Deep Understanding Shopper Behaviors and Interactions Using Computer Vision

Emanuele Frontoni, Rocco Pietrini

v. 05/18/2018

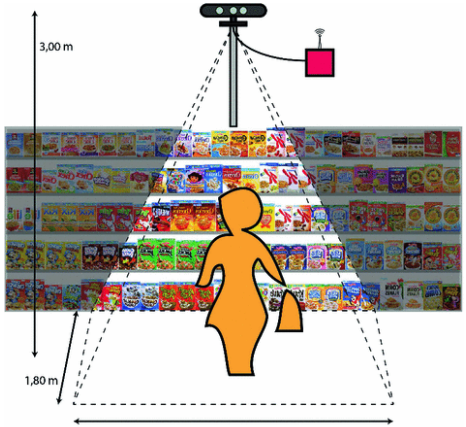
Shoppers Behavior: points of view

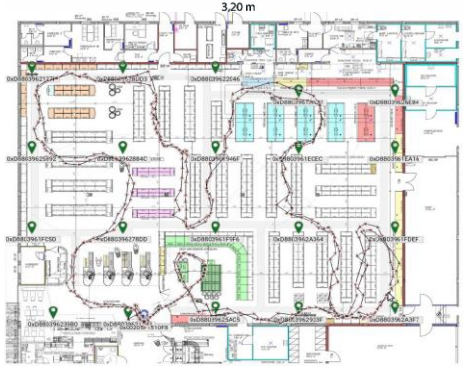
«Micro»

- What people do in front of the shelf?
- How people browse the shelf?

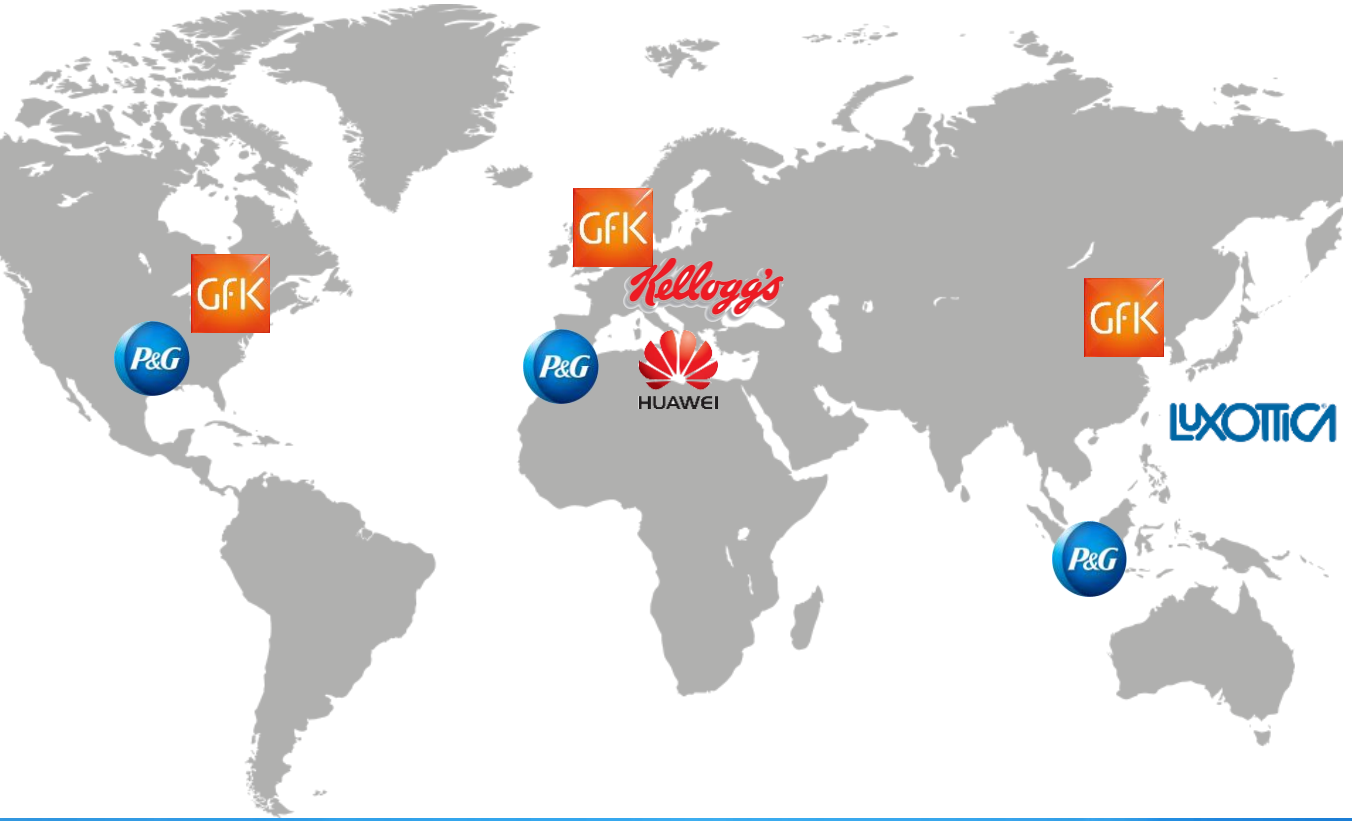
«Macro»

- How do people move across the category?
- How do people move across the store?





Background



Collected Data

2016-2018

People

10.323.710

Interactions

1.104.788

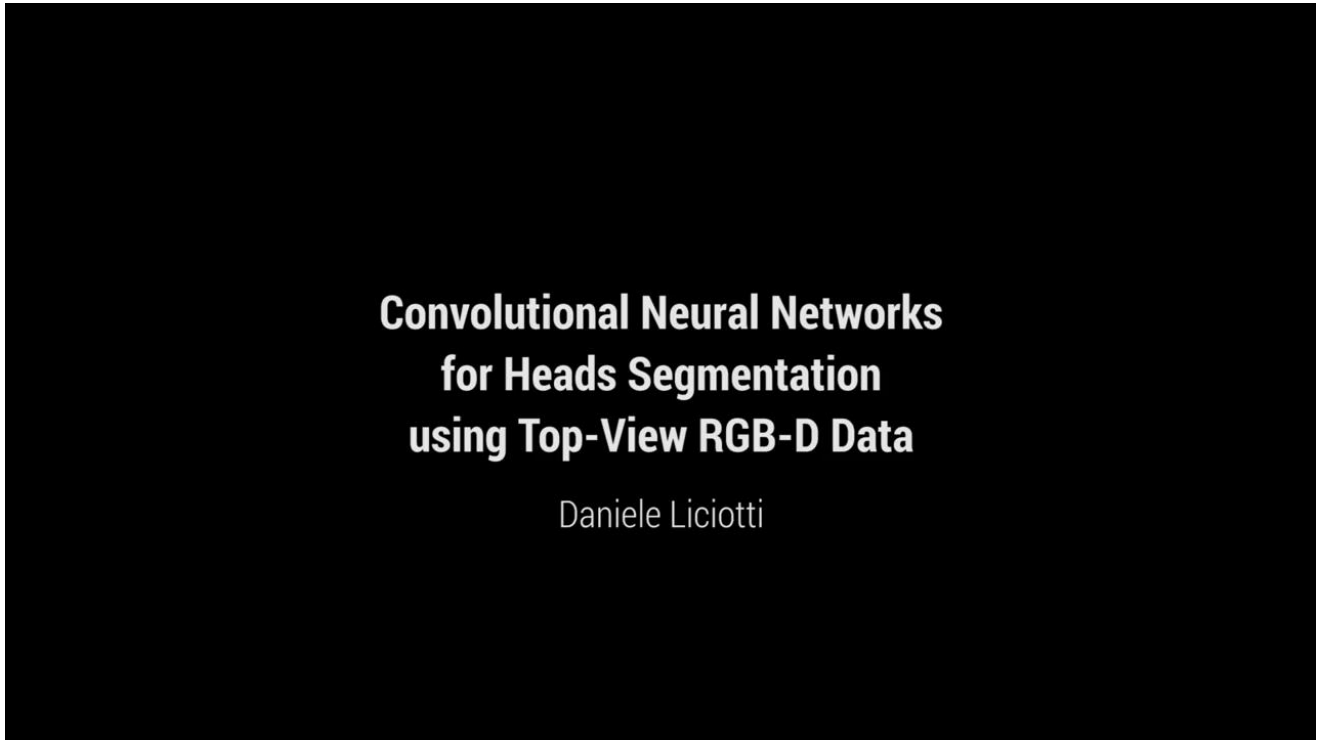
Shopper Analytics

https://www.youtube.com/watch?v=KhIHtKG4Wy8

Why top-view RGBD?

- Privacy compliant
- Occlusion Free
- Discreet, separate from the shelf
- Low resolution, on board elaboration (no recording)
- Live analysis on edge
- Cloud based multi-camera processing

TVHeads: 1815 top view RGBD people images with manually labeled heads location.

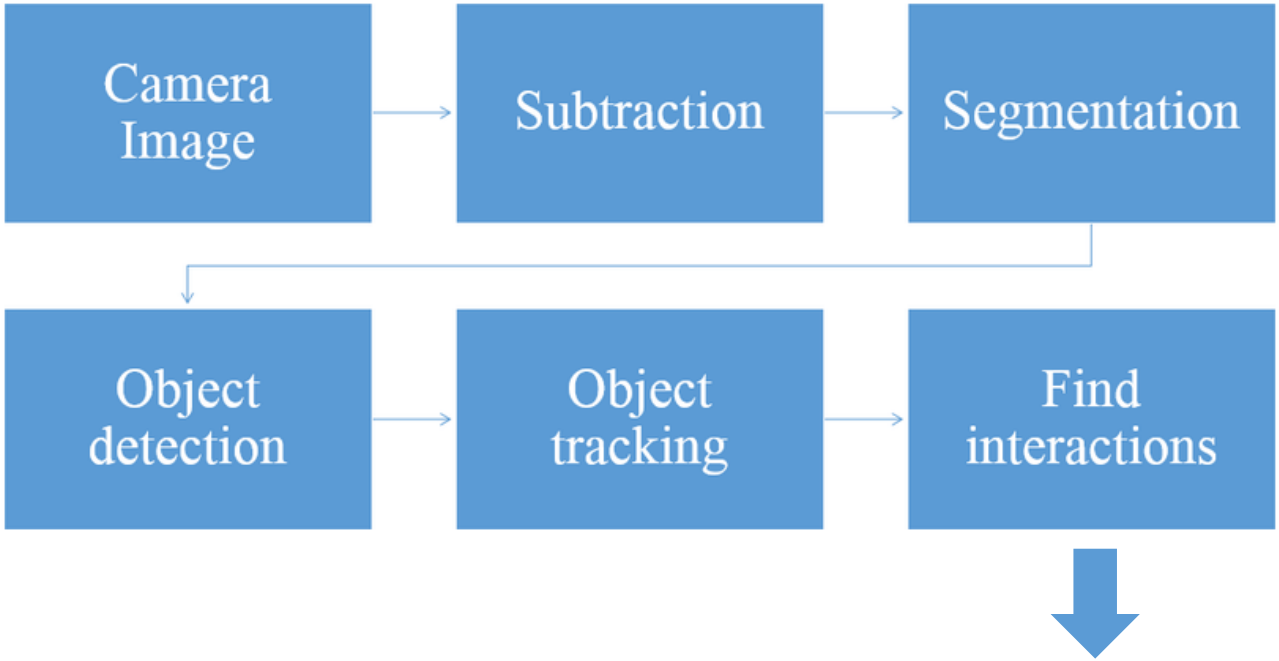


Convolutional Network for semantic Heads Segmentation using top-view RGB-D Data

Table: Jaccard and Dice indices of different CNN architectures.

Net	Bit	Jaccard *Train*	Jaccard *Validation*	Dice *Train*	Dice *Validation*
Fractal [Larsson et al., 2016]	8	0.960464	0.948000	0.979833	0.973306
	16	0.961636	0.947762	0.980443	0.973180
U-Net [Ronneberger et al., 2015]	8	0.896804	0.869399	0.945595	0.930138
	16	0.894410	0.869487	0.944262	0.930188
U-Net2 [Ravishankar et al., 2017]	8	0.923823	0.939086	0.960403	0.968586
	16	0.923537	0.938208	0.960249	0.968119
U-Net3	8	0.962520	0.931355	0.980902	0.964458
	16	0.961540	0.929924	0.980393	0.963690
SegNet [Badrinarayanan et al., 2015]	8	0.884182	0.823731	0.938531	0.903347
	16	0.884162	0.827745	0.938520	0.905756
ResNet [He et al., 2016]	8	0.932160	0.856337	0.964889	0.922609
	16	0.933436	0.848240	0.965572	0.917889

Steps



Interaction classification

geometric comparison vs deep learning

Interaction

Interaction are evaluated using pure geometry vs deep learning and classified as:

Positive

Negative

Neutral

Geometric comparisons drawbacks

- Low accuracy (76,4%)
- Small products (accuracy decrease by 20%)
- Bad surfaces for infrared light (glass, transparent plastic)

- Fake interactions due to:
  - Unintentional interaction threshold crossing
  - Object not present in the background too close to the shelf

→ **Deep Learning Approach** (in the interaction evaluation step)

HaDa Dataset

- Dataset of «hands» gathered from a real store
- +180.000 images 80x80 pixels
- Muti category / multi store / multi country
- Manual labelling & Crowd sourcing
- Network testing and fine tuning
- Evaluation

Hand classification, examples



Hand with a product



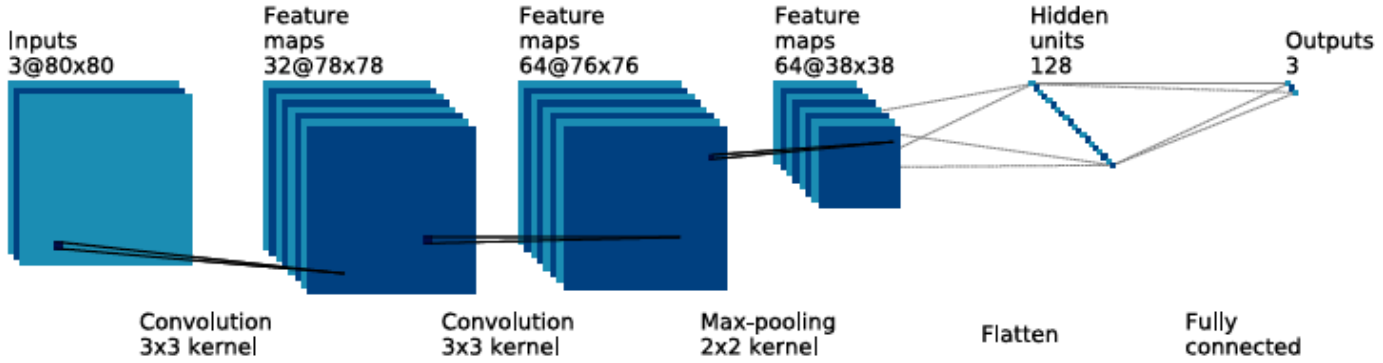
Empty Hand



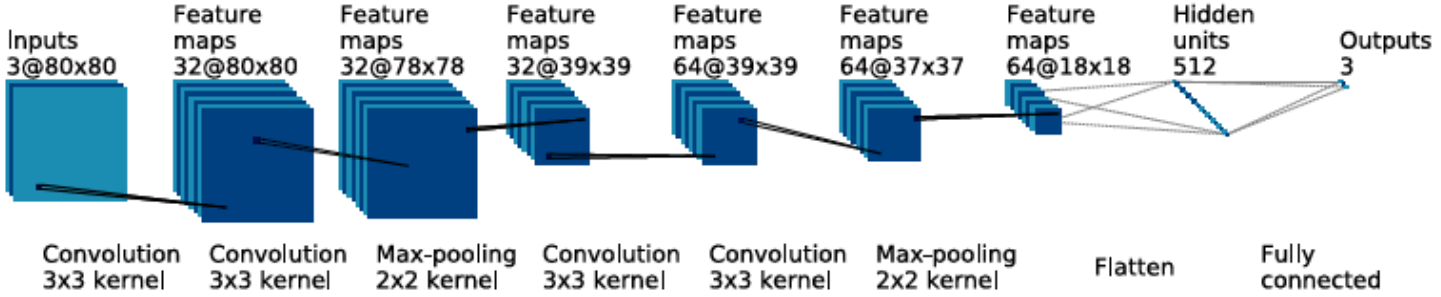
Other

CNNs

CNN



CNN_2

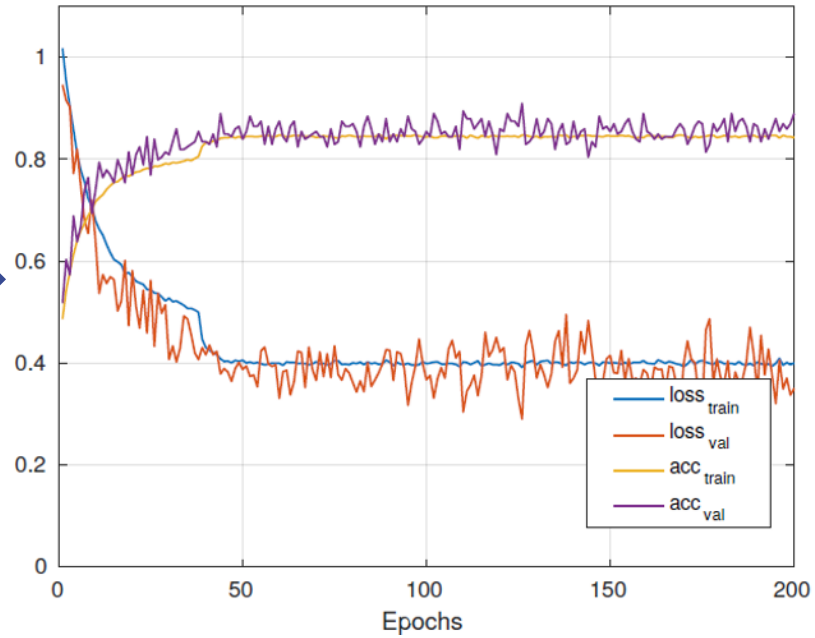


Deep NN accuracy for interaction classification

How good is our network on the hands dataset w.r.t. to known CNN architectures?

Net	Precision	Recall	F1-Score
CNN	0.780691	0.622234	0.691118
CNN_2	0.873640	0.816821	0.843801
AlexNet	0.771158	0.687371	0.726130
CaffeNet	0.899060	0.873149	0.885705

Performance on test set



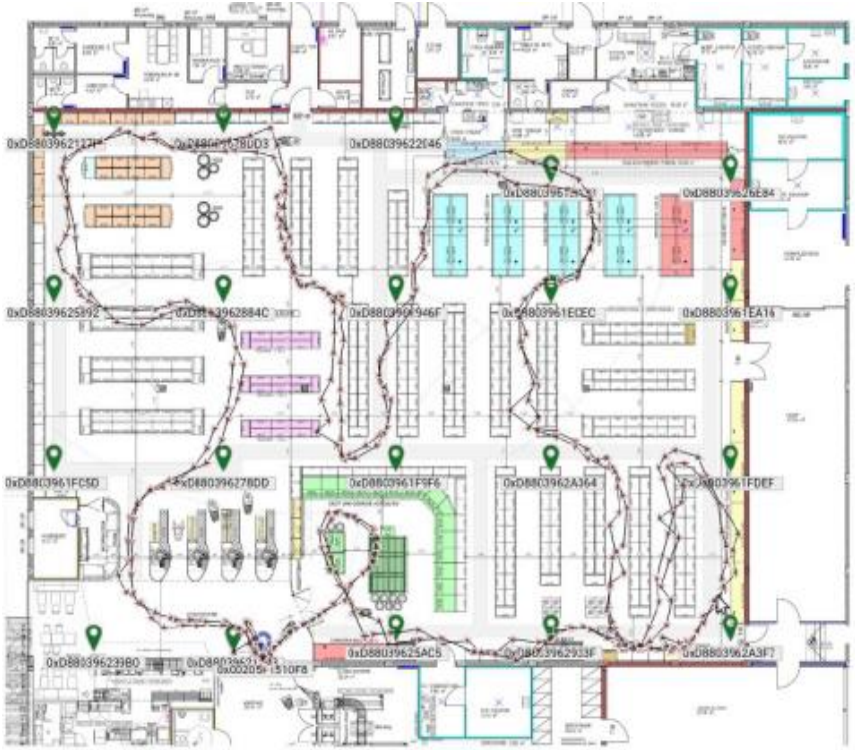
Macro & Re-Id

IDUnicCat

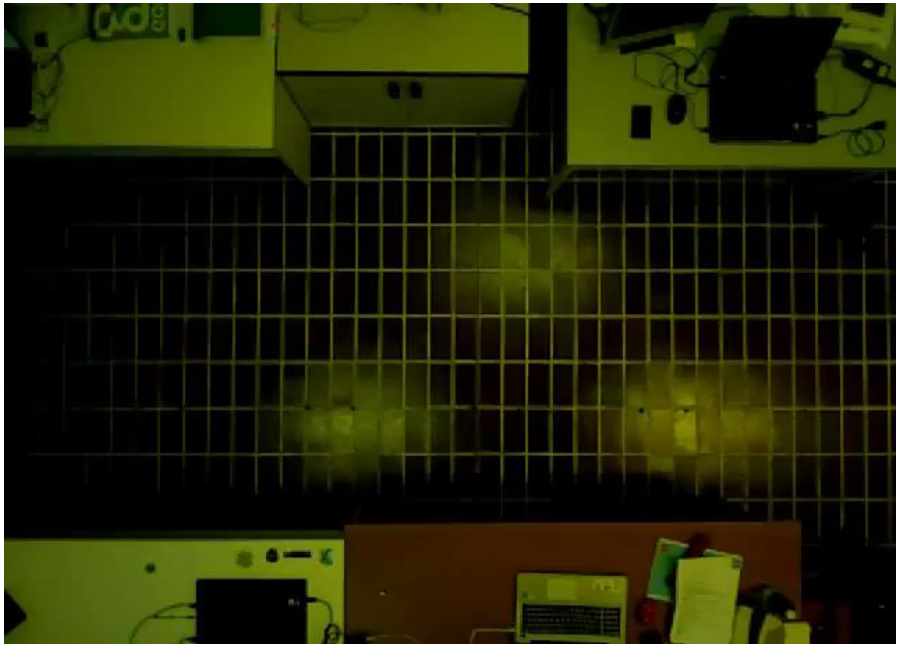
Fully covered area

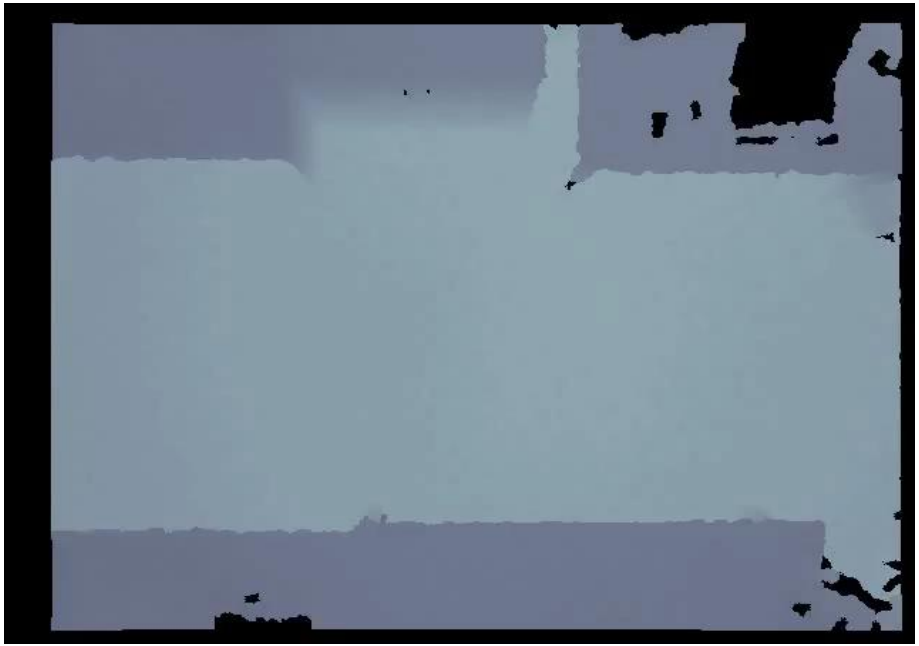
Re-ID

Sparse cameras



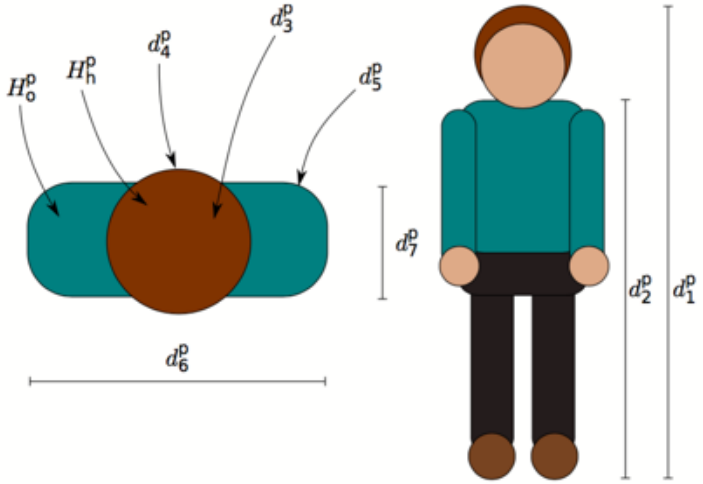
TVPR: Top View Person Re-Identification





Liciotti, D., Paolanti, M., Frontoni, E., Mancini, A., & Zingaretti, P. (2016, December). Person Re-identification Dataset with RGB-D Camera in a Top-View Configuration. In International Workshop on Face and Facial Expression Recognition from Real World Videos (pp. 1-11). Springer, Cham.
Paolanti, M., Liciotti, D., Cenci, A., Frontoni, E., Zingaretti, P. "Person re-identification with RGB-D camera in Top-View configuration" Submitted to Pattern Recognition Letters.

TVPR: Top View Person Re-Identification



- TVH is the colour descriptor:

$$TVH = \{H_h^p, H_o^p\}$$

- TVD is the depth descriptor:

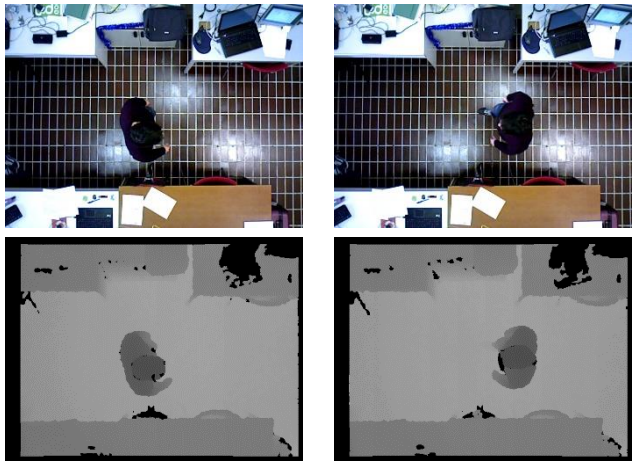
$$TVD = \{d_1^p, d_2^p, d_3^p, d_4^p, d_5^p, d_6^p, d_7^p\}$$

- Finally, $TVDH$ is the signature of a person defined as:

$$TVDH = \{d_1^p, d_2^p, d_3^p, d_4^p, d_5^p, d_6^p, d_7^p, H_h^p, H_o^p\}$$

Liciotti, D., Paolanti, M., Frontoni, E., Mancini, A., & Zingaretti, P. (2016, December). Person Re-identification Dataset with RGB-D Camera in a Top-View Configuration. In International Workshop on Face and Facial Expression Recognition from Real World Videos (pp. 1-11). Springer, Cham.
Paolanti, M., Liciotti, D., Cenci, A., Frontoni, E., Zingaretti, P. "Person re-identification with RGB-D camera in Top-View configuration" Submitted to Pattern Recognition Letters.

TVPR: Top View Person Re-Identification

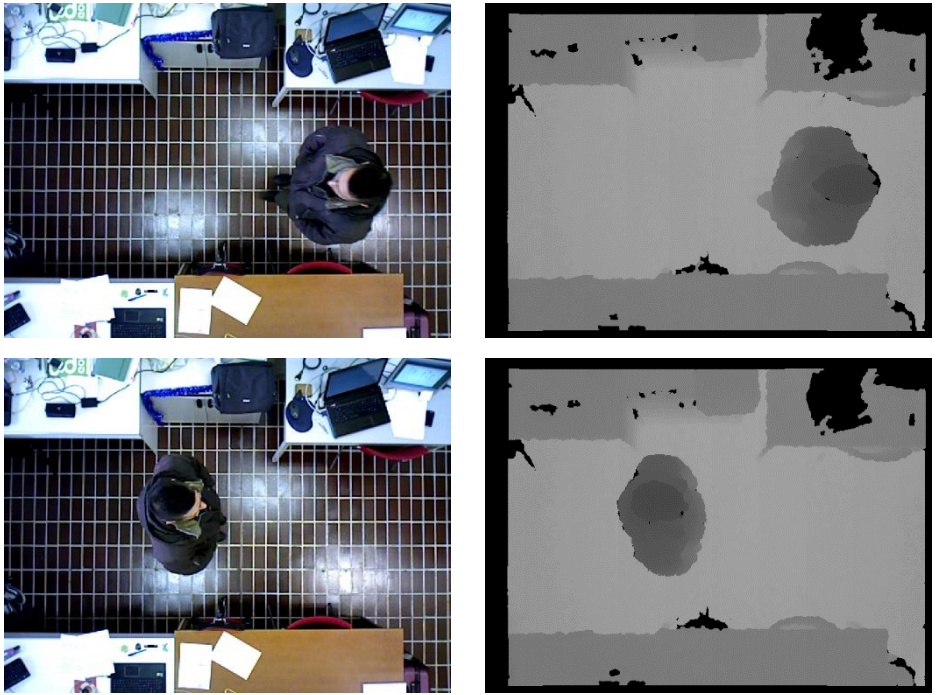


TVD	Classifier	Precision	Recall	F1-score
	KNN	0.35	0.32	0.31
	SVM	**0.48**	**0.43**	**0.42**
	Decision Tree	0.37	0.34	0.33
	Random Forest	0.46	0.43	0.42
TVH	**Classifier**	**Precision**	**Recall**	**F1-score**
	KNN	**0.75**	**0.73**	**0.71**
	SVM	0.70	0.67	0.64
	Decision Tree	0.49	0.46	0.45
	Random Forest	0.71	0.70	0.68
TVDH	**Classifier**	**Precision**	**Recall**	**F1-score**
	KNN	0.81	0.80	0.79
	SVM	**0.85**	**0.85**	**0.83**
	Decision Tree	0.52	0.50	0.48
	Random Forest	0.74	0.71	0.69

Liciotti, D., Paolanti, M., Frontoni, E., Mancini, A., & Zingaretti, P. (2016, December). Person Re-identification Dataset with RGB-D Camera in a Top-View Configuration. In International Workshop on Face and Facial Expression Recognition from Real World Videos (pp. 1-11). Springer, Cham.
Paolanti, M., Liciotti, D., Cenci, A., Frontoni, E., Zingaretti, P. "Person re-identification with RGB-D camera in Top-View configuration" Submitted to Pattern Recognition Letters.

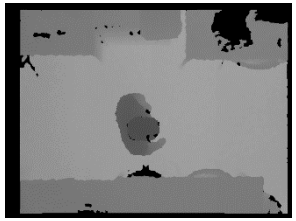
From feature based approach to Deep Learning

- **TVPR/2**: Labelled top view RGBD recordings of **1000** people passing by the camera.
- Resolution 320x240px 30fps

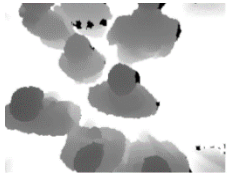


Resources: public datasets available

- **TVPR**: Labelled top view RGBD recordings of **100** people passing by the camera. Resolution 640x480px 30fps.
- **TVPR/2**: Labelled top view RGBD recordings of **1000** people passing by the camera. Resolution 640x480px 30fps.
- **HaDa**: +**180.000** RGBD images 80x80px of hands interacting with a shelf, manually labelled in the 3 categories.
- **TVHeads**: 1815 top view RGBD people images with manually labeled heads location.








http://vrai.dii.univpm.it

VRAI - Vision, Robotics and Artificial Intelligence

Summary

Understanding shopper behaviors is challenging!

Products in a supermarket shows a high variation in terms of dimension and how the shopper interact with. Going from edge computation on depth images using classical computer vision approaches to cloud-based deep learning shows a potential improvement.

A new annotated «hand dataset» has been built and made publicity available. Dataset improvement is however necessary, new examples from other shelf categories, to help newtork generalize.