COMPUTER VISION, AI & ROBOTICS
FOR VISUAL INTELLIGENCE

IROS 2018 Madrid, Keynote

Rita Cucchiara

AImageLab, Dipartimento di Ingegneria «Enzo Ferrari»

Università di Modena e Reggio Emilia, ITALIA
COMPUTER VISION

is a commodity
(i.e. a working software)

as a sense for ROBOTS

ROBOT

is a commodity
(i.e. a programmable, moving object)
as a moving sensor for VISION
Vision and AI are not only the 4 Rs
(Recognition, Reconstruction, Registration, Reorganization)

New computer vision results will provide computers, autonomous systems, robots, cars and objects with visual intelligence, towards an embodied intelligence
the ability to see what's there that others don't, to see what's not there that should be, to see the positives and the negatives, the opportunity, the invention, the upside, the warning signs, the quickest way, the way out, the win

the ability to visualize the world accurately, understand quickly the saliency and the important aspect, modify their surroundings based upon their perceptions, and recreate the aspects of their visual experiences.

People with high visual-spatial intelligence are good at remembering images, faces, and fine details. They are able to visualize objects from different angles and predict situations...
Can I take the spritz glass with overturning the plate?
Where do you put your attention?

What do you predict while driving?
What are these person doing?

How can I run out of the room?

Is that person going to move the chair?
1. SEE THE SALIENT (AND TELL ME BETTER WHAT YOU SEE)
   Thanks to Marcella Cornia, Lorenzo Baraldi, Andrea Palazzi [CVPR17, TMM17, TIP18, TPAMI18..]

2. SEE THE INVISIBLE (SEGMENT BY MOTION)
   Thanks to Guido Borghi and Roberto Vezzani [CVPR17, BMVC18, TPAMI19..]

3. SEE IN THE DARK (DOMAIN TRANSFER FROM RGB TO DEPTH)
   Thanks to Stefano Alletto, Davide Abati and Simone Calderara [TITS19]

4. SEE THE HIDDEN (BY ALLUCINATING OCCLUSIONS)
   Thanks to Matteo Fabbri, Fabio Lanzi, Simone Calderara [ECCV2018]
FROM VISION TO VISUAL INTELLIGENCE

1. SEE THE SALIENT (AND TELL ME BETTER WHAT YOU SEE)
Saliency: data-driven? memory and knowledge-based driven? or task-driven,?
In Neuroscience SALIENCY:

Saliency detection is considered to be a key **attentional** mechanism that facilitates **learning** and survival by enabling organisms to focus their limited **perceptual** and **cognitive** resources on the most pertinent subset of the available **sensory data**.

- **Itti and Koch**, The SALIENCY MAP (IEEE Trans on PAMI ’89)

**Two forms of visual attention:**

- Initial Bottom-up purely data driven
- Refined Task-driven and purposive
COMPUTATIONAL DETECTION OF SALIENCY MAPS BEFORE DL

LOW LEVEL FEATURES

• ‘80—2000 Itti Koch: combination color+gradient+orientation in a winner-take-all unsupervised neural network
• 2006 NIPS Perona et al. Graph-based Visual Saliency as a graph of low level features

ADDING MEMORY and knowledge-based higher level FEATURES (Faces, people, text..)

• 2009 ICCV Torralba et al
• 2012 ICPR Biorg ICPR
• 2013 ICCV Sclaroff et al.

Thanks to Marcella Cornia
THE DEEP LEARNING ERA

**Problems of annotated Data**

*2014 ICCV Vig et al.:* a three Convnet layer

*2015 ICRLW Kummerer et al.:* DeepGaze I with Alexnet (then 2016 ArXiv: Deepgaze II with VGG19)

*2015 CVPR Lin et al.:* data augmentation with image similarity

**DATASETS:**

MIT300 (Itti, Torralba et al) more than 70 competitors since 2014

SALICON (Jiang et al 2015), 10000 images;
	new competition CVPR-LSUN 2017

**TUTORIALs AND COMPETITIONS:**

2016 ECCV tutorial on Saliency

2017 ICME Competition on 360° Saliency

2017 CVPR New SALICON Competition

2018 CVPR competition

ARE COMPETITIONS/DATASET USEFUL?

**SALICON benchmark**

mit saliency benchmark
Trained in SALICON and in MIT300.. Now the net is exploring the world
SALIENCY DETECTION @AIMAGELAB ML-NET*

* [M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara. A Deep Multi-Level Network for Saliency Prediction, ICPR 2016.]
Saliency attentive Model
SAM*
Almagelab 2016-2018

* M. Cornia, L. Baraldi, G. Serra, R. Cucchiara
“Predicting Human Eye Fixations via an LSTM-based Saliency Attentive Model”
IEEE Transactions on Image Processing, 2018
SALIENCY DETECTION @IMAGELAB SAM

Saliency Attentive Model (SAM): ML-NET+ LSTMs

As a sort of Pre-attentive scan-path

The IDEA: define a new **CONV-LSTM** for scan the space and not the time
PERFORMANCE ANALYSIS

**SALICON Dataset (original release)**

<table>
<thead>
<tr>
<th></th>
<th>CC</th>
<th>sAUC</th>
<th>AUC</th>
<th>NSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAM</td>
<td>0.842</td>
<td>0.779</td>
<td>0.883</td>
<td>3.204</td>
</tr>
<tr>
<td>ML-Net [1]</td>
<td>0.743</td>
<td>0.768</td>
<td>0.866</td>
<td>2.789</td>
</tr>
<tr>
<td>SU [2]</td>
<td>0.780</td>
<td>0.760</td>
<td>0.880</td>
<td>2.610</td>
</tr>
<tr>
<td>SalNet [3]</td>
<td>0.622</td>
<td>0.724</td>
<td>0.858</td>
<td>1.859</td>
</tr>
<tr>
<td>DeepGazeII [4]</td>
<td>0.509</td>
<td>0.761</td>
<td>0.885</td>
<td>1.336</td>
</tr>
</tbody>
</table>

**SALICON Dataset (new release)**

<table>
<thead>
<tr>
<th></th>
<th>CC</th>
<th>sAUC</th>
<th>AUC</th>
<th>NSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAM</td>
<td>0.899</td>
<td>0.741</td>
<td>0.865</td>
<td>1.990</td>
</tr>
</tbody>
</table>


SAM : winner in the LSUN Challenge CVPR 2017
SALICON (original release)

Image

Groundtruth

SAM

SALICON (new release)

Groundtruth

SAM

Learning adaptation to data changes
Actions in the Eye (Hollywood2) dataset

<table>
<thead>
<tr>
<th></th>
<th>CC</th>
<th>Similarity</th>
<th>AUC</th>
<th>NSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAM</td>
<td>0.694</td>
<td>0.574</td>
<td>0.922</td>
<td>3.202</td>
</tr>
<tr>
<td>RMDN [1]</td>
<td>0.613</td>
<td>0.535</td>
<td>0.904</td>
<td>2.646</td>
</tr>
</tbody>
</table>

Few approaches to video

Bottom-up saliency

Task-driven saliency
Are the drivers looking at people when driving?

THE DR(EYE)VE PROJECT

Aimagelab.unimore.it/Dreyeve

• 3 months for acquisition
• Eytrackers and camera-car
• SIFT-based image registration frame by frame
• Automatic annotation of gaze, speed, GPS position

DR(eye)VE
a dataset for attention-based tasks with applications to autonomous and assisted driving.
DR(EYE)VE DATASET @AIMAGELAB [T PAMI2018]
(HUMAN) ATTENTIVE BEHAVIOR, MEASURED

Drivers do not care pedestrians..

A. Palazzi, D-Abati, S. Calderara, F. Solera, R. Cucchiara. «Predicting the Driver’s Focus of Attention: the DR(EYE)veProject IEEE Transactions on PAMI 2018
There are pedestrians around. How many? Is the DL-based detection correct?
There are pedestrians around.. How many? Where they are going?

We need Detection and Tracking together for Visual Intelligence.
USE SALIENCY FOR A BETTER TEXTUAL DESCRIPTION
AUTOMATIC IMAGE/VIDEO CAPTIONING

A fantastic compress form of description

From VISUAL DATA

To TEXT SEQUENCE

From MS COCO
GT1: a woman is slicing potatoes
GT2: a woman is cutting a potato into small pieces
GT3: a person is slicing a potato into pieces
GT4: a woman is slicing potatoes
GT5: a woman is cutting a potato

Pr: a person is cutting a potato
MANY ATTEMPTS TO CAPTIONING

.. a white shark swims in the ocean water..

CONV-NET + Recurrent NET (LSTM)

IMPORTANT REFERENCES
IMPROVING SENTENCE GENERATION IN LONG VIDEO

A new architecture for video captioning with the capacity of forgetting

RATIONALE:
Video captioning must be aware of the structure, not to mix words of consecutive shots, thanks to forget/reset mechanism

Keep in mind the consecutio temporum

A long training but now in real time

[L. Baraldi, C. Grana, and R. Cucchiara, Hierarchical Boundary-Aware Neural Encoder for Video Captioning CVPR 2017]
A SUITABLE MODIFICATION OF LSTM WITH BOUNDARY DETECTION

In a single end-to-end pipeline; if a boundary is detected, thus the memory is reset

Features from
ResNet 50 (Imagenet) +
C3D (Sport-1M for motion)

Vocabulary of
• M-VAD 6090 words (84.6 hours of 92 Hollywood movies, 46K video clips for impairs)
• MPII-MD 7198 words (94HD movies with 68K sentences)
• MSVD 4125 words (Microsoft 2K Youtube clips with 85K sentences)
• + <BOS> and <EOS>
<table>
<thead>
<tr>
<th>Model</th>
<th>METEOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>SA-GoogleNet+3D-CNN [49]</td>
<td>4.1</td>
</tr>
<tr>
<td>HRNE [22]</td>
<td>5.8</td>
</tr>
<tr>
<td>S2VT-RGB(VGG) [43]</td>
<td>6.7</td>
</tr>
<tr>
<td>HRNE with attention [22]</td>
<td>6.8</td>
</tr>
<tr>
<td>Venugopalan et al. [42]</td>
<td>6.8</td>
</tr>
<tr>
<td>LSTM encoder (C3D+ResNet)</td>
<td>6.7</td>
</tr>
<tr>
<td>Double-layer LSTM encoder (C3D+ResNet)</td>
<td>6.7</td>
</tr>
<tr>
<td>Boundary encoder on shots</td>
<td>7.1</td>
</tr>
<tr>
<td>Boundary-aware encoder (C3D+ResNet)</td>
<td><strong>7.3</strong></td>
</tr>
</tbody>
</table>

Table 1. Experiment results on the M-VAD dataset.

<table>
<thead>
<tr>
<th>Model</th>
<th>B@4</th>
<th>M</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>SA-GoogleNet+3D-CNN [49]</td>
<td>41.9</td>
<td>29.6</td>
<td>-</td>
</tr>
<tr>
<td>LSTM-YT [44]</td>
<td>33.3</td>
<td>29.1</td>
<td>-</td>
</tr>
<tr>
<td>S2VT [43]</td>
<td>-</td>
<td>29.8</td>
<td>-</td>
</tr>
<tr>
<td>LSTM-E [23]</td>
<td>45.3</td>
<td>31.0</td>
<td>-</td>
</tr>
<tr>
<td>HRNE [22]</td>
<td>46.7</td>
<td>33.9</td>
<td>-</td>
</tr>
<tr>
<td>Boundary-aware encoder</td>
<td>42.5</td>
<td>32.4</td>
<td>63.5</td>
</tr>
</tbody>
</table>

Table 3. Experiment results on the MSVD dataset.

<table>
<thead>
<tr>
<th>Model</th>
<th>CIDEr</th>
<th>B@4</th>
<th>R_L</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMT (best variant) [30]</td>
<td>8.1</td>
<td>0.5</td>
<td>13.2</td>
<td>5.6</td>
</tr>
<tr>
<td>SA-GoogleNet+3D-CNN [49]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>5.7</td>
</tr>
<tr>
<td>Venugopalan et al. [42]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>6.8</td>
</tr>
<tr>
<td>Rohrbach et al. [29]</td>
<td>10.0</td>
<td><strong>0.8</strong></td>
<td>16.0</td>
<td><strong>7.0</strong></td>
</tr>
<tr>
<td>LSTM encoder (C3D+ResNet)</td>
<td>10.5</td>
<td>0.7</td>
<td>16.1</td>
<td>6.4</td>
</tr>
<tr>
<td>Double-layer LSTM encoder (C3D+ResNet)</td>
<td>10.6</td>
<td>0.6</td>
<td>16.5</td>
<td>6.7</td>
</tr>
<tr>
<td>Boundary encoder on shots</td>
<td>10.3</td>
<td>0.7</td>
<td>16.3</td>
<td>6.6</td>
</tr>
<tr>
<td>Boundary-aware encoder (C3D+ResNet)</td>
<td><strong>10.8</strong></td>
<td><strong>0.8</strong></td>
<td><strong>16.7</strong></td>
<td><strong>7.0</strong></td>
</tr>
</tbody>
</table>

Table 2. Experiment results on the MPII-MD dataset.

![Example results on the MSVD dataset.](image)
GT: She gets out.
**LSTM encoder:** Someone stops.
**BA encoder (ours):** Someone gets out of the car.

GT: Shakes his head.
**LSTM encoder:** Someone gives her gaze.
**BA encoder (ours):** Someone looks at someone who shakes his head.

GT: He slows down in front of one house with a garage and box tree on the front.
**LSTM encoder:** Someone gets out of the car and walks out of the house.
**BA encoder (ours):** Someone drives up to the house.
PUTTING ALL TOGETHER: SALIENCY AND CAPTIONING
Video captioning.. and image captioning
A still un-completely solved problem

GT: he stands and offers her the small bouquet
Pr: someone looks up at someone

Attention in persons and not in the salient objects!

[M. Cornia, L.Baraldi, G. Serra, R.Cucxhiara Paying More Attention to Saliency: Image Captioning with Saliency and Context Attention ACM Transactions TOMM 2018]
EXPLOITING “MACHINE ATTENTION”

- **Machine attention mechanism**: a way of obtaining time-varying inputs for recurrent architectures.
- At each timestep the attention mechanism selects a region of the image, based on the previous LSTM state, and feeds it to the LSTM.
- The generation of a word is conditioned on that specific region, instead of being driven by the entire image.

---

Adding Saliency and Context as a bottom-up pre-attentive behavior

Some Experiments:
WHAT IS HIT BY SALIENCY (SAM)

Pascal c.a. 20K images, 400 labels
Cityscapes c.a. 5K images, 30 classes
LIP c.a. 50K images, 19 human body parts

1. Saliency looks at objects and less than 10% at background
2. Saliency is independent on the object size (thus importance is not related to size)

M Cornia, L. Baraldi, G. Serra, R. Cucchiara
Paying More Attention to Saliency: Image Captioning with Saliency and Context Attention
RATIONALE 3 FINAL PROPOSAL
ResNet -50, trained with Imagenet; 49 layers, output 2048 channel, +1 conv layer refined in the dataset → 512 filters
SAM (Saliency Attentive Model) defines saliency map and the contextual map. They are input for two LSTM for “Soft attention”, trained with the same weight.
QUALITATIVE RESULTS

GT: A large passenger jet sitting on top of an airport runway.
*With saliency & context:* A large jetliner sitting on top of an airport runway.
*Without:* A large airplane on a runway.

GT: Family of five people in a green canoe on a lake.
*With saliency & context:* A group of people sitting on a boat in a lake.
*Without:* A group of people sitting on top of a boat.

GT: Two people in Swarthmore College sweatshirts are playing frisbee.
*With saliency & context:* A man and a woman are playing frisbee on a field.
*Without:* A man standing next to a man holding a frisbee.
EXPERIMENTAL RESULTS

SALICON Dataset

<table>
<thead>
<tr>
<th></th>
<th>CNN</th>
<th>BLEU@1</th>
<th>BLEU@2</th>
<th>BLEU@3</th>
<th>BLEU@4</th>
<th>METEOR</th>
<th>ROUGE\textsubscript{L}</th>
<th>CIDEr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soft Attention</td>
<td>VGG-16</td>
<td>0.680</td>
<td>0.501</td>
<td>0.358</td>
<td>0.256</td>
<td>0.222</td>
<td>0.497</td>
<td>0.691</td>
</tr>
<tr>
<td>Saliency-Guided Attention</td>
<td>VGG-16</td>
<td>0.682</td>
<td>0.505</td>
<td>0.361</td>
<td>0.258</td>
<td>0.223</td>
<td>0.497</td>
<td>0.694</td>
</tr>
<tr>
<td>Saliency-Guided Att. (with GT saliency maps)</td>
<td>VGG-16</td>
<td>0.684</td>
<td>0.503</td>
<td>0.360</td>
<td>0.257</td>
<td>0.224</td>
<td>0.501</td>
<td>0.696</td>
</tr>
<tr>
<td>Soft Attention</td>
<td>ResNet-50</td>
<td>0.700</td>
<td>0.523</td>
<td>0.379</td>
<td>0.274</td>
<td>0.235</td>
<td>0.510</td>
<td>0.771</td>
</tr>
<tr>
<td>Saliency-Guided Attention</td>
<td>ResNet-50</td>
<td>0.709</td>
<td>0.534</td>
<td>0.388</td>
<td>0.280</td>
<td>0.233</td>
<td>0.513</td>
<td>0.774</td>
</tr>
<tr>
<td>Saliency-Guided Att. (with GT saliency maps)</td>
<td>ResNet-50</td>
<td>0.702</td>
<td>0.527</td>
<td>0.383</td>
<td>0.277</td>
<td>0.236</td>
<td>0.513</td>
<td>0.779</td>
</tr>
</tbody>
</table>

SALICON Dataset (subset of the Microsoft COCO dataset, composed by 20,000 images, largest available dataset for saliency prediction)

Microsoft COCO Dataset

<table>
<thead>
<tr>
<th></th>
<th>CNN</th>
<th>BLEU@1</th>
<th>BLEU@2</th>
<th>BLEU@3</th>
<th>BLEU@4</th>
<th>METEOR</th>
<th>ROUGE\textsubscript{L}</th>
<th>CIDEr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soft Attention</td>
<td>ResNet-50</td>
<td>0.717</td>
<td>0.546</td>
<td>0.402</td>
<td>0.294</td>
<td>0.253</td>
<td>0.529</td>
<td>0.939</td>
</tr>
<tr>
<td>Saliency-Guided Attention</td>
<td>ResNet-50</td>
<td>0.718</td>
<td>0.547</td>
<td>0.404</td>
<td>0.296</td>
<td>0.254</td>
<td>0.530</td>
<td>0.944</td>
</tr>
</tbody>
</table>

Microsoft COCO Dataset (composed by more than 120,000 images divided in training and validation sets each image is annotated with five sentences)
QUALITATIVE RESULTS

With saliency & context: A person riding a motorcycle on a road.

Without: A man on a bike with a bike in the background.
QUALITATIVE RESULTS

With saliency & context: A person taking a picture of himself in a bathroom.

Without: A bathroom with a sink and a sink.
With saliency & context: A man is looking inside of a refrigerator.

Without: A man is making a refrigerator in a kitchen.
From the Loomo Robot to Facebook GPU Server and vice-versa (about 15 frame per seconds)
“two standing women have a phone and a cup”
3. SEE THE HIDDEN (OCCLUSIONS)
THE NEW TREND
DETECTION BY POSE
MANY RELATED WORKS – POSE DETECTION 2017-2019

images
- **Associative Embedding**: End-to-End Learning for Joint Detection and Grouping [Newell et. Al, NIPS2017]
- **RMPE**: Regional Multi-person Pose Estimation [Fang et. Al, ICCV2018]
- **Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields**: [Cao et. Al, CVPR2017]
- **Cascaded Pyramid Network** for Multi-Person Pose Estimation [Chen et. Al, CVPR2018]

video
- **ArtTrack**: Articulated Multi-person Tracking in the Wild [Insafutdinov et. Al. CVPR2017]
- **PoseTrack**: Joint Multi-Person Pose Estimation and Tracking [Iqbal et al. CVPR2017, ECCV2018]
- Simple, efficient and effective **keypoint tracking** [Girdhar et. Al, ICCVws2017]
- **Towards Multi-Person Pose Tracking**: Bottom-up and Top-down Methods [Jin et. Al ICCVws2017]
- **OpenPose**: Real-time multi-person keypoint detection library for body, face, and hands estimation [S. Wei et al CVPR 2016, → CVPR 2017, CVPR 2018],
Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields [CVPR 2017]

- Two-branch multi-stage CNN inspired by CMP for 2D multi person-pose estimation that jointly learns and parts association
- Introduces a novel bottom-up approach for joints association via Part Affinity Fields (PAFs), a set of 2D vector fields that encode the location and orientation of limbs in the image domain.

[S. Wei V. Ramakrishna T.Kanade and Y. Sheikh Convolutional pose machines  CVPR 2017]
WHERE ARE THE MAIN CHALLENGES?

1. Reliable **DL architectures** for pose detection
2. Coping with **occlusions** (and body self-occlusions),
3. A global spatio-temporal association for detection, tracking and then prediction

1. A Large and General Annotated dataset
## RECENT ANNOTATED DATASETS FOR POSE DETECTION AND/OR TRACKING

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#Clips</th>
<th>#Frames</th>
<th>#PpF</th>
<th>3D</th>
<th>Occlusion</th>
<th>Tracking</th>
<th>Pose</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Penn Action</td>
<td>2,326</td>
<td>159,633</td>
<td>1</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>Sports</td>
</tr>
<tr>
<td>JHMDB</td>
<td>5,100</td>
<td>31838</td>
<td>1</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>Diverse</td>
</tr>
<tr>
<td>YouTube Pose</td>
<td>50</td>
<td>5,000</td>
<td>1</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>Diverse</td>
</tr>
<tr>
<td>Video Pose 2.0</td>
<td>44</td>
<td>1,286</td>
<td>1</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>Diverse</td>
</tr>
<tr>
<td>MOT-16</td>
<td>14</td>
<td>11,235</td>
<td>6-51</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>Urban</td>
</tr>
</tbody>
</table>

---

Penn Action | JHMDB | YouTube Pose | YouTube Pose 2.0
Whenever data annotation is extremely expensive…

Synthetic data become THE SOLUTION
GTA V SCRIPTHOOK

- Photorealistic
- Plausible dynamics
- Lifelike entity AI

+ Access to native GTA functions
+ Customizable
+ Extract all the information available to the game engine

+ PhD male students
EXTRACTING DATA FROM GAME ENGINE
• Bounding boxes
EXTRACTING DATA FROM GAME ENGINE

- Bounding boxes
- Instances
EXTRACTING DATA FROM GAME ENGINE

- Bounding boxes
- Instances
- Reciprocal distances
EXTRACTING DATA FROM GAME ENGINE

- Bounding boxes
- Instances
- Reciprocal distances
- Joints coordinates
THE NEW JTA DATASET BY AIMAGELAB UNIMORE 2018 [ECCV18]

About the Dataset

JTA (Joint Track Auto) is a huge dataset for pedestrian pose estimation and tracking in urban scenarios created by exploiting the highly photorealistic video game *Grand Theft Auto V* developed by Rockstar North. We collected a set of 512 full-HD videos (256 for training and 256 for testing), 30 seconds long, recorded at 30 fps.
### THE NEW JTA DATASET

<table>
<thead>
<tr>
<th></th>
<th>JTA</th>
<th>Posetrack</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data type</td>
<td>Synthetic</td>
<td>Real</td>
</tr>
<tr>
<td>#Clips</td>
<td>512</td>
<td>514</td>
</tr>
<tr>
<td>#Frames</td>
<td>&gt;460,000</td>
<td>&gt;22,000</td>
</tr>
<tr>
<td>#Poses</td>
<td>9,836,194</td>
<td>153,615</td>
</tr>
<tr>
<td>#PpF</td>
<td>0-60</td>
<td>1-13</td>
</tr>
<tr>
<td>3D</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Occlusion</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Tracking</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Pose</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Type</td>
<td>Urban</td>
<td>Diverse</td>
</tr>
<tr>
<td>Pose variation</td>
<td>Low</td>
<td>High</td>
</tr>
</tbody>
</table>

The J-GTA Dataset is available at [www.aimagelab.unimore.it/dataset](http://www.aimagelab.unimore.it/dataset)
Contact Matteo.Fabbri@unimore.it
NOW A DL-BASED PROPOSAL

1. Reliable DL architecture for pose detection
2. Coping with occlusions (and body self-occlusions)
3. A global spatio-temporal association for tracking
4. Large and General Annotated dataset

Starting from CPM by CMU
THOPA-NET FOR POSE ESTIMATION AND TRACKING

Detection by joints Heatmaps, Occlusions, Part Affinity Fields (PAFs) and Temporal Affinity Fields (TAF)

Tracking is then a NN-association by minimizing TAF scores in short term.

\[
l_x = \sum_{i} \sum_{x=1}^{n'} \sum_{y=1}^{n'} M(x,y) \odot (X_i^*(x,y) - X_i^*(x,y))^2
\]

\[
L = \sum_{s=1}^{S} (l_H^s + l_O^s + l_P^s + l_T^s)
\]

[M.Fabbri, F.Lanzi, S.Calderara, A.Palazzi, R.Vezzani, R.Cucchiara Learning to Detect and Track Visible and Occluded Body Joints in a Virtual World ECCV 2018]
HEATMAPS, OCCLUSIONS, PAFS AND TAFS

Heatmaps model the part locations and occlusions as Gaussian peaks in the map; One heatmap for each type of joint (eg. “nose”, “neck”, “left shoulder”, …) with Scale awareness without the need of multi-scale branches or pyramid inputs

The PAFs (Part Affinity Fields) are 2D versor encoding the direction that points from one joint to the other.

The TAF (Temporal Affinity Field) links corresponding joints of the same person in consecutive frames (for an unknown number of people).
Spatio-temporal association of joints to person is given by an optimization taking into account PAFs, TAFs and human anatomy constraints.

Automatically annotated joints, occluded joints, connections (as PAF) and temporal connections (TAF)

Detected joints, occluded joints, connections (as PAF) and temporal connections (TAF) (color means direction)
The network is able to **hallucinate** the position of not visible joints.

**Table 2.** Detection results on JTA Dataset

<table>
<thead>
<tr>
<th>Joints</th>
<th>Mean Average Prec.</th>
<th>Detection</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single Image no occ</td>
<td>50.9</td>
<td></td>
<td>81.5</td>
<td>64.1</td>
<td>71.6</td>
</tr>
<tr>
<td>Single Image + occ Complete</td>
<td>56.3</td>
<td></td>
<td>87.9</td>
<td>71.8</td>
<td>78.4</td>
</tr>
<tr>
<td>Complete</td>
<td><strong>59.3</strong></td>
<td></td>
<td><strong>92.1</strong></td>
<td><strong>77.4</strong></td>
<td><strong>83.9</strong></td>
</tr>
<tr>
<td>[7]</td>
<td>50.1</td>
<td></td>
<td>86.3</td>
<td>55.8</td>
<td>69.5</td>
</tr>
</tbody>
</table>

[7] [Cao et al. CVPR2017]
RESULTS – DETECTION AND TRACKING BY POSE AT ECCV2018

Table 3. Results on JTA dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>MOTA</th>
<th>IDF1</th>
<th>MT</th>
<th>ML</th>
<th>FP</th>
<th>FN</th>
<th>IDs</th>
<th>FRAG</th>
</tr>
</thead>
<tbody>
<tr>
<td>[3] + our det</td>
<td>57.4</td>
<td>57.3</td>
<td>45.3</td>
<td>21.7</td>
<td>40096</td>
<td>103831</td>
<td>15236</td>
<td>15560</td>
</tr>
<tr>
<td>[3] + DPM det</td>
<td>31.5</td>
<td>27.6</td>
<td>25.3</td>
<td>41.7</td>
<td>80096</td>
<td>170662</td>
<td>10575</td>
<td>19069</td>
</tr>
<tr>
<td>THOPA-net</td>
<td>59.3</td>
<td>63.2</td>
<td>48.1</td>
<td>19.4</td>
<td>40096</td>
<td>103662</td>
<td>10214</td>
<td>15211</td>
</tr>
</tbody>
</table>

Experiments of JTA THOPAnet in short term tacking (1 sec) w.r.t the detection and a standard nn tracking, and w.r.t a classical people detector+ nn tracking.

Table 4. Results on MOT-16 benchmark ranked by MOTA score

<table>
<thead>
<tr>
<th>Method</th>
<th>MOTA</th>
<th>IDF1</th>
<th>MT</th>
<th>ML</th>
<th>FP</th>
<th>FN</th>
<th>IDs</th>
<th>FRAG</th>
</tr>
</thead>
<tbody>
<tr>
<td>[25]</td>
<td>66.1</td>
<td>65.1</td>
<td>34.0</td>
<td>20.8</td>
<td>5061</td>
<td>55914</td>
<td>805</td>
<td>3093</td>
</tr>
<tr>
<td>[26]</td>
<td>61.4</td>
<td>62.2</td>
<td>32.8</td>
<td>18.2</td>
<td>12852</td>
<td>56668</td>
<td>781</td>
<td>2008</td>
</tr>
<tr>
<td>THOPA-net</td>
<td>56.0</td>
<td>29.2</td>
<td>25.2</td>
<td>27.9</td>
<td>9182</td>
<td>67059</td>
<td>4064</td>
<td>5557</td>
</tr>
<tr>
<td>[27]</td>
<td>47.2</td>
<td>46.3</td>
<td>14.0</td>
<td>41.6</td>
<td>2681</td>
<td>92856</td>
<td>774</td>
<td>1675</td>
</tr>
<tr>
<td>[28]</td>
<td>46.0</td>
<td>50.0</td>
<td>14.6</td>
<td>43.6</td>
<td>6895</td>
<td>91117</td>
<td>473</td>
<td>1422</td>
</tr>
<tr>
<td>[29]</td>
<td>43.9</td>
<td>45.1</td>
<td>10.7</td>
<td>44.4</td>
<td>6450</td>
<td>95175</td>
<td>676</td>
<td>1795</td>
</tr>
<tr>
<td>[30]</td>
<td>38.8</td>
<td>42.4</td>
<td>7.9</td>
<td>49.1</td>
<td>8114</td>
<td>102452</td>
<td>965</td>
<td>1657</td>
</tr>
</tbody>
</table>

Experiments of real data
With THOPA-net and final refinement


RESULTS ON TRACKING PEOPLE (JTA)
RESULTS ON TRACKING PEOPLE (DR(EYE)VE)
JUMP PROJECT – ACTION RECOGNITION
Actual, predicted joints, reconstruction (with a specific U-NET)
IN CONCLUSION...WHAT WE LEARNED, RECENTLY.

Computer vision, after about 30 years, with DL is still FASHINATING and COMPLEX

Each result requires TIME, PEOPLE, CPU/GPU, and STUDY

Data are necessary. Big Data are not

God gives bread to those who have no teeth and vice versa... invent solutions.

Vision (also starting from dataset) can be used around... Robotics and automotive are some examples

Enjoy with the others: Meet researchers, work together and organize each event as the best scientific party you can...
SEE YOU AT **ICPR 2020 MILANO, ITALY**
(AND AT **ECCV 2022 TELAVIV, ISRAEL**)

btw..**OPEN POSITIONS** at Aimagelab at University of Modena and Reggio Emilia, ITALY.
Send my your CV!

Thanks
MEASURE IN SALIENCY

Comparison as in [Bylinsky et al. ArXiv 2014]

- Similarity (SIM)
- Correlation Coefficient (CC)
- Area Under the ROC Curve (AUC)
- shuffled version of AUC (sAUC)
- NSS, Normalized Scanpath Saliency (NSS)
- Earth-Mover Distance (EMD).

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Location-based</th>
<th>Distribution-based</th>
</tr>
</thead>
<tbody>
<tr>
<td>Similarity</td>
<td>AUC, sAUC, NSS, IG</td>
<td>SIM, CC</td>
</tr>
<tr>
<td>Dissimilarity</td>
<td>EMD, KL</td>
<td></td>
</tr>
</tbody>
</table>

Image, Itti, ML-NET, GT
BLEU (bilingual evaluation understudy) is an algorithm for evaluating the quality of text which has been machine-translated from one natural language to another. BLEU uses a modified form of precision to compare a candidate translation against multiple reference translations.

METEOR (Metric for Evaluation of Translation with Explicit ORdering) is a metric for the evaluation of machine translation output. The metric is based on the harmonic mean of unigram precision and recall, with recall weighted higher than precision.

ROUGE, or Recall-Oriented Understudy for Gisting Evaluation,[1] is a set of metrics and a software package used for evaluating automatic summarization and machine translation software in natural language processing. ROUGE-N: N-gram[3] based co-occurrence statistics

CIDEn: Consensus-based Image Description Evaluation (from CVPR2016)