in: IEEE World Congress on Computational Intelligence (WCCI), Vancouver, Canada, pp. 267-276, IEEE 2016

Cooperative Multi-Scale Convolutional Neural Networks for Person Detection

Markus Eisenbach, Daniel Seichter, Tim Wengefeld, and Horst-Michael Gross
Ilmenau University of Technology, Neuroinformatics and Cognitive Robotics Lab
98684 Ilmenau, Germany
markus.eisenbach@tu-ilmenau.de

***Abstract*—Robust person detection is required by many computer vision applications. We present a deep learning approach, that combines three Convolutional Neural Networks to detect people at different scales, which is the first time that a multi-resolution model is combined with deep learning techniques in the pedestrian detection domain. The networks learn features from raw pixel information, which is also rare for pedestrian detection. Due to the use of multiple Convolutional Neural Networks at different scales, the learned features are specific for far, medium, and near scales respectively, and thus, the overall performance is improved. Furthermore, we show, that neural approaches can also be applied successfully for the remaining processing steps of classification and non-maximum suppression. The evaluation on the most popular Caltech pedestrian detection benchmark shows that the proposed method can compete with state of the art methods without using Caltech training data and without fine tuning. Therefore, it is shown that our method generalizes well on domains it is not trained on.**

I. Introduction

Detecting persons in images at high accuracy is indispensable for a wide range of applications, including pedestrian detection for car assistance systems [1], person detection for automatic surveillance video analysis [2], and potential user recognition for human robot interaction [3]. Therefore, a person detector should be generic and applicable to several domains. In the last years, a lot of approaches have been presented using different advanced hand-crafted features. Recently, deep learning approaches supplant these methods by learning superior features data-driven. Following this trend, we present an approach that is trained on a large dataset including samples from several domains. Therefore, it can cope with the state of the art without fine tuning on a specific benchmark. We implemented all processing steps, i.e., feature extraction, classification, and non-maximum suppression by means of Convolutional Neuronal Networks. Therefore, we can claim that it is sufficient to use neural approaches only.

II. Related Work

Visual Pedestrian detection is a wide field of research with a large range of approaches, developed over recent years. According to [4], the approaches can be divided into three families of solutions:

- **Full body detection methods applying feature selection**: Decision Forest (DF) approaches [5], [6] typically generate a very large pool of features during the training phase. Therefore, they systematically select sums over rectangular regions [7] or haar like features [8] over different channels of the input image. AdaBoost is then used to train a classifier while it is simultaneously finding the best features for the classification problem.
- **Body part detectors**: The philosophy of deformable parts models (DPM) [9], [10] is to detect smaller patches, e.g. a leg or head of a pedestrian and to combine them in relation to their spatial dependencies. These approaches usually use gradient features for part description and a star model to the root of the person to model spatial relations.
- **Deep Learning approaches**: The family of deep learning (DL) approaches make use of deep neural network architectures to learn features rather than using designed ones either from raw pixels [11] or edge and color channels [12], [13].

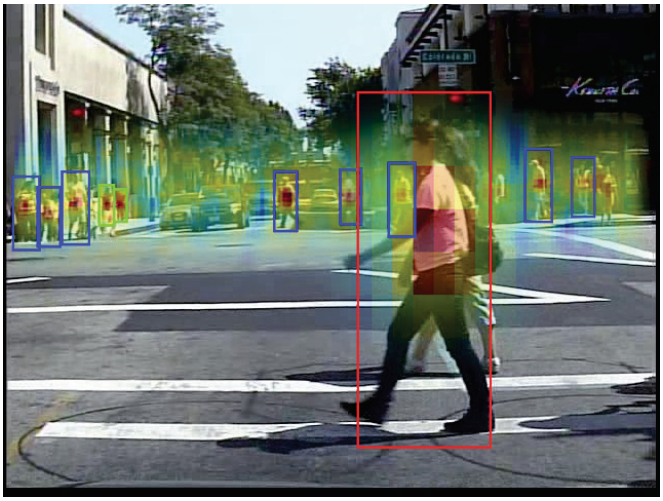


Fig. 1: Person detections by Convolutional Neural Networks at multiple scales shown as red, blue, and green boxes with overlaid network output.

A well known generic option for object detectors taken up by us is the usage of multi-resolution models. Since features of the same object class differ with the distance to the sensor, it is favorable to train different classifiers for different image resolutions. This improvement was successfully adapted to DF [5], [14] and DPM [15], [9] but is not yet exploited by deep learning approaches.

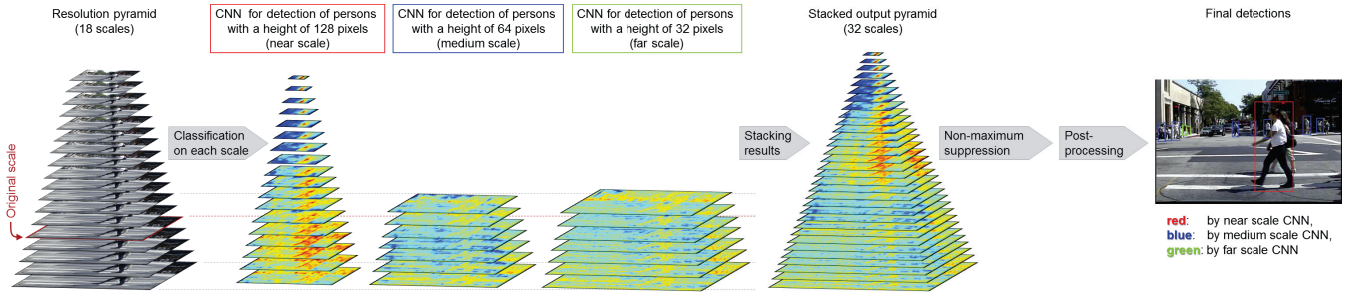


Fig. 2: Processing chain in the application phase. To create the resolution pyramid, the input image is scaled by a factor of 0.9057237 to exactly half the image size after seven scales. Thus, each of the three CNNs has to process seven scales. The near scale CNN (red box, second from left) additionally processes smaller scales to detect larger persons. The classification results are stacked and non-maximum suppression[1] is applied to find the best fitting positions and scales for all persons in the scene. Finally, detections are post-processed to remove some false positives.

Our contribution is to combine a multi-resolution model with deep learning techniques using raw pixel information as input. Therefore, we are able to learn problem-specific features for every resolution without any feature design decisions.

Since the data used for training and evaluation have a huge impact on the quality of the detector, the large but challenging Caltech dataset [1] has evolved as the standard benchmark for pedestrian detection. In [4] it is shown that nearly every top performing approach on the Caltech benchmark uses Caltech training samples while those who do not perform significantly worse. Hence, often approaches tend to be fine tuned on this dataset but do not generalize well. We show that it is possible to achieve top performance on the Caltech dataset without using Caltech training data and thereby generalize better on other domains.

III. Multi-Scale Person Detection by CNNs

To detect persons at different scales using a sliding window approach, there are three possibilities:

- Using a resolution pyramid and apply a single detector to each scale.
- Using several detectors for finding persons with different sizes and apply each of them on the non-scaled input image.
- Using a hybrid approach with few detectors for finding persons of different sizes and apply them to parts of the resolution pyramid.

We decided in favor of the hybrid approach, known as multi-resolution model, that uses a resolution pyramid in combination with detectors at different scales. Thus, it is fast in the application phase, and the number of neural networks to be trained is manageable.

A. System Overview

Our approach uses three Convolutional Neural Networks to handle different scales. We trained these networks on cropped images showing persons (positive class), other objects, and typical false detections (e.g. sub-images containing only parts of a person or persons that fill only parts of the image section = negative class). After network training, fully connected layers were converted to convolutional layers to be able to process images of any size without the need to shift a sliding window to several locations. Thus, we achieve a great speedup in the application phase.

Fig. 2 shows the processing chain of the application phase. The requirements and design decisions for this network architecture are described subsequently. Each of the Convolutional Neural Networks works on multiple scales of the resolution pyramid. For each image scale, an output map is calculated. High neural activations suggest that persons are present in that region of the image (see Fig.1). When these classifications have been done, the full output pyramid can be constructed. Then, a 3D non-maximum suppression (NMS) is applied to find persons in the scene and at the best fitting scale. For NMS, we implemented an approximation of the mean-shift algorithm as a single 3D pooling layer. In a last step, we post-process all detections. Therefore, we filter out detections that do not fit the ground plane and those ones, that do appear only once in consecutive frames.

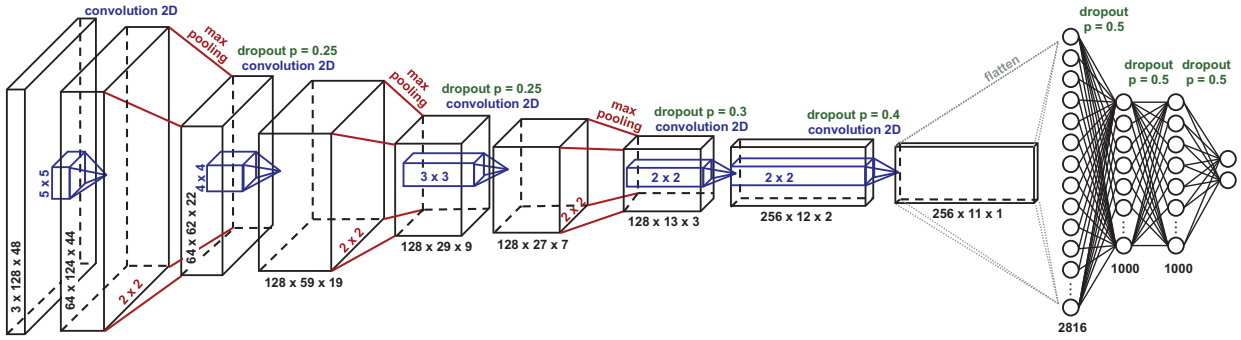
For implementation, we used the Theano framework [17], [18] in combination with Keras[2]. The hardware, used for training, is a PC with a Core i7 CPU, 16 GB RAM, and a single NVIDIA Titan X GPU.

B. Related Work: Multi-Scale Neural Networks

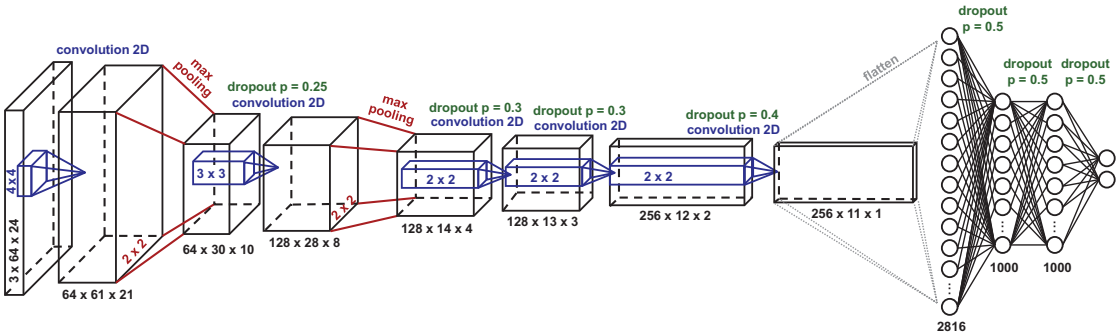
In recent work, there are some neural approaches that address the task of object recognition in multiple scales, which in the following will be described briefly. One part of approaches use special pooling layers to represent feature maps at different scales, e.g. [19] apply pyramidal pooling layers. In [20] the outputs of multiple layers are connected with fully connected layers to represent features of different scales. Another part of approaches apply multiple sub-networks with identical topology to differently scaled input images. In [21] these sub-networks are fused by connecting them at the fully connected layers. In [22] the fusion is done later by connection at the softmax layer. The approach most similar to ours is [23]:

[1]For the basic idea and an overview on non-maximum suppression, we refer to [16].

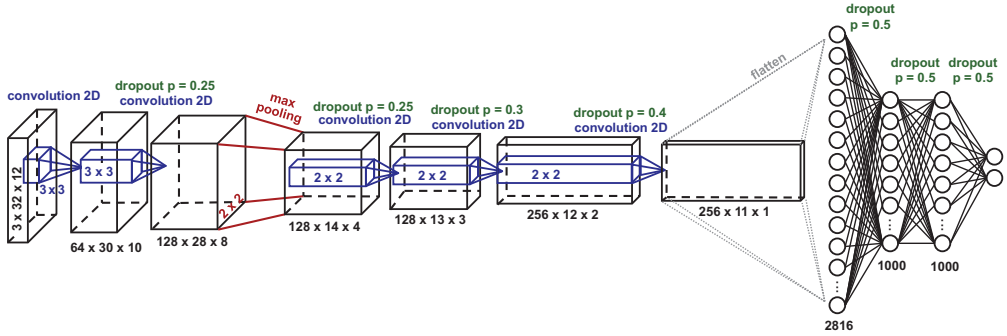
[2]Keras: Deep Learning library for Theano and TensorFlow http://keras.io/



(a) CNN for detecting persons with a height of 128 pixels. It processes near scales and, thus, detects persons of height 80 pixels and larger.



(b) CNN for detecting persons with a height of 64 pixels. It processes medium scales and, thus, detects persons of height 40–79 pixels.



(c) CNN for detecting persons with a height of 32 pixels. It processes far scales and, thus, detects persons of height 20–39 pixels.

Fig. 3: Convolutional Neural Network topologies for detecting persons at different scales.

Multiple CNNs that are applied to different scales produce multiple output maps that encode positions where objects of interest are likely to be. Then, all maps are scaled to the same size to make them comparable. Finally, a complex refinement step is applied to find out where the objects are and at which scale. In contrast, we use a top down design to construct the networks for the differently scaled inputs such that their output maps are easily comparable. Therefore a simple 3D max pooling is sufficient to locate the persons in the input image. Details will be described in the following.

C. Network Architecture and Processing Chain

The objective of the network design is to get network outputs for different scales that are easily comparable such that a 3D NMS is sufficient to locate persons in the input image. The outputs would be comparable if a single network would produce all output maps. But a single network would not be flexible enough to detect persons at all scales adequately. Therefore, we decided in favor of a collection with three CNNs to detect persons at near, medium, and far scale, each producing parts of the output pyramid. The three CNNs process input image patches of different size, so each one specializes on detecting persons of a specific size. However, in order to get output maps that appear to be produced by the same network a proper network architecture for each of the networks must be chosen.

The input patch sizes for the three networks are chosen as follows: The far scale CNN processes image patches of height 32 pixels. It is applied to several scales and, thus, can detect persons of height 20–39 pixels. The medium scale CNN detects persons twice that size. Thus, its inputs are image patches with a height of 64 pixels. Therefore, it is responsible

for detecting persons of height 40–79 pixels. The near scale CNN again detects persons twice the size of the medium scale CNN. Its input patches are 128 pixels high. Thus, it detects persons who are at least 80 pixels high. To apply the scheme presented above, the resolution pyramid is scaled by a factor of 0.9057237 to exactly half the image size after seven scale layers, and each of the three CNNs processes seven scales as shown in Fig. 2. The near scale CNN additionally processes smaller scales to detect larger persons.

If this scheme would be applied using identical architectures for all CNNs, the size of the output maps in the application phase would not match the required size. The medium scale CNN would produce the same output size for an input image as the near scale CNN. This would be caused by identical filter and pooling sizes that are applied on the whole image in the application phase. The expected output size for the medium scale CNN would be twice the size of the near scale CNN's output. In our case, this can be achieved by using one pooling layer less[3]. Therefore, the near scale CNN (input height 128 pixels) has three pooling layers, the medium scale CNN (input height 64 pixels) two pooling layers, and the far scale CNN (input height 32 pixels) has only one pooling layer. The final network topologies are shown in Fig. 3. All neurons use scalar product activation and a ReLU output function. The input coding is described in the next section. The output is a softmax layer with two neurons representing non-persons and persons.

D. Input Coding

As input we take the pixels of the image to be classified in RGB color space. We normalize the input such that a black pixel is represented by three neurons (RGB) with activation -1, white ones by three neurons with activation 1, and medium gray pixels by three neurons with activation 0. Thus, the latter have no influence to subsequent layers. We decided in favor of this input coding to accomplish that normalization has zero mean. By gray world assumption the expected mean color is gray.

Additionally, we use zero padding for training samples that do not fill the complete patch (e.g. when persons are near the camera and legs are outside the image). Thus, the added regions do not have any influence on subsequent layers.

E. Network Training

For training the Convolutional Neural Networks, we use cropped images of size 128×48 showing both persons and non-persons. Therefore, we first composed a large database incorporating multiple benchmark datasets (see below). Training images were identical for each of the three Convolutional Neural Networks but are downscaled for two of them (to 64×24 and 32×12 respectively).

1) Training Data: To get a large, versatile, general purpose training dataset, we combined 22 datasets from pedestrian detection and person re-identification domain. Tab. I lists all datasets used here: Positive samples (persons) are drawn from all datasets utilized. The number of images taken from each of the datasets is listed in Tab. I. We took care to avoid sampling too many images of identical persons. Therefore, if more than 20 images per subject were available, we removed samples by k-medoids[4] clustering based on color histograms of the image patches to ensure different lighting and other environmental conditions.

Negative samples were taken from the INRIA [31], NICTA dataset [35] and from publicly available images without persons showing landscapes and urban scenes. To simulate typical false detections, we extracted mis-aligned patches of the SAIVT-SoftBio dataset [40] showing at most half of a person. Therefore, we used the ground truth and shifted the bounding boxes by half in four directions (up, down, left, right). Additionally, we scaled the boxes in half and doubled their size to incorporate incorrect scales. This ensures that the networks learn to separate persons such that all persons standing nearby can be detected when non-maximum suppression (NMS) is applied.

To also collect real false detections that are made by state of the art detectors (we used [10], [14] and [48]), we recorded data with a mobile robot. The autonomous robot drove through a clinic [49], [50] and a faculty building [51] when no people were present, so every detection represents a false detection. Then, the robot drove through the corridor of the clinic when lots of persons were present. A laser-based detection approach [52] and the map of the scene were used to check the detections for plausibility [3]. We additionally checked them manually. These negative samples shall help the networks to avoid typical errors.

Summarized, we collected a relatively large training dataset containing 100,107 positive and 628,636 negative samples. This is crucial to learn proper features and classifiers by deep learning.

Note that only 5.6% of the positive samples are drawn from domains that are similar to our testing benchmark dataset Caltech. These are the 5,644 samples from PedCut, PPSS, and PRID450S dataset (see Tab. I). Furthermore, none of the negative samples are drawn from domains similar to Caltech. Thus, our training dataset differs significantly from the Caltech dataset. Hence, if the trained network works well on Caltech, we can claim that it generalizes well on domains it is not trained at.

2) Parameter Optimization and Regularization: As training algorithm, we use stochastic gradient descend (SGD) with mini-batches and momentum. To avoid overfitting, we use dropout for regularization [53], [54]. It is applied to all layers, except the input layer, using relatively large dropout rates. This was found to be necessary, since the error landscape seems to

[3]Theoretical background: The number of pooling layers P affects the stride S in the original image ($S = 2^P$) and thus, the size of the network output. Note that this formula is only valid if the convolution filters are applied with a stride of 1 and the pooling regions have a size of 2×2 with stride 2, which is fulfilled in our CNNs.

[4]K-medoids clustering is similar to k-means clustering but with input samples as centroids, see [46].

Dataset	Short description	Location	Camera view	#images	Source
3DPeS	3D People Surveillance Dataset: People viewed from different angles, good resolution, no occlusions, different lighting condition	Campus, outdoor	Surveillance	866	[24]
CAVIAR4REID	Clips from shopping center in Portugal of CAVIAR project: People captured with low resolution camera, some occlusions, frontal, side, and back views	Mall, indoor	Surveillance	1220	[25][5]
CUHK01	Students walking on campus, observed from different views, few occlusions	Campus, outdoor	Surveillance	4403	[26]
ETHZ	Pedestrians in pedestrian zone observed over longer timespan, from low to high resolutions, some occlusions, mainly frontal and back views	Pedestrian zone, outdoor	At ground level	2784	[27], [28]
GRID	Surveillance camera footage of a subway station, bad image quality, very noisy, very dark	Subway, indoor	Surveillance	1275	[29]
iLIDS	Surveillance camera footage of an airport terminal, very different perspectives, low resolution, noisy images	Airport, indoor	Surveillance	476	[30]
INRIA	Photo collection, holiday and sports activities, mainly urban scenes	(varies), outdoor	At ground level	1704	[31]
Mall crowd counting	Lots of far away persons in low resolution, many occlusions	Mall, indoor	Surveillance	1356	[32]
Market-1501	Good image quality, persons in several different poses	Urban, outdoor	Surveillance	25259	[33]
MIT	Pedestrians recorded in good resolution, only frontal and back views, no occlusions	Urban, outdoor	At ground level	888	[34]
NICTA	Large collection of pedestrians in urban scenes	Urban, outdoor	At ground level	44223	[35]
PedCut	Daimler pedestrian segmentation dataset, pedestrians walk near streets	Car traffic, outdoor	At ground level	785	[36]
PPSS	Pedestrian Parsing in Surveillance Scenes Dataset, pedestrians walking near streets	Car traffic, outdoor	Surveillance	3961	[37]
PRID450S	Person Re-ID 450S, Pedestrians crossing street, side views, bad color calibration of images	Cross-walk, outdoor	Surveillance	898	[38]
RAiD	Re-identification Across indoor-outdoor Dataset, different lighting conditions, no occlusions, good image quality	Campus, indoor, outdoor	Surveillance	865	[39]
ROREAS	ROREAS-robot driving through a rehab clinic, patients using walking aids	Clinic, indoor	At ground level	2501	[3]
SAIVT-SoftBio	Surveillance camera footage, good image quality and resolution	Airport, indoor	Surveillance	3040	[40]
SARC3D	Persons in four poses (front, left, right, back view), no occlusions, good image quality and illumination	Campus, outdoor	Surveillance	200	[41]
Town Centre	Pedestrians in urban scenes, no occlusions	Pedestrian zone, outdoor	Surveillance	878	[42]
V-47	People in office, different views (front, side, back), many partial occlusions	Office, indoor	At ground level	662	[43]
VIPeR	Pedestrians, no occlusions, varying lighting conditions	Urban, outdoor	At ground level	1264	[44]
WARD	People walking on campus, good image quality, no occlusions	Campus, outdoor	Surveillance	599	[45]

TABLE I: Collection of datasets used to extract positive samples (images of persons).

be very cluttered with lots of suboptimal local minima. The dropout rates increase from input to output layer with highest dropout rates of 0.5 for the fully connected layers. For each layer's dropout rate, we refer to Fig. 3.

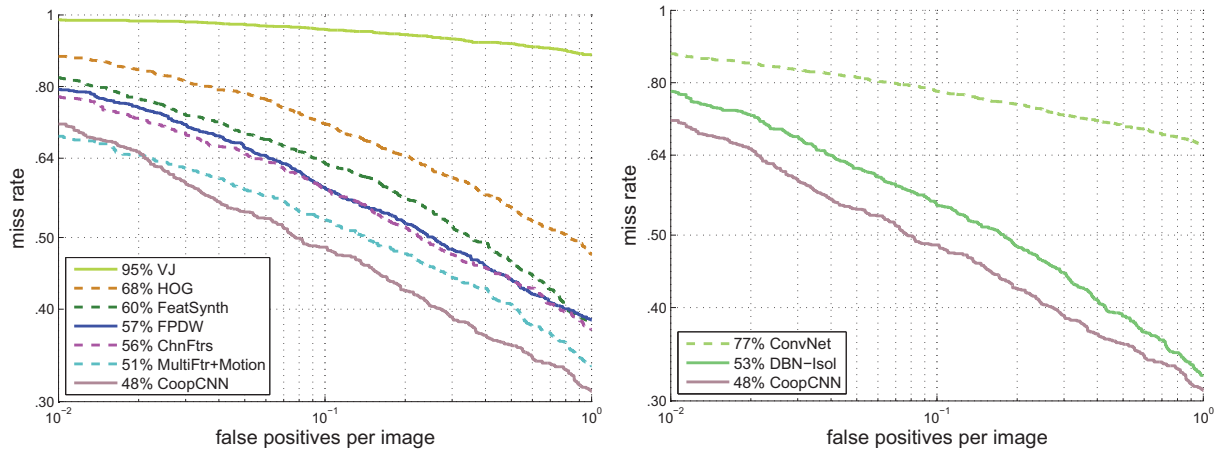
The applied SGD training is controlled by three parameters: learning rate, mini-batch size, and momentum. These parameters were chosen as follows:

- Learning rate: starting at 0.01, linear decreasing to 0.0001 within 2,000 epochs
- Mini-batch size: 256
- Momentum: relatively high, starting at 0.9, linear increasing to 0.999 within 2,000 epochs

[5]Extracted from the CAVIAR-Dataset: http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/

F. Reshaping CNNs for Application Phase

After network training, we reshaped the CNNs such that they are applicable to different image sizes. Therefore, we first removed all dropout layers and used the linear output of the last layer instead of the softmax output. Then, the weights of the fully connected layers were reshaped to three dimensions such that they can be used as filter kernels. This avoids the need of shifting the sliding window to several positions and, thus, speeds up the processing in the application phase. For the first fully connected layer this can be done easily by ignoring



(a) Comparison to state of the art

(b) Comparison to deep learning approaches

Fig. 4: Evaluation results under reasonable conditions (unoccluded and partially occluded persons taller than 50 pixels) on the Caltech Pedestrian Dataset. Our approach, referred to as CoopCNN, is compared to the best methods in [47] and the standard methods Viola & Jones (VJ) and HOG (a) and to other deep learning approaches (b). For a visualization of additional state of the art methods see Fig. 8.

the step shown as 'flatten' in Fig. 3. This and succeeding layers produce a number of feature maps that equal their number of neurons. Thus, reshaping all other layers is easy, too. Finally, the softmax function was applied to the two linear output maps.

G. Non-Maximum Suppression

When CNNs are reshaped, they can be used directly to calculate output maps from differently scaled full images instead of patches. Then, the outputs of the three CNNs can be stacked to create the output pyramid. Finally, non-maximum suppression has to be applied to find the best positions and scales for all persons in the scene. Therefore, we implemented a single 3D max-pooling layer as approximation of the mean-shift algorithm.

The 3D pooling is applied to the scale of interest and the respective five scales above and beneath, each filtered with a 2×2 2D max-pooling with stride 1×1 and one pixel zero padding (to avoid aliasing affects) and then rescaled to the size of the output of the scale of interest. For max-pooling, we use a pooling region of depth 11 (scales), width 3 (x-direction) and height 7 (y-direction) with zero-padding of 0, 1, and 3 in scale, x-, and y-direction respectively. The values in x- and y-direction are chosen because of the fact that persons can appear next to each other (x-direction) but unlikely above each other (y-direction). If the maximum is found within the scale of interest, a person is detected in that scale at the position of the maximum. The position in the output map can then be converted to the position in the original image by computing the output neurons' receptive field in the input layer, which matches the detected person's bounding box. For our network architecture, this is easily done considering the stride as given by the number of max-pooling layers in the responsible CNN at this scale (see Sect. III-C).

H. Post-processing

As post-processing steps, we filter out detections that do not fit the ground plane and those ones that do appear only once in consecutive frames.

1) Ground Plane Assumption: Assuming that all people stand on the ground, the height is a linear function of the base point of the hypothesis' bounding box. Therefore, for each detection we calculate the expected height $\hat{h}$ from the base point (only y-coordinate) and compare if the detected height h_{det} is within the interval $\frac{\hat{h}}{1.6} \ldots \hat{h} \cdot 1.6$. If not, we remove this detection.

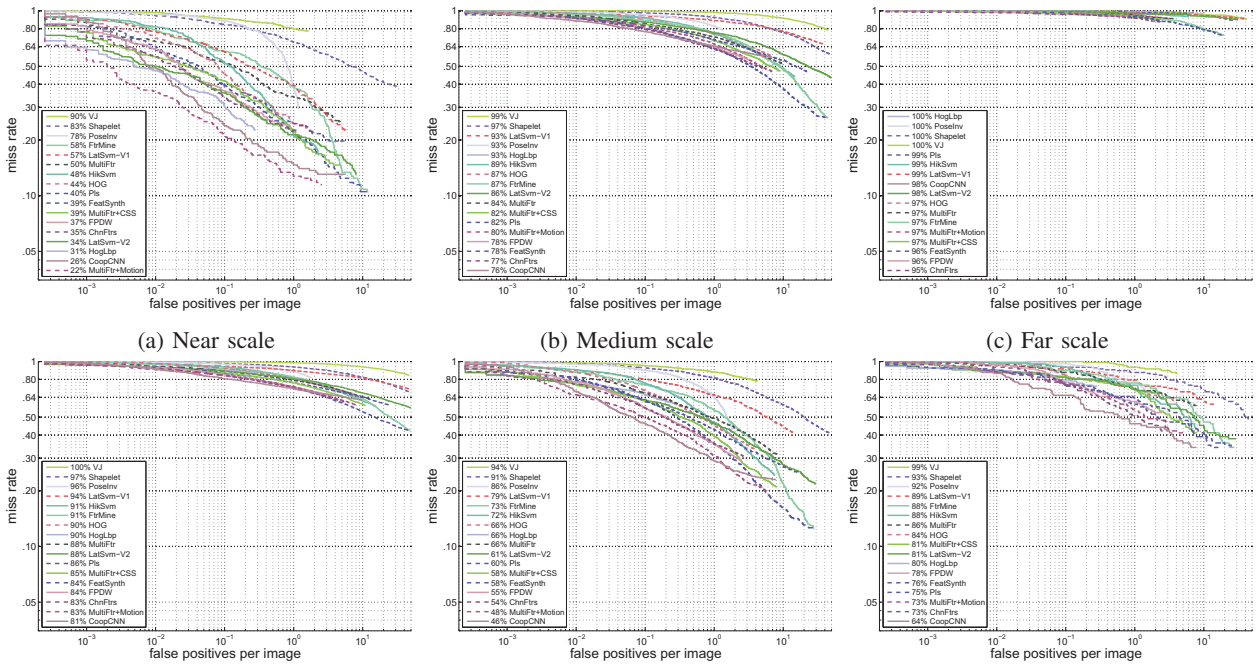
2) Temporal Filtering: Real applications as well as most datasets (e.g. Caltech, see Sect. IV) provide videos instead of incoherent images. To incorporate the temporal context, we filter out detections that are found in the considered frame only but not in the three frames before and after. Correspondences are found by comparing the ratio of the overlap to the union of the bounding boxes and check if it is above a threshold of 0.5 (same as used in Caltech evaluation protocol).

IV. Experiments

For evaluation of the proposed method, we choose the most popular Caltech Pedestrian Detection Benchmark [1]. Its training data are not included in our training data to show that we have designed a Convolutional Neural Network collection that is able to cope with the state of the art on any application without fine tuning.

The Caltech dataset consists of video streams recorded from a car driving through urban streets. The objective is to detect all pedestrians who are at least 20 pixels high. Due to bad image quality, low resolution, lots of occlusions, and diverse backgrounds, it is one of the most challenging pedestrian detection dataset and, thus, the most popular benchmark.

The current evaluation protocol [47] will be explained briefly in the following. Evaluation is performed for every



(a) Near scale (b) Medium scale (c) Far scale

(d) Overall (e) No occlusion (f) Partial occlusion

Fig. 5: Detailed evaluation results under different conditions on the Caltech Pedestrian Dataset. Our approach is referred to as CoopCNN. (a) Performance on unoccluded persons taller than 80 pixels, (b) on unoccluded persons of size 30–80 pixels, and (c) on unoccluded persons of size 20–30 pixels. (d) Overall performance for all annotated pedestrians taller than 20 pixels that are not fully occluded. (e) Part of the reasonable evaluation (persons taller than 50 pixels) where persons are unoccluded and (f) where they are partially occluded.

30th video frame (or one image per second respectively). Assignments of detections with the ground truth are found by calculating the ratio of the intersection of two bounding boxes to the union of the boxes. If the ratio is above a threshold of 0.5, the detection is considered to be a match. If multiple boxes match the ground truth, only the one with the highest confidence is taken. Assigned detections are true positives, not assigned detections are false positives, and not assigned ground truth boxes are false negatives.

The relevant scenario, referred to as 'reasonable', is defined as follows [47]: Persons' heights have to be at least 50 pixels. Both unoccluded and partially occluded persons are considered. Heavily and fully occluded persons are excluded from evaluation.

We used the publicly available MATLAB code for evaluation. It constructs the detection error tradeoff (DET) curve (which equals the ROC curve with flipped ordinate) with the abscissa as false positives per image (fppi) and the ordinate as miss rate (1 - true positive rate) and shows this on a double logarithmic plot. The most relevant working point is defined as position where fppi = 10^{-1}, which means, only one false detection every ten images is allowed. The miss rate for each of the detectors at this working point is shown in the legend. For further evaluation details, we refer to [47].

Fig. 4a shows the detection capability of our approach, referred to as CoopCNN, in comparison to the best state of the art approaches evaluated in [47]. Fig. 4b compares the proposed method with other deep learning approaches that also excluded Caltech training data from their training. It can be seen that our approach performs better than the state of the art and significantly better than the deep learning approaches ConvNet [11] and DBN-Isol [13] in the relevant range of fppi = $10^{-1} \ldots 10^{0}$. Note that the computer vision approaches evaluated in [47] use hand-crafted features but train a SVM for classification in a similar way we train our networks. The other deep learning approaches also use a fully supervised training scheme. So, the kind of training is comparable. However, the training datasets differ.

Fig. 5 shows a detailed analysis of the performance. Our method is compared to the state of the art in several categories:

- Near scale (Fig. 5a) includes only large persons with a height of 80 pixels and above that are not occluded. The proposed method is the second best. The only better approach benefits from a multi-frame tracking approach.
- Medium scale (Fig. 5b) includes only medium sized persons with a height of 30–80 pixels that are not occluded. Our method is best at the relevant working point and within the best approaches for more false positive per image.
- Far scale (Fig. 5c) includes only small persons with a height of 20–30 pixels that are not occluded. As other approaches, our method clearly fails.
- Overall (Fig. 5d) includes all persons with a height of at least 20 pixels that are not fully occluded. Our method is

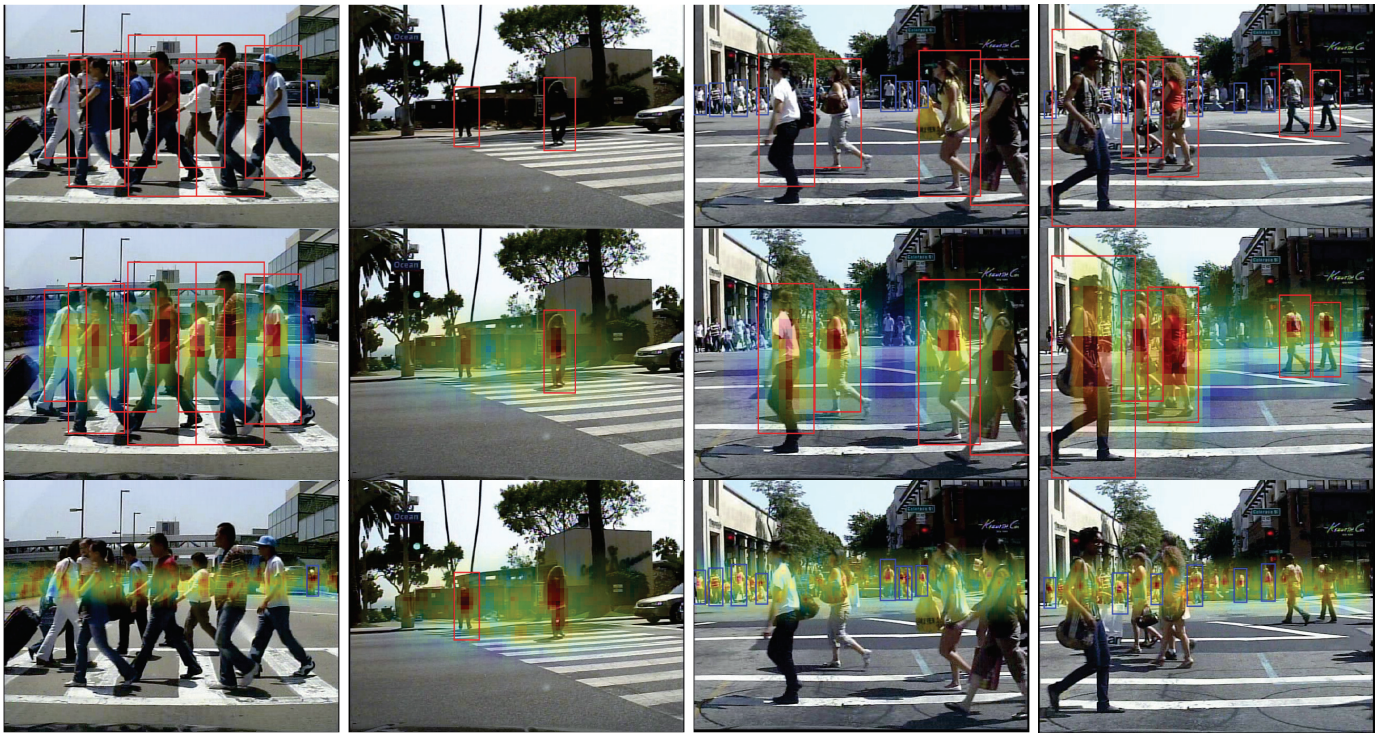


Fig. 6: Examples of Caltech dataset. Top row: Detections of the proposed method are shown in red and blue to highlight which CNN was responsible for the detection (near or medium scale CNN). Center row: Linear network output for larger persons overlaid. Red regions represent a high output for person class while yellow/green/blue regions represent a low output in this scale. Bottom row: Linear network output for smaller persons overlaid. Note that visualizations of linear outputs for different scales and images are scaled separately such that the highest activation is show dark red and the lowest activation as blue. After application of the softmax function and non-maximum suppression, only the dark red activations remain as person detections.

the best in the relevant range of fppi = $10^{-1} \ldots 10^{0}$.

- No occlusion (Fig. 5e) shows the part of the reasonable evaluation where persons are not occluded. Again, our method is the best in the relevant range of fppi = $10^{-1} \ldots 10^{0}$.
- Partial occlusion (Fig. 5f) shows the part of the reasonable evaluation where persons are partially occluded. In this category our method clearly outperforms the state of the art. The benefits of the proposed negative sampling for the training dataset become apparent.

To evaluate the visual results, in Fig. 6 detections are shown for four examples of the Caltech dataset. Additionally, the linear network output is shown for different scales. The three cooperative Convolutional Neural Networks perform very well in detecting nearly all persons at different scales in the scene while false positive detections are not present.

Fig. 7 shows the 64 filters of the first convolutional layer learned by the near scale CNN. The network learned typical color filters and textural filters in different color channels. It is remarkable, that it additionally learned some specialized filters. Blue boxes highlight filters that search for skin color and edges of elliptical structure simultaneously. These can be used to find the contour of a face. Red boxes highlight filters that search for defined lines which may be used to find facial structure. Note that these filters are very similar to haar-like features learned by the famous Viola & Jones face detector. The face position was not labeled in the data but were learned as underlying structure that strongly indicates the presence of a person. Since none of the filters is large enough to detect a face at a whole, the network learned to combine multiple part features to master this hidden task of finding faces.

As a consequence, we observed some typical false detections that strongly respond to round (face-like) objects with

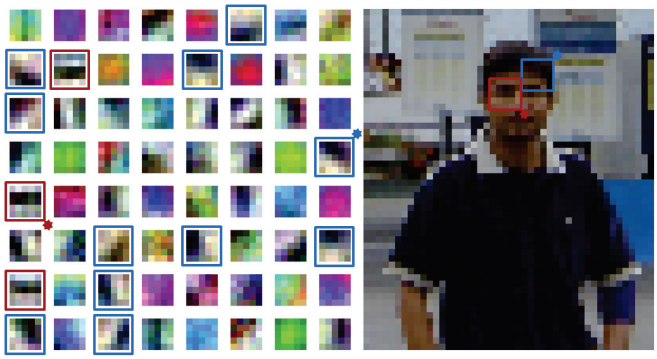
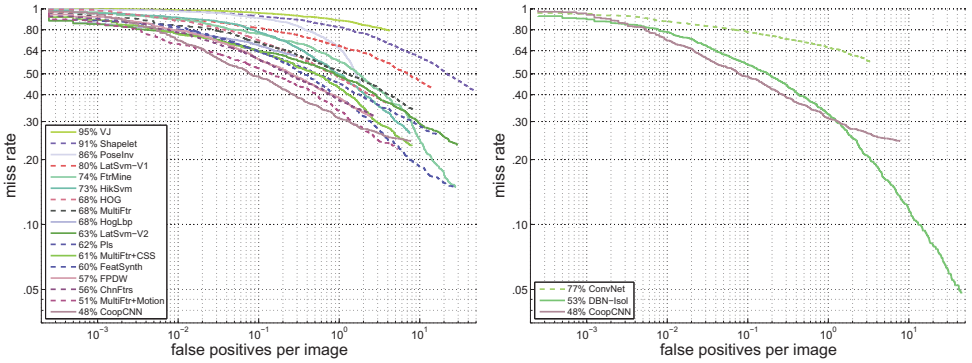


Fig. 7: Filters of the first convolutional layer learned by the near scale CNN. Left: The 64 learned filters. Right: Sample image from the INRIA dataset to set the filter size in relation to the patch size. Red and blue boxes highlight specialized filters. For two filters good fits in the sample are show.



(a) Comparison to state of the art

(b) Comparison to deep learning approaches

Fig. 8: Evaluation results under reasonable conditions (unoccluded and partially occluded persons taller than 50 pixels) on the Caltech Pedestrian Dataset on a wider fppi range than in Fig. 4. Additionally, methods that are not listed there are shown. Our approach is referred to as CoopCNN.

red shades (skin-colors) in the upper third of the patch. In the Caltech road traffic scenario, these are back lights of cars and red traffic lights. These objects caused about one third of all false positive detections. Note that green traffic lights do not lead to false detections. Other typical false detections result from very structured regions with defined edges and simultaneous color changes. These kind of inputs are under-represented in the training data. Specific re-training may eliminate these weaknesses of our detector. This will be the focus of our future work.

The computation in the application phase on a NVIDIA Titan X GPU took 2.061 seconds on average per image of size 640×480 pixels when persons of at least 20 pixels height should be detected. In future work, we plan to reduce this time dramatically by cutting the resolution pyramid to only relevant regions based on the ground plane. Additionally, if far scale persons are not relevant for the application, the computation time can be significantly reduced to 0.594 seconds, and if only near scale persons matter (height ≥ 80 pixels), the computation time can be reduced to 0.231 seconds. In comparison to the state of the art methods evaluated in [1], our method is the second fastest, although this comparison is not fair, since our method runs on a high performance GPU and all others on a CPU. But the method with the next best miss rate, MultiFtr+Motion [55], is more than 100 times slower then the proposed method and thus, would not catch up by using a GPU.

V. Conclusion

In this paper, we have presented a deep learning approach that combines three Convolutional Neural Networks to detect people at different scales. This is the first deep learning implementation of a multi-resolution model in the pedestrian detection domain. The networks learn features from raw pixel information. Due to the use of multiple Convolutional Neural Networks at different scales, the learned features are specific for the respective resolutions the particular CNNs are applied to, which improves the performance significantly. Furthermore, we successfully applied neural approaches for the remaining processing steps of classification and non-maximum suppression. The evaluation on the most popular Caltech pedestrian detection benchmark shows that the proposed method beats state of the art methods although it does not use Caltech data for network training. Other deep learning approaches that exclude Caltech training data as well are outperformed significantly. Detailed experiments show that the proposed method performs best or second best on all sub-evaluations under varying conditions. It is far the best method when persons are partially occluded. The fact that we can compete with the state of the art without training on scenario-specific data shows that the proposed deep learning approach generalizes very well to unseen domains.

References

[1] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: A benchmark," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 304–311.

[2] A. Kolarow, K. Schenk, M. Eisenbach, M. Dose, M. Brauckmann, K. Debes, and H.-M. Gross, "APFel: The intelligent video analysis and surveillance system for assisting human operators," in *IEEE Int. Conf. on Advanced Video and Signal-Based Surveillance (AVSS)*. IEEE, 2013, pp. 195–201.

[3] M. Eisenbach, A. Vorndran, S. Sorge, and H.-M. Gross, "User recognition for guiding and following people with a mobile robot in a clinical environment," in *Int. Conf. on Intelligent Robots and Systems (IROS)*, 2015, pp. 3600–3607.

[4] R. Benenson, M. Omran, J. Hosang, and B. Schiele, "Ten years of pedestrian detection, what have we learned?" in *ECCV Workshop on Computer Vision for Road Scene Understanding and Autonomous Driving (CVRSUAD)*, 2014, pp. 613–627.

[5] R. Benenson, M. Mathias, T. Tuytelaars, and L. Van Gool, "Seeking the strongest rigid detector," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 3666–3673.

[6] P. Dollar, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 8, pp. 1532–1545, 2014.

[7] P. Dollar, Z. Tu, P. Perona, and S. Belongie, "Integral channel features," in *British Machine Vision Conf. (BMVC)*, 2009, pp. 91.1–91.11.

[8] S. Zhang, C. Bauckhage, and A. B. Cremers, "Informed haar-like features improve pedestrian detection," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 947–954.

[9] J. Yan, X. Zhang, Z. Lei, S. Liao, and S. Z. Li, "Robust multi-resolution pedestrian detection in traffic scenes," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 3033–3040.

[10] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part based models," *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 32, no. 9, pp. 1627–1645, 2010.

[11] P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. LeCun, "Pedestrian detection with unsupervised multi-stage feature learning," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 3626–3633.

[12] W. Ouyang and X. Wang, "Joint deep learning for pedestrian detection," in *Int. Conf. on Computer Vision (ICCV)*, 2013, pp. 2056–2063.

[13] W. Ouyang and X. Wang, "A discriminative deep model for pedestrian detection with occlusion handling," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 3258–3265.

[14] P. Dollar, S. Belongie, and P. Perona, "The fastest pedestrian detector in the west," in *British Machine Vision Conf. (BMVC)*, 2010, pp. 68.1–68.11.

[15] D. Park, D. Ramanan, and C. Fowlkes, "Multiresolution models for object detection," in *European Conf. on Computer Vision (ECCV)*, 2010, pp. 241–254.

[16] A. Neubeck and L. Van Gool, "Efficient non-maximum suppression," in *Int. Conf. on Pattern Recognition (ICPR)*, vol. 3. IEEE, 2006, pp. 850–855.

[17] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, and Y. Bengio, "Theano: a CPU and GPU math expression compiler," in *Proceedings of the Python for Scientific Computing Conference (SciPy)*, Jun. 2010.

[18] F. Bastien, P. Lamblin, R. Pascanu, J. Bergstra, I. J. Goodfellow, A. Bergeron, N. Bouchard, and Y. Bengio, "Theano: new features and speed improvements," Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop, 2012.

[19] J. Masci, U. Meier, G. Fricout, and J. Schmidhuber, "Multi-scale pyramidal pooling network for generic steel defect classification," in *Int. Joint Conf. on Neural Networks (IJCNN)*. IEEE, 2013, pp. 1–8.

[20] P. Sermanet and Y. LeCun, "Traffic sign recognition with multi-scale convolutional networks," in *Int. Joint Conf. on Neural Networks (IJCNN)*. IEEE, 2011, pp. 2809–2813.

[21] K. Kamnitsas, C. Ledig, V. F. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, and B. Glocker, "Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation," *arXiv preprint arXiv:1603.05959*, 2016.

[22] H. Dou and X. Wu, "Coarse-to-fine trained multi-scale convolutional neural networks for image classification," in *Int. Joint Conf. on Neural Networks (IJCNN)*. IEEE, 2015, pp. 1–7.

[23] C. Szegedy, A. Toshev, and D. Erhan, "Deep neural networks for object detection," in *Advances in Neural Information Processing Systems*, 2013, pp. 2553–2561.

[24] D. Baltieri, R. Vezzani, and R. Cucchiara, "3dpes: 3d people dataset for surveillance and forensics," in *Int. ACM Workshop on Multimedia access to 3D Human Objects*, 2011, pp. 59–64.

[25] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino, "Custom pictorial structures for re-identification," in *British Machine Vision Conf. (BMVC)*, 2011, pp. 68.1–68.11.

[26] W. Li, R. Zhao, and X. Wang, "Human reidentification with transferred metric learning," in *Asian Conf. on Computer Vision (ACCV)*, 2012, pp. 31–44.

[27] A. Ess, B. Leibe, K. Schindler, and L. van Gool, "A mobile vision system for robust multi-person tracking," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2008, pp. 1–8.

[28] W. Schwartz and L. Davis, "Learning discriminative appearance-based models using partial least squares," in *Brazilian Symp. on Computer Graphics and Image Processing (SIBGRAPI)*, 2009, pp. 11–14.

[29] C. Loy, T. Xiang, and S. Gong, "Time-delayed correlation analysis for multi-camera activity understanding," *Int. Jour. of Computer Vision (IJCV)*, vol. 90, pp. 106–129, 2010.

[30] UK Home Office, "i-LIDS multiple camera tracking scenario definition," 2008.

[31] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2005, pp. 886–893.

[32] K. Chen, C. Loy, S. Gong, and T. Xiang, "Feature mining for localised crowd counting," in *British Machine Vision Conf. (BMVC)*, 2012.

[33] L. Zheng, S. Wang, L. Shen, L. Tian, J. Bu, and Q. Tian, "Person re-identification meets image search," Tsinghua University, Tech. Rep., 2015.

[34] M. Oren, C. Papageorgiou, P. Sinha, E. Osuna, and T. Poggio, "Pedestrian detection using wavelet templates," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 1997, pp. 193–99.

[35] G. Overett, L. Petersson, N. Brewer, L. Andersson, and N. Pettersson, "A new pedestrian dataset for supervised learning," in *Intelligent Vehicles Symposium (IV)*, 2008, pp. 373–378.

[36] F. Flohr and D. M. Gavrila, "Pedcut: an iterative framework for pedestrian segmentation combining shape models and multiple data cues," in *British Machine Vision Conf. (BMVC)*, 2013.

[37] P. Luo, X. Wang, and X. Tang, "Pedestrian parsing via deep decompositional neural network," in *Int. Conf. on Computer Vision (ICCV)*, 2013, pp. 2648–2655.

[38] P. M. Roth, M. Hirzer, M. Koestinger, C. Beleznai, and H. Bischof, "Mahalanobis distance learning for person re-identification," in *Person Re-Identification*, ser. Advances in Computer Vision and Pattern Recognition, S. Gong, M. Cristani, S. Yan, and C. C. Loy, Eds. London, United Kingdom: Springer, 2014, pp. 247–267.

[39] A. Das, A. Chakraborty, and A. Roy-Chowdhury, "Consistent re-identification in a camera network," in *European Conf. on Computer Vision (ECCV)*, 2014, pp. 330–345.

[40] M. Halstead, S. Denman, S. Sridharan, and C. B. Fookes, "Locating people in video from semantic descriptions : A new database and approach," in *Int. Conf. on Pattern Recognition (ICPR)*, 2014, pp. 24–28.

[41] D. Baltieri, R. Vezzani, and R. Cucchiara, "Sarc3d: a new 3d body model for people tracking and re-identification," in *Int. Conf. on Image Analysis and Processing (ICIAP)*, 2011, pp. 197–206.

[42] B. Benfold and I. Reid, "Guiding visual surveillance by tracking human attention." in *British Machine Vision Conf. (BMVC)*, 2009, pp. 1–11.

[43] S. Wang, M. Lewandowski, J. Annesley, and J. Orwell, "Re-identification of pedestrians with variable occlusion and scale," in *Int. Conf. on Computer Vision Workshops (ICCV Workshops)*, 2011, pp. 1876–1882.

[44] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," *Int. Workshop on Performance Evaluation for Tracking and Surveillance (PETS)*, 2007.

[45] N. Martinel, C. Micheloni, and C. Piciarelli, "Distributed signature fusion for person re-identification," in *Int. Conf. on Distributed Smart Cameras (ICDSC)*, 2012, pp. 1–6.

[46] H.-S. Park and C.-H. Jun, "A simple and fast algorithm for k-medoids clustering," *Expert Systems with Applications*, vol. 36, no. 2, pp. 3336–3341, 2009.

[47] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 34, no. 4, pp. 743–761, 2012.

[48] C. Weinrich, C. Vollmer, and H.-M. Gross, "Estimation of human upper body orientation for mobile robotics using an svm decision tree on monocular images," in *Int. Conf. on Intelligent Robots and Systems (IROS)*, 2012, pp. 2147–2152.

[49] H.-M. Gross, K. Debes, E. Einhorn, St. Mueller, A. Scheidig, Ch. Weinrich, A. Bley, and Ch. Martin, "Mobile robotic rehabilitation assistant for walking and orientation training of stroke patients: A report on work in progress," in *IEEE Int. Conf. on Systems, Man, and Cybernetics (SMC)*. IEEE, 2014, pp. 1880–1887.

[50] H.-M. Gross, A. Scheidig, K. Debes, E. Einhorn, M. Eisenbach, St. Mueller, Th. Schmiedel, T. Q. Trinh, Ch. Weinrich, T. Wengefeld, A. Bley, and Ch. Martin, "Roreas: robot coach for walking and orientation training in clinical post-stroke rehabilitation: Prototype implementation and evaluation in field trials," *Autonomous Robots*, pp. 1–20, 2016.

[51] R. Stricker, St. Mueller, E. Einhorn, Ch. Schroeter, M. Volkhardt, K. Debes, and H.-M. Gross, "Konrad and Suse, two robots guiding visitors in a university building," in *Autonomous Mobile Systems (AMS)*, 2012, pp. 49–58.

[52] C. Weinrich, T. Wengefeld, C. Schroeter, and H.-M. Gross, "Generic distance-invariant features for detection of people with walking aid in 2d range data," in *Int. Symp. on Robot and Human Interactive Communication (RO-MAN)*, 2014, pp. 767–773.

[53] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv:1207.0580*, 2012.

[54] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[55] S. Walk, N. Majer, K. Schindler, and B. Schiele, "New features and insights for pedestrian detection," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 1030–1037.