint}_i^p(f) \]

\[ C^c(f) = \frac{1}{m} \sum_{i=1}^{m} \text{joint}_i^c(f) \]
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
Segmentation/detection in surveillance

Segmentation by motion

Detection with learning and classifiers

Differential Motion

HOG, part based, CNNs People detectors

Background suppression


End-To-End People Detection in Crowded Scenes
Russell Stewart, Mykhaylo Andriluka, Andrew Y. Ng CVPR2016
- Tracking few people in a constrained environment: «solved problem» 😊

Detection and tracking in commercial systems

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

In semi-constrained world Tracking is possible
..and tracking single (people) target

- Is tracking a solved problem? 😅 😅 😅
- We tried to answer this question in an “experimental evaluation”
- Even in case of single target tracking*
- - a very large dataset
- - of 14 categories of challenges
- - a large set of performance measures
- - a large experimentation
- (with code available over 3 clusters in 3 labs)

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

http://www.alov300.org
http://imagelab.ing.unimo.it/dsm

MOTA; OTA; Deviation....
F-Measure
SURVIVAL CURVES..

19 trackers
BASELINES
STATE OF THE ART
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-OCLUSION
12-MOVINGCAMERA
13-ZOOMINGCAMERA
14-LONGDURATION
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%
Tracking by detection with transductive learning *

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

Tracking by detection with structural SVM **

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


Fig. 2. Irregular partitioning of the image area through Voronoi diagrams: (a) Reports the first regular division of the image (50 x 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.
HBU Around...
If the trajectories of every pedestrian in the scene (more or less) were available, would we be able to discern the behaviour of groups?
Some Results

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

Results in Multicamera MTT

• UNIMRE and Duke University
• Duke dataset

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 Tornasi, Rita Cucchiara
HBU in cars

Let’s come back to self-behavior understanding
• Understanding the human pose/gaze
• Understanding the human attention/distraction
• Learn the human driving behavior

...use both car and human intelligences together.

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

**Ground Truth:**
- Attentional map integrated over 25 frames (1 sec)
- Speed/GPS and driving course information
• Image registration and synchronization
• synchronization
• 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!
### Other datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Frames</th>
<th>Drivers</th>
<th>Scenarios</th>
<th>Annotations</th>
<th>Real-world</th>
<th>Public</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pugeault et al.</td>
<td>158,668</td>
<td>n.d.</td>
<td>Countryside, Highway, Downtown</td>
<td>9 classes in Environment, Road, Junction, Attributes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Simon et al.</td>
<td>40</td>
<td>30</td>
<td>Downtown</td>
<td>Gaze Maps</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Underwood et al.</td>
<td>120</td>
<td>77</td>
<td>Urban Motorway</td>
<td>n.d.</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Fridman et al.</td>
<td>1,860,761</td>
<td>50</td>
<td>Highway</td>
<td>6 Gaze Location Classes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Dr(eye)ve</td>
<td>555,000</td>
<td>8</td>
<td>Countryside, Highway, Downtown</td>
<td>Gaze Maps</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Frames</th>
<th>Differentiation</th>
<th>Real world</th>
<th>Benchmark</th>
<th>Addressed tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dr(eye)ve</td>
<td>500000</td>
<td>landscapes, weather, daytime</td>
<td>yes</td>
<td>no</td>
<td>visual saliency</td>
</tr>
<tr>
<td>Cityscapes</td>
<td>25000</td>
<td>cities</td>
<td>yes</td>
<td>yes</td>
<td>semantic segmentation, instance level segmentation</td>
</tr>
<tr>
<td>Kitti</td>
<td>depends on the benchmark</td>
<td>landscapes</td>
<td>yes</td>
<td>yes</td>
<td>stereo, optical flow, visual odometry, object detection, tracking</td>
</tr>
<tr>
<td>PfD</td>
<td>25000</td>
<td>weather, daytime</td>
<td>no</td>
<td>yes</td>
<td>semantic segmentation</td>
</tr>
</tbody>
</table>
Dr(Eye)Ve project @Imagelab
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](http://imagelab.ing.unimore.it/dreyeve)
Some results: Attentive Behavior, measured

* Under submission
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) $50 \leq \text{km/h} \leq 70$
(e) $70 \leq \text{km/h}$

Fig. 9. Architecture of the coarse prediction module. The first part of the network performs a feature encoding of the input videoclip. The input videoclip is a tensor of size $3 \times 16 \times 112 \times 112$ that undergoes a sequence of conv3D and maxpool layers that gradually squeeze it to size $512 \times 7 \times 7$. All conv3D have kernel size $(3,3,3)$ and ReLU activation units; all pool3D have pool size $(2,2,2)$ 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 feature representation is decoded through a series of intertwined layers of batch normalization, conv2D and $x2$ upsampling on the spatial dimension. conv2D 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 tensor of size $1 \times 112 \times 112$, 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!
• 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

• A new approach:*
  only row depth images
POSEidon a trident CNN

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)
### Preliminary results

- Preliminary comparative results

<table>
<thead>
<tr>
<th>Method</th>
<th>Year</th>
<th>Data</th>
<th>Location</th>
<th>Pitch</th>
<th>Roll</th>
<th>Yaw</th>
<th>Mean</th>
<th>Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fanelli [16]*</td>
<td>2011</td>
<td>Depth</td>
<td>14.0</td>
<td>8.5 ± 9.9</td>
<td>7.9 ± 8.3</td>
<td>8.9 ± 13.6</td>
<td>16 ± 10.4</td>
<td>0.790%</td>
</tr>
<tr>
<td>Yang [51]</td>
<td>2012</td>
<td>RGB + Depth</td>
<td>3.97 ± 2.18</td>
<td>9.1 ± 7.4</td>
<td>7.4 ± 4.9</td>
<td>8.9 ± 11.6</td>
<td>8.5 ± 6.9</td>
<td>-</td>
</tr>
<tr>
<td>Padeleris [33]</td>
<td>2012</td>
<td>Depth</td>
<td>13.8</td>
<td>6.6</td>
<td>6.7</td>
<td>8.1</td>
<td>-</td>
<td>76.0</td>
</tr>
<tr>
<td>Rekik [38]</td>
<td>2013</td>
<td>RGB + Depth</td>
<td>5.1</td>
<td>4.3</td>
<td>5.2</td>
<td>4.9</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Baltrusaitis [2]</td>
<td>2012</td>
<td>RGB + Depth</td>
<td>7.6</td>
<td>5.1</td>
<td>11.2</td>
<td>6.3</td>
<td>7.6</td>
<td>-</td>
</tr>
<tr>
<td>Ahn [1]*</td>
<td>2014</td>
<td>RGB</td>
<td>-</td>
<td>3.4 ± 2.9</td>
<td>2.9</td>
<td>2.8 ± 2.4</td>
<td>2.9 ± 2.6</td>
<td>-</td>
</tr>
<tr>
<td>Martin [28]*</td>
<td>2014</td>
<td>Depth</td>
<td>5.8</td>
<td>2.5</td>
<td>3.6</td>
<td>2.9</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Saeed [39]</td>
<td>2015</td>
<td>RGB + Depth</td>
<td>-</td>
<td>5.0 ± 5.6</td>
<td>5.5 ± 4.6</td>
<td>3.9 ± 4.2</td>
<td>4.4 ± 4.9</td>
<td>-</td>
</tr>
<tr>
<td>Papazov [35]</td>
<td>2015</td>
<td>Depth</td>
<td>8.4</td>
<td>2.5 ± 4.8</td>
<td>3.8 ± 16.0</td>
<td>3.0 ± 9.6</td>
<td>4.0 ± 11.0</td>
<td>-</td>
</tr>
<tr>
<td>Drouard [13]</td>
<td>2015</td>
<td>RGB</td>
<td>-</td>
<td>5.9 ± 2.8</td>
<td>4.7 ± 4.6</td>
<td>4.9 ± 4.1</td>
<td>5.2 ± 4.5</td>
<td>-</td>
</tr>
<tr>
<td>Meyer [30]</td>
<td>2015</td>
<td>RGB</td>
<td>-</td>
<td>2.4</td>
<td>2.1</td>
<td>2.1</td>
<td>2.2</td>
<td>0.946</td>
</tr>
<tr>
<td>Liu [25]</td>
<td>2016</td>
<td>RGB</td>
<td>6.0 ± 5.8</td>
<td>5.7 ± 7.3</td>
<td>6.1 ± 5.2</td>
<td>5.9 ± 6.1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>POSEidon</td>
<td>2016</td>
<td>Depth</td>
<td>5.3 ± 9</td>
<td>1.6 ± 1.7</td>
<td>1.8 ± 1.8</td>
<td>1.7 ± 1.5</td>
<td>1.7 ± 1.7</td>
<td>0.950</td>
</tr>
</tbody>
</table>

Table 4. Results on Biwi Dataset. In this case, no head 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)
Depth-to-face an impressive side effect

Learned by the POSEidon Net @Imagelab

POSEidon learned something more...

To have a mental image on what it didn’t see
To imagine face by depth!
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

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

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

\[
\mu, \Sigma = \left[ \left( \frac{R}{C} \right)^2, \left( \frac{R}{\alpha} \frac{C}{\beta} \right) \left( \frac{R}{\alpha} \frac{C}{\beta} \right)^2 \right]
\]

---

Pierre Baldi Autoencoders, Unsupervised Learning, and Deep Architectures JMLR W 2012
3D Pandora dataset @Imagelab
2D Pandora dataset @Imagelab
Learned by the Poseydon Net @Imagelab
Very very preliminary results

- Results on PANDORA Dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>Fusion</th>
<th>Pitch</th>
<th>Head Roll</th>
<th>Yaw</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>depth (no crop)</td>
<td>-</td>
<td>8.0 ± 7.0</td>
<td>6.2 ± 5.3</td>
<td>11.7 ± 8.2</td>
<td>0.553</td>
</tr>
<tr>
<td>depth</td>
<td>-</td>
<td>6.5 ± 6.6</td>
<td>5.4 ± 5.1</td>
<td>10.3 ± 11.8</td>
<td>0.646</td>
</tr>
<tr>
<td>gray</td>
<td>-</td>
<td>6.8 ± 7.0</td>
<td>5.7 ± 5.7</td>
<td>10.3 ± 14.6</td>
<td>0.647</td>
</tr>
<tr>
<td>rgb</td>
<td>-</td>
<td>7.1 ± 6.6</td>
<td>5.6 ± 5.3</td>
<td>9.0 ± 10.9</td>
<td>0.639</td>
</tr>
<tr>
<td>optical flow</td>
<td>-</td>
<td>7.7 ± 7.5</td>
<td>5.3 ± 5.7</td>
<td>10.0 ± 12.5</td>
<td>0.609</td>
</tr>
<tr>
<td>depth + gray</td>
<td>concat</td>
<td>5.6 ± 5.0</td>
<td>4.1 ± 5.0</td>
<td>9.8 ± 13.4</td>
<td>0.698</td>
</tr>
<tr>
<td>depth + OF</td>
<td>concat</td>
<td>6.0 ± 6.1</td>
<td>4.5 ± 4.8</td>
<td>9.2 ± 11.5</td>
<td>0.690</td>
</tr>
<tr>
<td>POSEidon</td>
<td>conv</td>
<td>5.7 ± 5.6</td>
<td>4.9 ± 5.1</td>
<td>9.0 ± 11.9</td>
<td>0.715</td>
</tr>
<tr>
<td>POSEidon</td>
<td>concat</td>
<td>6.3 ± 6.1</td>
<td>5.0 ± 5.0</td>
<td>10.6 ± 14.2</td>
<td>0.657</td>
</tr>
<tr>
<td>POSEidon</td>
<td>mul+concat</td>
<td>5.6 ± 5.6</td>
<td>4.9 ± 5.2</td>
<td>9.1 ± 11.9</td>
<td>0.712</td>
</tr>
</tbody>
</table>
A demo
• 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

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