Improving Measurements of Pedestrian Dynamics using Deep Neural Networks

Alessandro Corbetta*, Vlado Menkovski† and Federico Toschi‡

TU/e Technische Universiteit Eindhoven University of Technology

*Dept. of Applied Physics TU/e, a.corbetta@tue.nl
http://corbetta.phys.tue.nl/
†Dept. of Mathematics and Computer Science, TU/e v.menkovski@tue.nl
‡Dept. of Applied Physics, TU/e f.toschi@tue.nl

NWO Nederlandse Organisatie voor Wetenschappelijk Onderzoek

Introduction

Highly accurate pedestrian trajectory measurements are paramount for quantitative analyses of the crowd behavior and validation of models.

Due to the extreme variability of pedestrian dynamics, statistical resolution is essential for reliable analyses including rare events [1]. Statistical resolution demands for datasets containing tens of thousands of trajectories, i.e. prolonged and continuous monitoring of real environments (cf. setup in Fig. 1, see [2]). **Extracting accurate localization and tracking data** from such campaigns is, *per se*, a scientific and technological **challenge**.

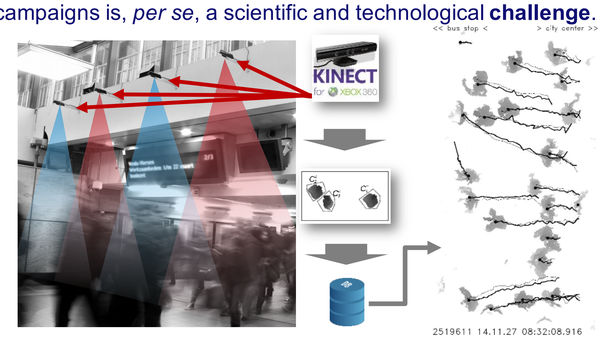


Figure 1: Example of a real-life pedestrian tracking setup at Eindhoven train station. Raw data, as depth maps, is collected via grids of Microsoft KinectTM overhead sensors. Ad hoc, clustering-based algorithms are used for pedestrian localization.

In previous years, overhead Microsoft Kinect sensors successfully allowed real-life data collection [1-4]. Their output, in form of depth maps, is both non privacy intrusive and efficient for the purpose of pedestrian tracking.

Localization algorithms employed so far [1-4] are mostly handcrafted and based on hierarchical clustering. Notably they lose accuracy as the pedestrian density or the scenario complexity (e.g. mixed presence of adults/kids, or of humans/bikes/carts) increase (cf. sample pitfalls in Fig. 2).

Deep Learning (DL) models and, specifically, **Deep Convolutional Neural Networks** (DCNN), do not require handcrafted solutions and demonstrated particular success in image analysis [5].

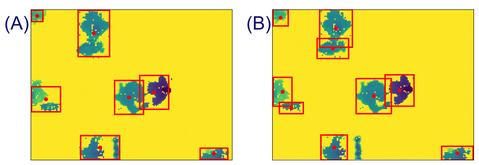


Figure 2: (A) Example scenario of under-performing localization employing a clustering (CL) approach. The comparison with the ground truth in (B) shows some typical mistakes, e.g. the inability to disentangle close subjects, especially when they are small in size (as infants, cf. top and left of the scene). CL approaches cannot distinguish shapes, thus the localization at the bottom includes a static obstacle.

Their hierarchical structure allows models to build complex features and form efficient representation of the input data. However, they require exceeding amount of manually annotated training examples to achieve human-level performance. These annotations are difficult and time consuming to produce.

Motivated by the necessity of performing accurate pedestrian tracking in complex real-life environments, we develop:

1. **A DL model for pedestrian localization in overhead depth maps, which leverages DCNN to achieve a better disentanglement of multiple nearby individuals and/or objects in the scene;**
2. **An efficient model training approach with *weak* supervision from an expert.**

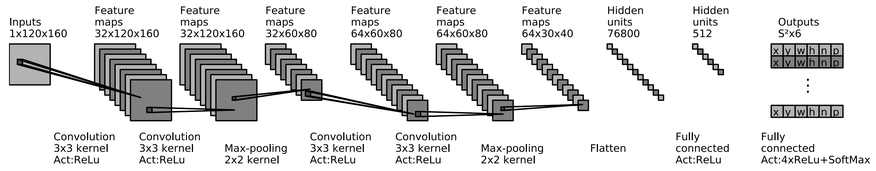


Figure 3: Diagram of the DCNN proposed. Kernel sizes and activation functions (Act) are below each layer.

Method

The architecture of the proposed DCNN model, reported in Fig. 3, is related to the efficient YOLO object localization approach [6]. Its salient features are:

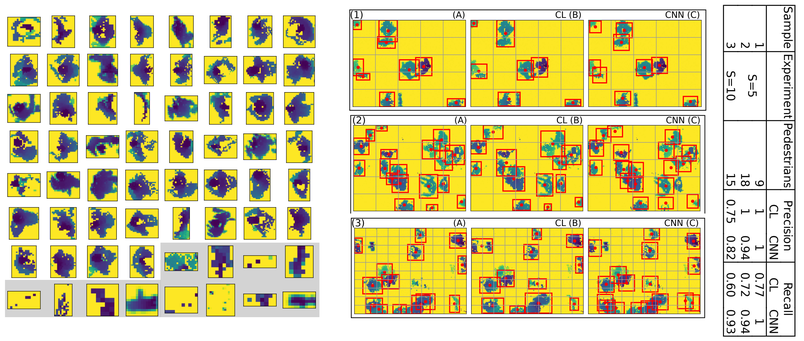
- Localization is performed in a **single pass** producing a set of **bounding boxes** (BB) for each element detected in a depth image.
- The model **overlays a grid** over the image, and produces a **binary detection decision** for each cell in the grid plus the BB coordinates.
- For a $S \times S$ regular grid, the model produces S^2 vectors reading
$$z_i = (x_i, y_i, w_i, h_i, p_i) = (\vec{b}_i, p_i), \qquad i = 1, \dots, S^2$$
where x_i, y_i are the Cartesian coordinates of the BB at the i-th tile, whose width and height are, respectively, w_i and h_i. p_i denotes the probability that z_i is actually a BB.

We train the network to minimize the **loss function**
$L = \sum_{i=1}^{S^2} \| \vec{b}_i - \vec{b}_i^t \|_{L^2} + \mathcal{H}(p_i | p_i^t)$, where the "$t$" denotes the ground truth data and $\mathcal{H}$ is the cross-entropy.

Training

To train the hundred of thousands of parameters in the network one needs millions of hand annotated sample depth maps. To bypass this issue, **we generate realistic artificial depth maps** exploiting the overhead depth map geometry. We proceed as follows:

1. we extract sample **"depth patches"** containing pedestrians of various sizes and shapes as well as objects and depth artifacts from real-life data (cf. Fig.4L);



Sample	Experiment	Pedestrians	Precision CL	Precision CNN	Recall CL	Recall CNN
1	S=5	9	1	1	0.77	1
2		18	1	0.94	0.72	0.94
3	S=10	15	0.75	0.82	0.60	0.93

Figure 4: (L) Sample of "pedestrian patches" adopted. (R) Localization algorithms in action in our two experiments. Comparison of synthetic data (A), clustering-based algorithm (B), and our CNN (C). Precision and recall data for the three samples are in the table. A higher recall for the CNN approach can be observed as side by side subjects and static objects are typically correctly recognized.

2. we **augment** the set of depth patches introducing rotations, flipping, depth translations and different noisy perturbations;
3. we combine augmented patches producing **millions of random training images**. **This is possible because of simple geometric combination rules of depth maps [2,7].**

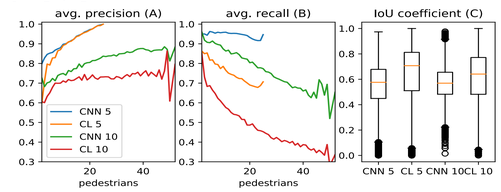


Figure 5: Comparison between DCNN and CL localization approaches in case $S = 5$ and $S = 10$. We include average precision (A) and average recall (B), both conditioned to the number of pedestrians observed, and intersection over union coefficient (C).

Experiments

We compare our trained DCNN with a clustering method [1,2] making two experiments over an area of $2.9m \times 2.2m = 6.4m^2$. Respectively, $S = 5$ (avg. $\sim 1.9 ped/m^2$) and $S = 10$ (avg. $\sim 3.1 ped/m^2$, min dist. $\sim 0.28m$) hold. Results for visual inspection are reported in Fig. 4R(C). In Fig. 5 we plot agglomerated performance data in terms of decision *precision* ($\frac{TP}{TP+FP}$), decision *recall* ($\frac{TP}{TP+TN}$) and BB *intersection over union (IoU)*. (TP, FP, TN stand for number of true positive, false positive and true negative decisions).

Conclusions

1. **We presented a deep convolutional neural network approach for pedestrian localization in overhead depth maps.**
2. **The network shows significantly higher recall performance than clustering methods, that are typically employed for the task.** In fact, it learns the distinctive shape of individuals, therefore it can disentangle neighboring subjects and diminish false positive outputs.
3. **To bypass the difficulties imposed by the need for a large number of annotated training examples, we developed a procedure producing synthetic training data as a mean for efficient delivery of ``weak'' human supervision.**

References:

[1] A. Corbetta et al. *Fluctuations around mean walking behaviours in diluted pedestrian flows* Phys.Rev.E, 032316 2017
[2] A. Corbetta et al. *Continuous measurements of real-life bidirectional pedestrian flows on a wide walkway*, PED 2016
[3] S. Seer et al. *Kinect and human kinetics: A new approach for studying pedestrian behavior*, Transp.Res.C. 2014
[4] D. Brscic et al. *Person tracking in large public spaces using 3-D range sensors*, IEEE.Trans.Hum.Mach.Syst 2013
[5] O. Russakovsky et al. *Imagenet large scale visual recognition challenge*, Int.J.Comput.Vis. 115.3 2015
[6] J. Redmon et al. *You only look once: Unified, real-time object detection*, IEEE ICVPR 2016.

More details: [7] A. Corbetta, V. Menkovski, F. Toschi, *Weakly supervised training of deep convolutional neural networks for overhead pedestrian localization in depth fields,* 2017 14th IEEE International Conference on Advanced Video and Signal based Surveillance (AVSSW 17), https://arxiv.org/abs/1706.02850