Health, demographic change and wellbeing

Personalising health and care: Advancing active and healthy ageing

H2020-PHC-19-2014

Research and Innovation Action

ACANTO

A CyberphysicAl social NeTwOrk using robot friends

Deliverable 3.6

Interpreting the social context (final): Interpretation of the social context from the platform view and site-wide social context

Deliverable due date: April 2017	**Actual submission date: 21.08.2017**
Start date of project: February 1, 2015	**Duration: 36 months**
Lead beneficiary for this deliverable: UNITN	**Revision: —-**
Authors: Radu-Laurențiu Vieriu (UNITN)	
Internal reviewer: Josef Alois Birchbauer (SIEMENS)	

The research leading to these results has received funding from the European Union's H2020 Research and Innovation Programme - Societal Challenge 1 (DG CONNECT/H) under grant agreement n°643644

Dissemination Level		
PU	**Public**	
CO	**Confidential, only for members of the consortium (including the Commission Services)**	X

The contents of this deliverable reflect only the authors' views and the European Union is not liable for any use that may be made of the information contained therein.

Contents

Executive Summary

This deliverable builds upon D3.5 and extends the work of analyzing the social context from both platform view, as well as the site-wide view. More specifically, we focus our research and development efforts to answer the following questions: (i) how can one locate a person of interest in a given surveillance camera network? (ii) how can one identify unusual behavior in surveillance footage? (iii) how can monitoring low level facial appearance cues help modeling social behavior from the platform's perspective?

The answer to the first question is person re-identification (re-ID). In this deliverable, we dive deep into the underlying components of person re-ID and bring contributions on several levels. We first address detecting and tracking of pedestrians using surveillance cameras. We review existing work in pedestrian detection, commonly used benchmarks along with evaluation protocols and investigate the performance of a general purpose publicly available object detection system on pedestrian detection. By leveraging existing multimodal pedestrian data collections along with recent deep learning developments, we further present results of a novel cross-modal deep representation framework designed to “robustify” pedestrian detection under difficult recording conditions. Finally, we conduct a systematic analysis of person re-identification approaches and gain important insights into trends and good practices used in person re-ID.

To answer the second question, we conducted a research study to identify common approaches for anomaly detection in videos and pushed state-of-the-art further by developing a deep learning framework that automatically learns feature representations while combining appearance and motion from surveillance videos.

Finally, we extend social modeling from the platform's point of view to include face-to-face interactions with walker users. Here we base our statistics on face analysis components and combine them into a rule-based system able to cast probabilistic predictions about potential face-to-face social interactions.

1. Introduction

Recalling from D3.5, one of the main goals of task T3.3, to which both D3.5 and D3.6 (current document) subscribe, is to contribute to a high level of *situational awareness* meant to support the correct functionality of the FriWalk. In other words, T3.3 is concerned with finding relevant pieces of information from the vast pool of social context cues, that would lay the foundation for any activity execution and monitoring logic.

While sensing the environment, we are relying in ACANTO not only on sensors placed on the FriWalk itself, but also on distributed sensors, such as surveillance cameras that can be accessed in public places, like museums or shopping malls. Indeed, surveillance cameras offer a much wider and complete view on what is happening in the environment, when compared with pure platform perspective. In such surveillance scenarios, being able to recognize certain types of personnel (*e.g.* a nearby policeman, a shop assistant or a close friend) is sometimes a matter of critical importance. Imagine for a second our FriWalk user lost and disoriented in a crowded mall, or even worse, falling a victim of a theft on an empty pedestrian sidewalk, looking helpless as the thieve runs away with his/her valuables. In order to approach situations like these, we propose technological solutions to assist the elderly in finding the right person at the right time. Towards this goal, we investigate recent work in person re-identification and dive deep into its structural components, such as pedestrian detection and tracking, as well as pedestrian retrieval.

We first take a closer look at object detection with a focus on detecting pedestrians and evaluate state-of-the-art approaches on typically used pedestrian benchmarks. A subsequent person re-identification module enables identifying persons of interest in the form of a single image query retrieval. Secondly, we leverage existing multimodal pedestrian data collections along with recent deep learning developments to propose a cross-modal deep representation framework for robust pedestrian detection under difficult recording conditions. Finally, we complement the collection of context-related services with a component able to automatically locate anomalous dynamic behavior in surveillance videos. Our deep learning-based approach combines appearance and motion cues in a novel late fusion strategy.

To deal with social interactions, we switch perspective to the platform's point of view and study the potential of facial cues in revealing face-to-face interactions. We combine head pose and speech patterns into a probabilistic model which we then deploy in a real-time service ready to be integrated into the ACANTO framework.

2. Person re-identification using surveillance cameras

Person re-identification (for short person re-ID) is a fundamental aspect of multi-camera tracking and is concerned with establishing correspondences between images of a person captured from different cameras. From a technical perspective, person re-ID can be broken down into three sub-modules: person detection, person tracking and person retrieval. The first two components are considered independent computer vision tasks, such that most work done on person re-ID focuses on person retrieval, *i.e.* answering the question of whether a person of interest (query) is present or not in a gallery of person instances and if so, retrieve the instance(s) that match the query. In practice though, this question is typically implemented as retrieving the top *k* best matches from the gallery, closest (according to some distance metric) to the query. Person re-ID has gained a great deal of attention in the last years, evidence of this being the increasing share of papers accepted at major CV venues (see Fig. 2.1).

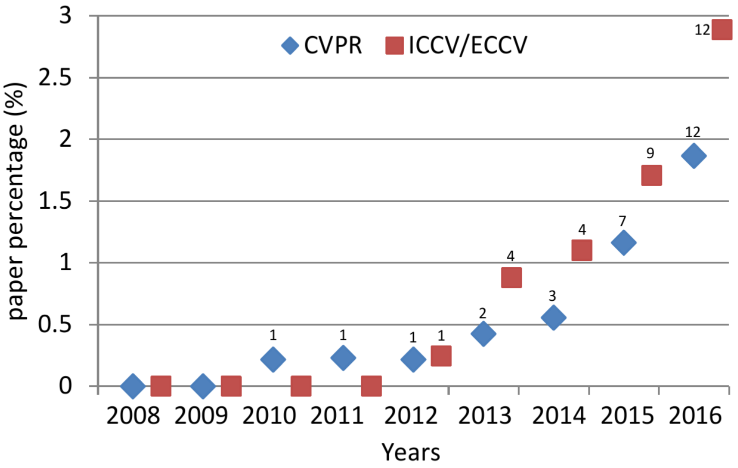


Fig. 2.1. Trends in the number of publications on person re-ID in top computer vision venues (courtesy of [1])

Interestingly, the "explosion" of publications in this field seems to be perfectly aligned temporally with the Deep Learning revolution. In fact, there's no surprise that most, if not all, recent approaches to person re-ID make some use of deep models.

We note that in ACANTO, for solving person re-ID -related tasks, such as retrieving persons of interest (*e.g.* policeman, medical caregiver, etc.), one needs only pedestrian detection and retrieval. Tracking is typically used in situations in which the temporal dimension is of interest. We therefore focus here only on detection and retrieval of pedestrians.

2.1. Pedestrian detection from surveillance cameras – an overview

In order to address the problem of people re-identification in a surveillance scenario, we have performed, as a first step, a study aimed at determining the performance of existing models on the task of pedestrian detection in low resolution images. We partly motivate the setup shift from omni-directional to surveillance cameras by the limitations of the former approach (which experienced accuracy difficulties) and partly by the recent progress of deep convolutional models for object detection. We argue that the increase in processing complexity for analyzing pedestrians is a negligible drawback considering the hardware support provided in cloud computing nowadays and comparing with the benefits in accuracy gained by deploying deep models.

Pedestrian detection is seen as a canonical instance of the much larger problem of object detection (OD) with large applicability in the automotive industry, video surveillance and robotics [2]. It is often used as a playground for exploring ideas that sound promising for generic OD. Despite extensive research done in the field, recent papers still report significant improvements, suggesting that there's still room to reach a saturation point. On the other hand, recent general OD systems have become quite competitive in performing pedestrian detection alone, due to a dominant representation of the pedestrian class in general OD benchmarks, such as PASCAL VOC [3], [4] , IMAGENET [5] or MSCOCO [6]. According to recent surveys [2], the driving force for performance in pedestrian detection has been the attention to features. The authors also stress the importance of optical flow and context information as complementary sources of information that are likely to boost the accuracy. Needless to say, as in many other vision challenges, state-of-the-art in pedestrian detection is nowadays claimed by Deep Learning models [7]–[9] and most of the research effort is channeled in exploiting additional cues or slightly modified architectures to improve a baseline that is already competitive.

In what follows, we present results of a preliminary study aimed at identifying available OD systems that perform well on pedestrian detection, focusing on solutions for which source code is available. A similar recent analysis [10] investigating the potential of Faster R-CNN for pedestrian detection has shown that by compensating for the insufficient resolution of feature maps for handling small instances as well as for the lack of any bootstrapping strategy for mining hard negative examples, one can turn a general OD system into a state-of-the-art pedestrian detector. We first review common benchmarks along with typical performance evaluation protocols.

a) Pedestrian detection benchmarks

One of the oldest datasets for pedestrian detection is INRIA [11] which stands out thanks to high quality annotations of 1800+ pedestrians in various settings (*e.g.* city, beach, mountains). While this is particularly advantageous for training (in fact, INRIA seems to be a diverse enough dataset to allow good generalization to other collections [2]), INRIA is not among the most commonly used benchmarks for pedestrian detection. Instead, Caltech-USA [12] has gained a lot of popularity in the recent years. Caltech brings along 10 hours of video footage recorded at 30Hz from a vehicle driving in regular traffic on the streets of LA, US. Annotations sum up to 350k bounding boxes (BBs) corresponding to roughly 2300 unique pedestrians. Few example frames can be seen in Fig. 2.1.1, in which one can spot two kinds of annotations: solid green rectangles, corresponding to full size pedestrians, and dashed yellow boxes depicting the visible areas of occluded pedestrians.

As the authors notice [12], the probability map encoding the likelihood of a pixel to be occluded given that the pedestrian is occluded is highly biased towards the lower part of the BB, the kind of additional information one can exploit in order to improve detection. In addition, Caltech contains temporal correspondences between BBs as well as detailed occlusion labels. The authors have released source code implementing the evaluation protocol and are also keeping track of recent work, by allowing other authors to submit their results online. All these features have made Caltech-USA the most popular benchmark for pedestrian detection and tracking.

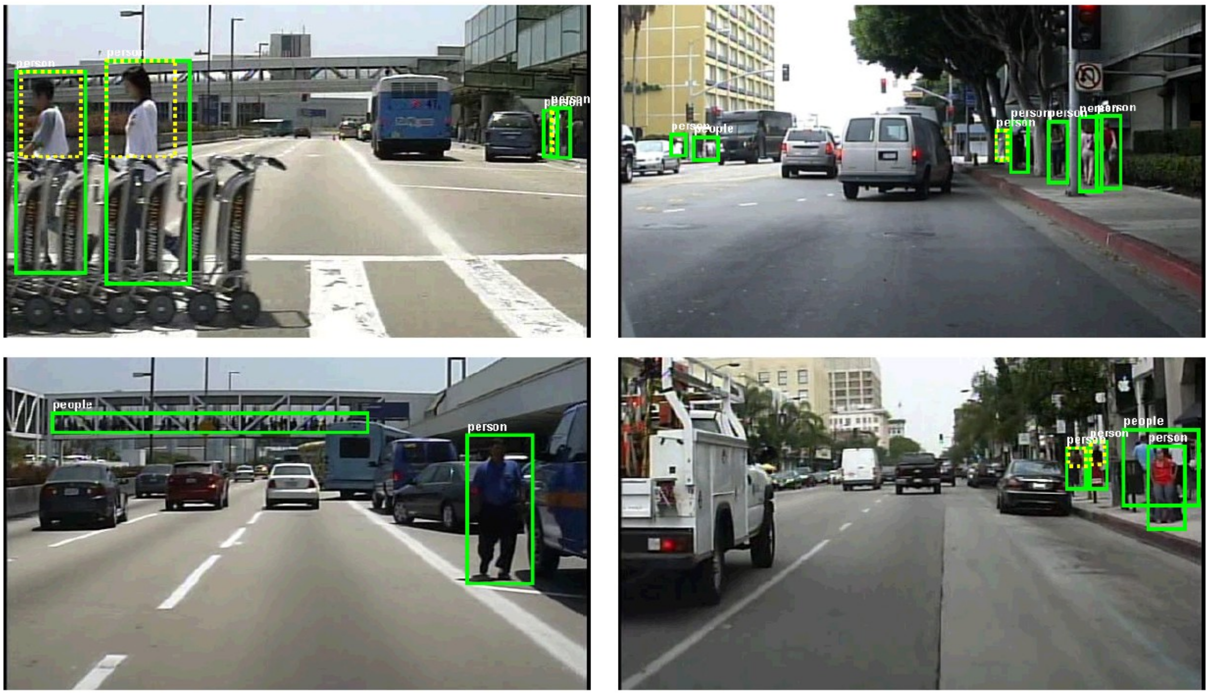


Figure 2.1.1. Sample images from Caltech-USA dataset along with ground truth annotations. The annotations contain full pedestrian bounding boxes (ignoring occlusions) – green rectangles, as well as boxes showing the visible area – yellow dashed rectangles.

b) Performance measures

Performance evaluation tools for Caltech have been initially developed in [13] and subsequently reviewed in [12]. The current protocol defines four evaluation scenarios by splitting the 11 recorded sessions (S0-S10) into two groups: S0-S5 and S6-S10. First two scenarios, *ext0* and *ext1* allow authors to develop systems on any external data and test them on S0-S5 and S6-S10 respectively. The third scenario (*cal0*) asks for a 6-fold cross-validation over S0-S5, while in the fourth (*cal1*), authors are asked to train on S0-S5 and report results on S6-S10. The richness of the evaluation scenarios allows for previous systems, trained on existing data collections, to be tested on Caltech.

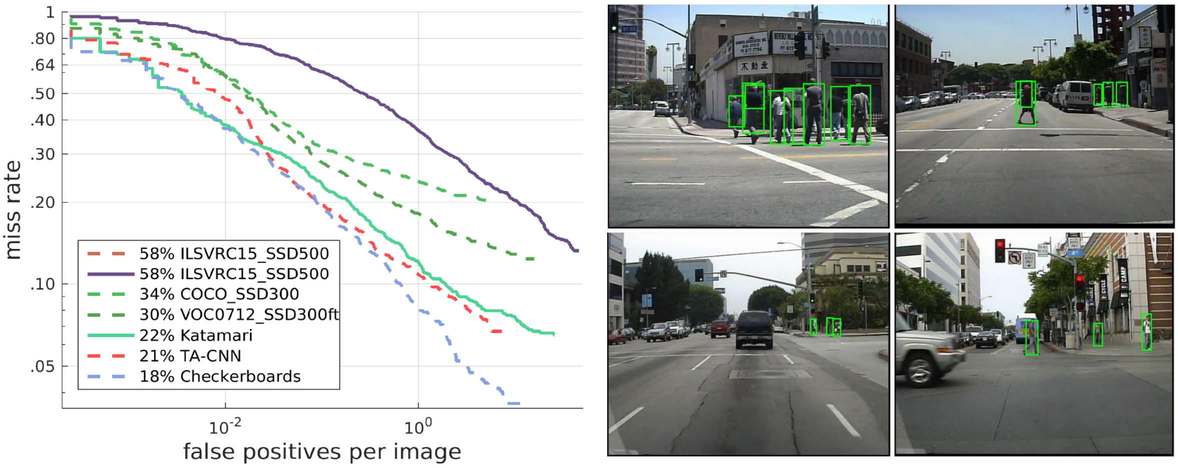
The core performance measure of the evaluation process in Caltech is the overlapping area between a detected bounding box BB_{pred} and the corresponding ground truth box BB_{gt}. As in PASCAL challenge, this measure is computed as the intersection over the reunion (IoU between the two BBs and, in order for a detection to be counted as a potential match, IoU needs to exceed 0.5. Detections with highest confidence are matched first, whereas ambiguities are solved in a greedy fashion. At the end, unmatched BB_{pred} boxes are counted as false positives, while unmatched BB_{gt} as false negatives. From the above counts, the miss rate (MR) versus the number of false positives per image (FPPI) is built in log-log scale, by varying the threshold on the detection confidence. Additionally, the log-average miss rate is computed by averaging MR over evenly spaced FPPI values in the range $10^{-2} - 10^{0}$, to summarize the performance of a detector by one single value. In practice, this average value is close (if not similar) to the MR value of the detector at FPPI equal to 10^{-1}.

Another important aspect of the evaluation protocol is that the authors have isolated a set of ground truth samples for which the BBs of the pedestrians are at least 50 pixels tall and correspond to cases of no or partial occlusions. This subset is called *reasonable* and it is widely used in literature to report pedestrian detection performance.

c) SSD as a pedestrian detector

A recent study [14] has introduced a unified deep neural network framework for general object detection based on discretizing the output space of BBs into a set of default boxes over different aspect ratios and scales per feature map location. Their system, called Single Shot multibox Detector (SSD), is interesting because it completely eliminates traditional bounding-box proposal generation as well as feature resampling, making it much faster at testing time, while maintaining competitive accuracy. This makes it very appealing for practical scenarios such as in ACANTO, where monitoring pedestrians in real time is a valuable asset.

The core of SSD is predicting category scores and box offsets for a fixed set of default BBs using small convolutional filters applied to feature maps. SSD has been validated on several general OD data collections (including PASCAL VOC [3], [4] , IMAGENET [5] and MSCOCO [6]), all of which contain a class for pedestrians, under various settings. Most of the models are available online for testing. We took the liberty of testing them on detecting pedestrians using Caltech-USA as benchmark and the testing protocol explained in 2.1.b). Figure 2.1.2 a) highlights the MR vs. FPPI curves of several models (including some recent ones, specifically developed for the task [2], [8], [15]). Despite achieving only moderate results on the MR scale (which are still higher than the baselines reported in [12]), SSD models perform pedestrian detection reasonably well (in particular the ones trained on PASCAL datasets) considering they are only general OD systems. Figure 2.1.2 b) is a qualitative proof of this statement. We are, however, still far from what is reported as human performance on Caltech [16], *i.e.* 5.6% miss rate in the vicinity of 1 false positive per 10 images (against approx. 30% obtained with the best SSD model).



a) b)

Fig. 2.1.2 Performance comparison of SSD models on Caltech pedestrian detection benchmark. The numbers represent the percentage of miss rate (MR) at a false positive per image (FPPI) of 1 every 10 images. SSDs are compared with Katamari, TA-CNN and Checkerboards, all which, to the best knowledge of the authors, are not publicly available. As a reference, a human operator misses, at the same FPPI rate, approx. 5.6% true positives, according to some studies.

This motivates a further study on how to improve pedestrian detection, especially under challenging conditions (such as crowded public spaces).

2.2. Exploiting multiple modalities for improved pedestrian detection

Looking back on pedestrian detection research, particularly the more realistic playgrounds in which pedestrians are captured in cluttered background or undergo substantial occlusions, as we expect to be the case in ACANTO's shopping malls and museum scenarios, we notice two main trends that drive recent advancements: Deep Learning – undoubtedly one of the key engines to improve performance [8], [9], [17] and the adoption of additional sensors (such as thermal or depth cameras) that bring complementary information to tackle problems such as adverse illumination conditions and occlusions [18], [19][20]. However, the vast majority of wide camera networks in surveillance systems still employ traditional RGB sensors and detecting pedestrians in case of illumination variation, shadow, and low external light remains a challenging open issue.

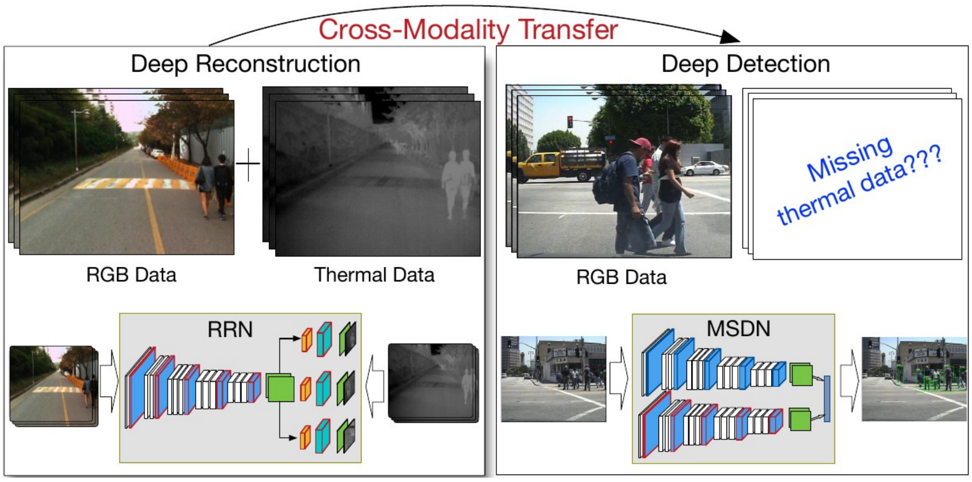


Fig. 2.2.1. Overview of the proposed framework. Our approach relies on two networks. The first network is used to learn a nonlinear feature mapping between RGB and thermal image pairs. Then, the learned model is transferred to a target domain where thermal inputs are no longer available and a second network is used for learning a RGB-based pedestrian detector.

We take inspiration from recent work showing the benefits of leveraging cross-modal data in solving detection and recognition tasks [21][22] and combine it with recent unsupervised deep learning techniques to develop a CNN-based approach for learning cross-modal representations for pedestrian detection which does not require bounding box annotations. More specifically, we use multispectral data and CNNs to learn a mapping from RGB space to thermal representation without human supervision (see Fig. 2.2.1 for an overview). Importantly, thermal data are not needed at testing time to perform pedestrian detection from RGB images. The intuition behind using thermal data is that by doing so, there's a good chance to increase the discrimination power of the classifier by addressing difficult images that look like pedestrians in the RGB space (such as electric poles or trees) but have a completely different representation (and thus are much easier to discriminate) in the thermal space (see Fig. 2.2.2. for few examples). In other words, we aim at addressing hard negative samples by looking at them in a space that makes them easy negatives.

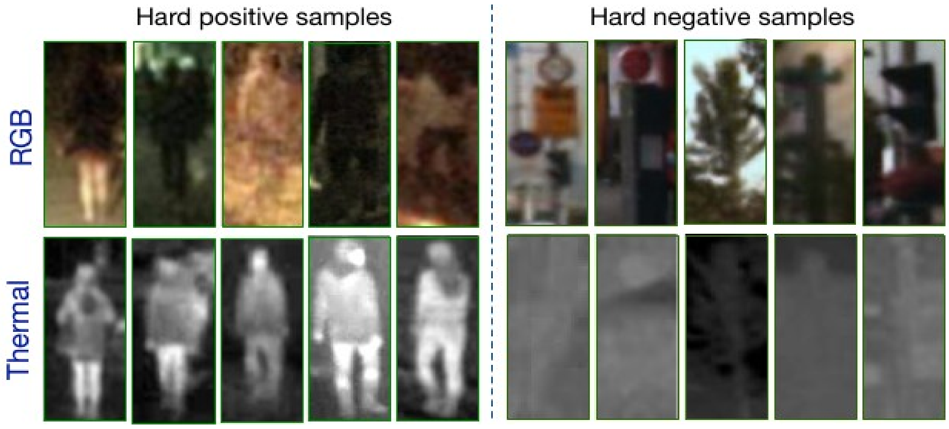


Fig. 2.2.2. Exploiting thermal data in addition to RGB samples makes it is easier to discriminate among pedestrians and background clutter.

Our proposed architecture is based on two different CNN networks, associated to the reconstruction and the detection tasks, respectively. The first deep model, *i.e.* the Region Reconstruction Network (RRN), is a fully convolutional network trained on pedestrian proposals collected from RGB-thermal image pairs in an unsupervised manner. RRN is used to learn a non-linear mapping from the RGB channels to the thermal channel. In the target domain, only RGB data are available and a second deep network, the Multi-Scale Detection Network (MSDN), embedding the parameters transferred from RRN, is used for robust pedestrian detection. MSDN takes a whole RGB image and a number of pedestrian proposals as input and outputs the detected BBs with associated scores. In the test phase, detection is performed with MSDN and only RGB inputs are needed.

The training process involves two main phases. In the first phase, RRN is trained on multispectral data (*i.e.* pairs of RGB-thermal images). The front-end convolutional layers of RRN are initialized using the parameters of the 13 convolutional layers of the VGG-16 model [23] pretrained on ImageNet dataset. The remaining parameters are randomly initialized. Stochastic Gradient Descent (SGD) is used to learn the network parameters. In the second phase, the parameters of MSDN are optimized using RGB data and pedestrian bounding box annotations from the target domain. MSDN contains two sub-networks (Subnet A and B), one of which is a copy of the convolutional layers from RRN. These layers, learned in the first phase, are used to initialize part of MSDN in the second phase. Fine-tunning for MSDN is performed on the target pedestrian dataset using back-propagation and SGD.

Datasets

We validate our framework on two public datasets: KAIST multispectral pedestrian dataset [18] and Caltech-USA. KAIST contains images captured under various traffic scenes with different illumination conditions (*i.e.* data recorded both during day and night). The dataset consists of 95k aligned RGB-thermal image pairs, of which 50.2k samples are used for training and the rest for testing. A total of 103,128 dense annotations corresponding to 1,182 unique pedestrians are available. We follow the evaluation protocol outlined in [18] in our experiments. The performance is evaluated on three different test sets, denoted as *Reasonable all*, *Reasonable day* and *Reasonable night*. Reasonable here indicates that the pedestrians are at most partially occluded and contain more than 55 pixels height. The *day* and *night* sets are obtained from the *Reasonable all* set according to the capture time. From Caltech we used the Caltech-All and Caltech-Reasonable settings.

Experimental details

We implemented our framework using *Caffe* [24] on an Intel(R) Xeon(R) CPU E5- 2630 with a single CPU core (2.40GHz), 64GB RAM and an NVIDIA Tesla K80 GPU. We employ ACF [25] to generate pedestrian proposals for training both the reconstruction and the detection network with a low detection threshold of -70 as in [9] to obtain a high recall of pedestrian regions. In the test phase we also use ACF and consider the test proposals available online[1]. It is worth nothing that, while we focus on ACF, our cross-modality learning approach can be used in combination with an arbitrary proposal method.

For training the reconstruction network, we use the whole training set of the KAIST dataset. As thermal images captured from an infrared device have relatively low contrast and significant noise, we perform some basic processing, such as adaptive histogram equalization and denoising. By computing pedestrian proposals applying ACF, we end up creating a dataset of about 20K frames for training the reconstruction network. All the frames are then horizontally flipped for data augmentation. The mini-batch size is set to 2 and a fixed learning rate $\lambda_r = 10^{-9}$ is used to guarantee smooth convergence. We train the RRN for about 10 epochs.

For training the detection network, we follow previous work [15] and for the Caltech dataset we construct a training set where every every 3rd frame is used. Instead, for the KAIST dataset we adopt the standard training protocol and every 20th frame is considered. For both datasets, we use the same protocol for training MSDN. Similarly to RRN training, the data are flipped horizontally for data augmentation. Each mini-batch consists of 80 pedestrian proposals randomly chosen from one training image. Positive samples with a ratio of 25% are taken from the proposals which have an IoU overlap with the ground truth of more than 0.5, while negative samples are obtained when the IoU overlap is in the range of [0, 0.5]. SGD is used to optimize MSDN with the momentum and the weight decay parameters set to 0.9 and 0.0005, respectively. The network is trained with 8 epochs using an initial learning rate of 0.001 and drop by 10 times at the 5th epoch.

Results on KAIST

On KAIST dataset we derive a series of model variations for our CMT-CNN (Cross-Modality Transfer CNN), using different settings:

- CMT-CNN-SA – only Subnet A from MSDN is used for predicting pedestrian BBs.
- CMT-CNN-SA-SB (ImageNet) – both MSDN subnets are used but they are both initialized with VGG16 convolutional weights pretrained on ImageNet
- CMT-CNN-SA-SB (random) – similar to previous case, with the difference being that the weights of Subnet B are randomly initialized
- CMT-CNN – finally, this is the model for which Subnet B of MSDN is initialized with the weights of the RRN, previously pretrained on pairs of RGB-thermal images from KAIST

[1] http://www.vision.caltech.edu/Image$_$Datasets/CaltechPedestrians/

Table 2.2.1: Comparison of different methods on the KAIST multi-spectral datasets including reasonable all, reasonable day and reasonable night settings

Model	All	Day	Night
CMT-CNN-SA	54.2%	52.4%	58.9%
CMT-CNN-SA-SB (random)	56.7%	54.8%	61.2%
CMT-CNN-SA-SB (ImageNet)	52.1%	50.7%	57.6%
CMT-CNN	49.5%	47.3%	54.7%

Table 2.2.1 shows the results of the above models on KAIST dataset. The table reveals the benefits of using thermal data for improving pedestrian detection. Qualitative results are shown in Fig. 2.2.3, where we compare our CMT-CNN model (last row) with a standard pedestrian detector (ACF, top row) as well as one variation of our model (CMT-CNN-SA, middle row). We note a great reduction in false positives, as we move down the rows, particularly in the last one.

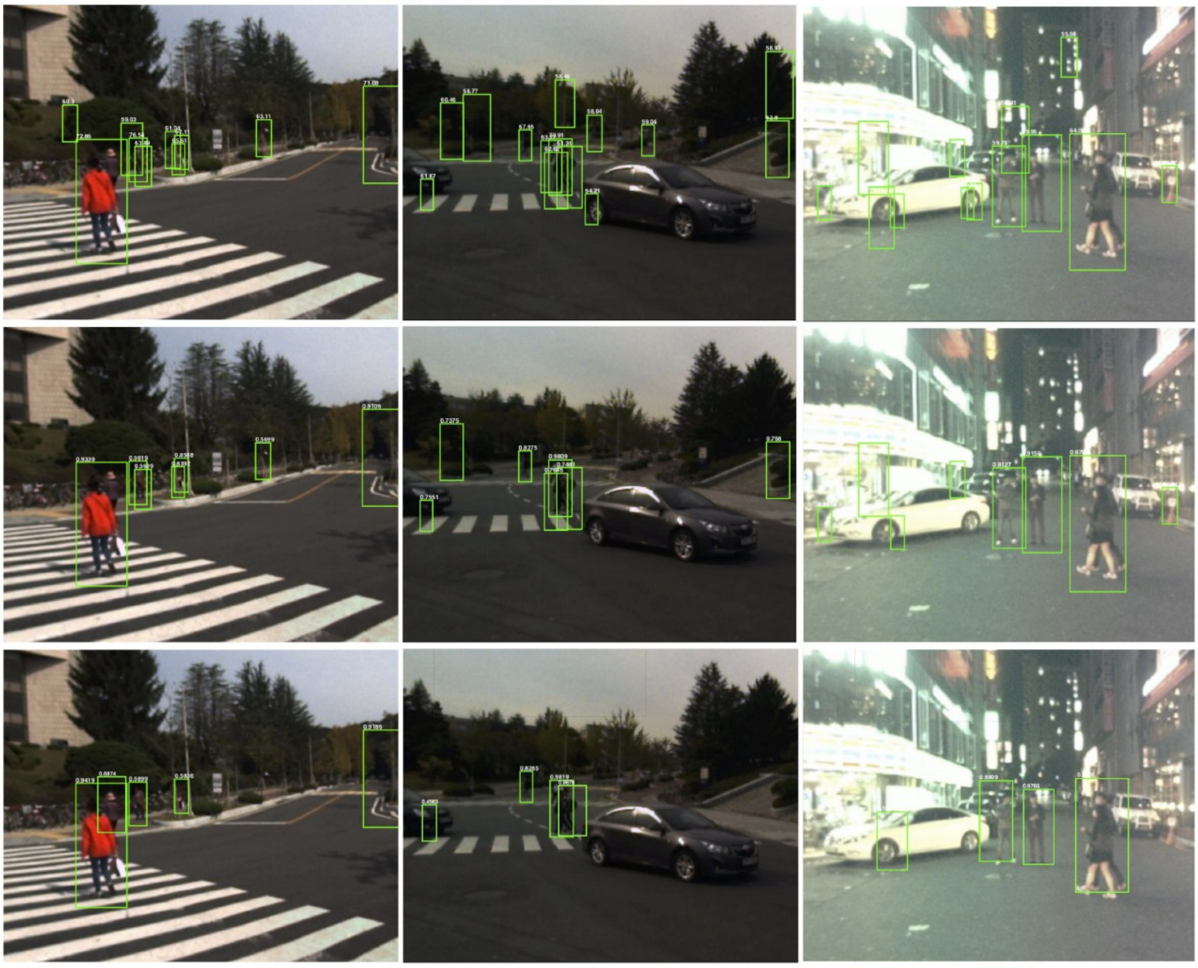


Fig. 2.2.3. Examples of pedestrian detection results under different illumination conditions on the KAIST multispectral pedestrian dataset: (top) ACF detector, (middle) CMT-CNN-SA, (bottom) CMT-CNN

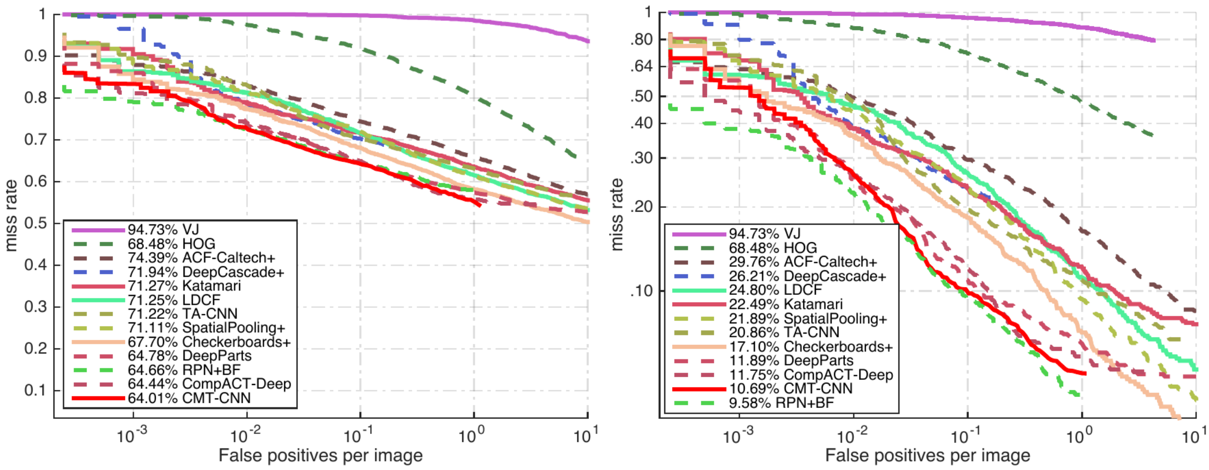
Results on Caltech

We tested the same model variants described above, to which we added CMT-CNN-SA-SB (RGB-KAIST), in which the weights of VGG16 used to initialize Subnet B of MSDN were pretrained on KAIST (using only RGB images). The results of the comparison are shown in Tab 2.2.2. Again, the numbers in the table confirm the effectiveness of our approach.

Table 2.2.2: Comparison of different variants of our method on the Caltech-Reasonable dataset. Performance are evaluated in terms of log-average miss-rate

Model	Average Miss Rate
CMT-CNN-SA	13.76%
CMT-CNN-SA-SB (random)	15.89%
CMT-CNN-SA-SB (ImageNet)	13.01%
CMT-CNN-SA-SB (RGB KAIST)	12.51%
CMT-CNN	10.69%

We further compare CMT-CNN against a large number of published results on Caltech. We consider both *All* and *Reasonable* subsets of Caltech and the following approaches: Viola-Jones (VJ) [26], Histograms of Oriented Gradients (HOG) [11], DeepCascade+ [17], LDCF [27], SCF+AlexNet [7], Katamari [2], SpatialPooling+ [28], SCCPriors [29], TA-CNN [8], CCF and CCF+CF [30], Checkerboards and Checkerboards+ [15], DeepParts [31], CompACT-Deep [32] and RPN+BF [10]. Results can be visualized in Fig. 2.2.4.



a) b)

Fig. 2.2.4. Quantitative results on Caltech dataset (*All* – a) and *Reasonable* – b))

2.3. Building pedestrian heat maps from detections

Recalling from D3.5, heat maps were created using omnidirectional cameras, by first detecting people's head (using motion cues and segmentation approaches) and inferring their feet coordinates (step called foot correlator), followed by a dewarping process of these coordinates, knowing camera calibration parameters, and finally updating the heat map along the temporal axis. One of the main drawbacks of using omnidirectional settings was the fact that the heat maps were much less accurate at the periphery, due to perspective distortion. In addition, the presence of objects in the scene that would occlude people's feet was an additional factor affecting the quality of the heat maps. We focused on improving estimating the positions of the feet by means of geometric tools, however we noted a very marginal improvement in the final quality of the heat maps. We therefore decided to switch to general surveillance cameras.

What we propose now in the surveillance setting is using the detection bounding boxes (BBs), alongside with scene geometry and few assumptions in order to accumulate pedestrian trajectories into heat maps. More specifically, for every detected pedestrian in a surveillance image, we consider the bottom midpoint on the BB as an anchor point (similarly to a projection point) and use this to represent a person on the flat surface captured by the camera. We therefore assume (as before, in the omnidirectional case) that the surface onto which pedestrians are walking is flat. Knowing camera geometry, we can project a given pixel from the camera plane onto a new coordinate space where we compensate for the camera perspective distortion. In a very simplified approach, we are looking at solving a standard planar homography problem, where the objective is to estimate a transformation 3×3 matrix H, that maps pixel coordinates from the camera plane to a new set of coordinates in which the perspective distortion has been compensated for (alongside a given planar surface). An example of such a transformation can be seen in Fig. 2.3.1. Estimating matrix H is done by providing a set of corresponding 2D points between the two views.

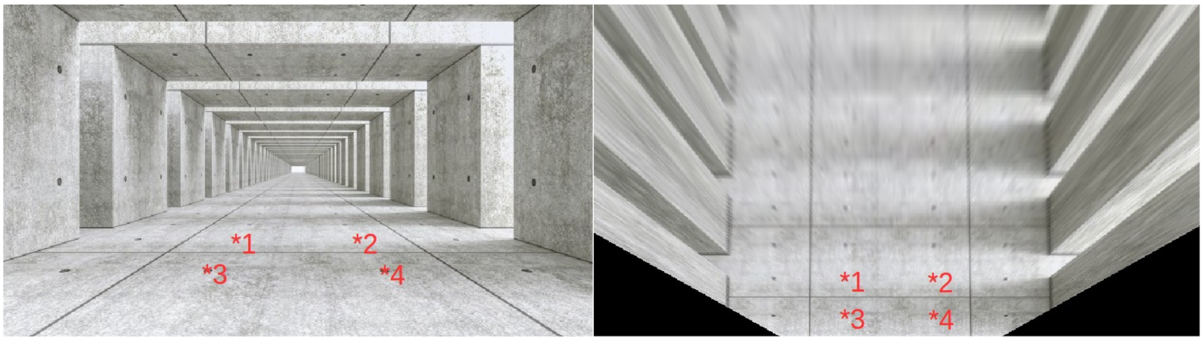


Fig. 2.3.1. Example of a planar homography. Original image (left) is homographically transformed so that the floor plane appears with no perspective distortion (right). Matrix H is determined by solving the matching between the 4 corresponding points.

Having the floor compensated for perspective distortion gives us the base for the heat map, onto which the accumulated pedestrian traces will be overlaid. Similarly to D3.5, we assume the pedestrian footprint to be a circle and for each pair of coordinates in the original image (corresponding to a detected BB), we “paint” an area in the heat map that correspond to the considered circle. Accumulating these footprints over a temporal window gives us the final heat map.

Figure 2.3.2 shows a real world example of a heatmap created using our proposed approach. The images were taken from the “Mall Dataset” [33], containing surveillance footage of people walking along a mall corridor. We first extract pedestrian bounding boxes and represent each box using the bottom midpoint as anchor, as described above. We then take a reference frame from the surveillance collection that contains as few pedestrians as possible and back-project it to correct for the perspective distortion. Onto that image, we paint the pedestrian trajectories over time (knowing their 2D locations), using a “brush” in the form of a disc proportional to the footprint left by a pedestrian.

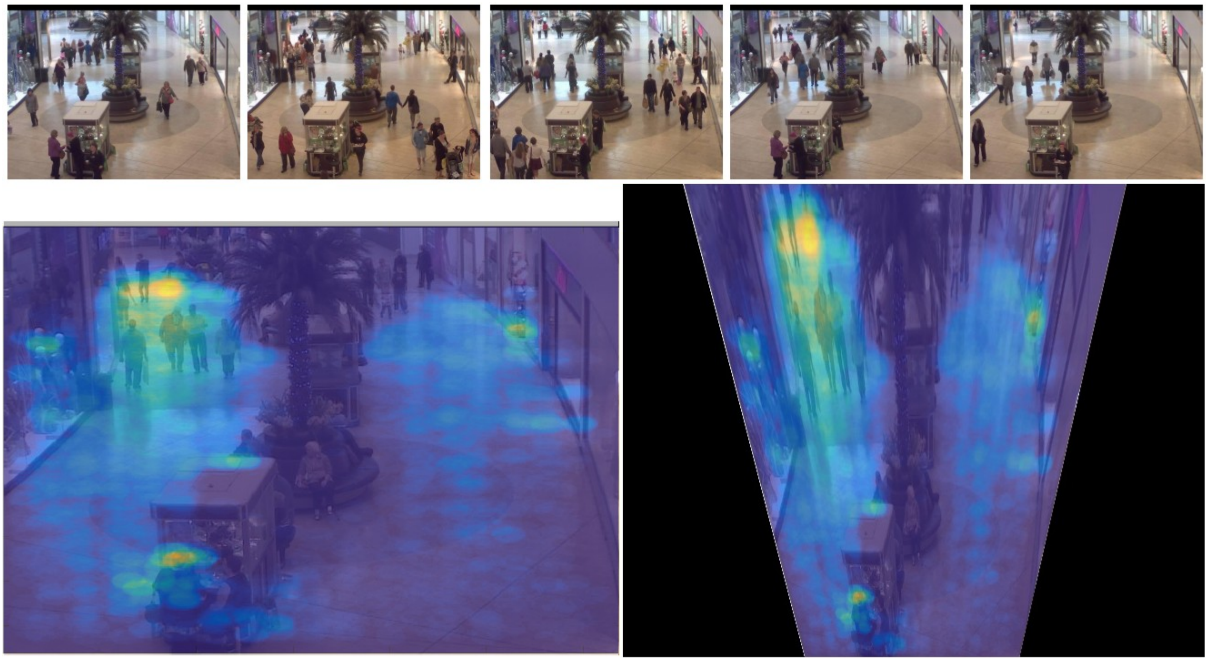


Fig. 2.3.2. Sample images from the Mall Dataset (up row). Heat map overlaid onto a reference image (left) and the same heat map projected to compensate for the perspective distortions (right).

2.4. Person re-identification – an overview

From a computer vision perspective, one of the main difficulties person re-ID research has to overcome is finding a good representation of a person, *i.e.* extracting discriminative enough features such as to have a robust enough representation able to control the within-class and between-class variances in order to have the former smaller than the latter. Assuming that detection and tracking are possible, then the first objective towards person re-ID is learning a visual descriptor. The simplest low-level visual descriptors are appearance-based. They rely on visual cues such as color and texture to describe the appearance. These descriptors are most sensitive to changes in illumination, pose or camera viewpoint. If we add to the mix the unconstrained nature of the recording environment (which translates into lack of cooperativeness or potential presence of occlusions or background clutter), as well as the difficulty of ensuring high quality data (resolution, frame rate), then we have a clearer picture of why this task is particularly difficult. While much effort has been channeled into finding a good representation for re-ID, other works have been shifting the focus towards distance metric learning. These methods aim at learning appropriate distance metrics that can maximize the matching accuracy regardless of the choice of appearance representation. The main idea here is to learn a metric in the space defined by image features that keep those coming from the same class closer, while the ones coming from different classes further apart. In the context of re-ID, the image features are appearance descriptors across camera views and the aim is to learn a distance metric in the appearance space that maximizes the distance between descriptors of different people and minimizes the distance for descriptors of the same person.

In spite of all the challenges, the community has found ways to push the advancements in person re-ID, often exploiting context information (such as camera geometry), but mostly taking advantage of deep learning (ever since the first two publications [34], [35] appeared in

2014), which offers a highly versatile way of representing data, given that enough of them are available for training. In what follows, we present a small survey of most recent work published at CVPR this year, in order to identify trends in dealing with the challenges in person re-identification. First though, we will present common benchmarks used in the community (summarized in Tab. 2.4.1.), as well as most frequently reported evaluation metrics.

Commonly used datasets:

- **VIPeR**: VIPeR [36] is the most tested benchmark in person re-ID. The dataset contains 632 pedestrian image pairs taken by two different cameras. The cameras have different viewpoints and is not free from illumination variations. The images are cropped and scaled to be 128 × 48 pixels. This is considered one of the most challenging datasets for automated person Re-ID. Typically, 10 random train/test splits are used for stable performance, and each split has 316 different identities in both the training and testing partitions.
- **iLIDS** [37]: captured at an airport arrival, this dataset contains 476 images of 119 person identities captured with two non-overlapping cameras. On average, each identity is represented by 4 different images, with a minimum of 2. The dataset has considerable illumination variations and occlusions across the two cameras. All images are normalized to 128 × 64 pixels.
- **PRID 450S** [38]: this dataset consists of 450 images pairs of pedestrians captured by two non-overlapping cameras. The main challenges are related to changes in viewpoint, pose as well as significant differences in background and illumination. The widely adopted experimental protocol on this datasets is that a random selection of half the number of persons is used for training and the rest for testing. The procedure is repeated for 10 times, then the average performance is reported.
- **CUHK01** [39]: contains 971 identities captured from two camera views (A and B) in a campus environment. Camera view A captures frontal or back views of a person while camera B captures the person's profile views. Each person identity has 4 images with two from each camera, summing up a total of 3,884 images.
- **CUHK03** [35]: contains 1,360 identities captured by six surveillance cameras in a similar campus environment. Each identity is captured by two disjoint cameras. In total there are 13,164 images with each identity having on average 4.8 images. Differently from previous datasets, this one provides two types of annotations, including manually annotated bounding boxes (called *labeled*) and bounding boxes detected using DPMs (*detected*).
- **Market-1501** [40]: this is currently the largest benchmark in person re-identification, comprising 1,501 identities and 32,668 images captured using 6 cameras in front of a supermarket in Tsinghua University. The bounding boxes were detected using DPMs. Half of these identities (about 13k images) are used for training, while the remaining half (about 19k images) are used for testing. During testing 3,368 images are taken as the probe to identify the correct identities on the testing set.

Evaluation metrics: The most commonly used evaluation metric for person re-ID is the cumulative matching characteristic curve (CMC). CMC reflects the fact that person re-ID is formulated as a ranking problem, where elements from the gallery are ranked based on their comparison with the probe. The CMC curve shows the probability that a query identity appears in different-sized candidate lists. This evaluation method is only accurate if there is only one single ground truth image in the given query, which seems to be the case for most benchmarks. However, in some cases (*e.g.* Market-1501) there are multiple ground truth samples for each query. In these situations, CMC cannot distinguish between algorithms experiencing different recall rates. As a result, Rank-1 recognition rates along with mean average precision (mAP) values have been adopted in many papers.

Tab. 2.4.1 Summary of most commonly used benchmarks for person re-ID

Dataset Name	Release Year	no. IDs	no. Images	no. Cameras	Label type
VIPeR [36]	2007	632	1264	2	manual
ILIDS [37]	2009	119	476	2	manual
PRID450S [38]	2014	450	900	2	manual
CUHK01 [39]	2012	971	3884	2	manual
CUHK03 [35]	2014	1467	13164	2	manual/DPM
Market-1501 [40]	2015	1501	32668	6	DPM

A systematic overview of the most recent work published in a top computer vision gathering (such as CVPR17) reveals that the trends in person re-ID research follow (with few exceptions) the two main directions defined at the beginning of this paragraph, namely pushing for more discriminative data representations and investigating novel, more informative distance metrics. In the first category, all the works are dominated by deep learning approaches. For instance [41] emphasizes the need of capturing small scale visual cues for increased discriminative power. To this goal, they propose to use Multi-Scale Context-Aware Networks (MSCAN), which contain multi-scale convolutional layers, to encode image context information. Secondly, Spatial Transformer Nets equipped with problem-specific spatial constraints are used to localize pedestrian parts. The two resulting representations are fused and jointly optimized in an end-to-end approach. Along the same line, SpindleNet [42] uses a region proposal network to drive feature extraction of seven body sub-regions (in order to ensure correspondence between feature locations of different images). Slightly differently, [43] takes one step back looking at the whole camera network and aims at optimizing performance at this level, instead of focusing on pairs of cameras. To this goal, the authors propose a consistent-aware deep learning (CADL) approach, whose objective is finding a global optimum matching for the entire camera network.

Moving towards distance metric learning, we identify transition works that attempt to combine discriminative representations with some non-standard metrics. The work from [44] aims at improving data representation by introducing a margin-based online hard negative mining quadruplet loss designed to reduce the intra-class feature variation while enlarging the inter-class one. Similarly, [45] combines the discrimination power of deep models with the versatility of a novel metric learning based on a point-to-set (P2S) similarity comparison. Incorporating a pair-wise term (for overfitting robustness), a triplet term (for controlling distances between positive pairs and negative pairs) and a regularization term (for smoothing parameters and numerical stability), the P2S metric proves consistently superior to the point-to-point (P2P) counterpart.

Focusing more on metric learning, [46] takes a closer look at the manifold onto which images reside. In an attempt to smooth the learned metric wrt. the local geometry of the data manifold, the authors propose a general purpose manifold-preserving approach that handle similarities between two instances in the context of other pairs, thus better reflecting the geometry structure of the manifold. Their algorithm, called Supervised Smoothed Manifold

(SSM) shows superior performance on person re-ID, especially on large datasets (CUHK03, Market-1501). Chen et al. [47] address cross-camera variations by using a hashing based method that transforms the original feature representation into compact identity-preserving binary codes. A more systematic approach to distance metric learning is presented in [48] where the problem of color variation among different cameras is tackled in a one-shot-learning approach, which divides the metric learning into two components: a texture component (applied on intensity images) and a color component. For texture, a color-invariant deep representation (CNN-based) is being learned. Color, however, is being incorporated into the model using handcrafted color features for which a color metric is being learned for each camera pair using a ColorChecker chart along with a one-shot learning formulation. To learn this metric, only one example per camera is being required, allowing for easy scalability to large camera networks. Experimental results validate this approach in the context of semi-supervised or unsupervised person re-ID.

Tab. 2.4.2. Performance evaluation of most of CVPR17 person re-ID publications

Method	Rank-1 score						Obs.
	VIPeR	iLIDS	CUHK01	CUHK03	PRID 450S	Market 1501	
MSCAN [41]	-	-	-	74.2	-	80.3	Feature learning
SpindleNet [42]	53.8	66.3	79.9	88.5	-	76.9	Feature learning
CADL [43]	-	-	-	-	-	80.8	Feature learning
BL [44]	49.0	-	62.5	75.5	-	80.3	Feature + metric
P2S [45]	-	-	77.3	-	-	70.7	Feature + metric
SSM [46]	53.7	-	-	76.6	72.9	82.2	Metric learning
CSBT [47]	33.1	-	48.0	46.2	-	42.9	Hashing
One-Shot ML [48]	34.3	51.2	45.6	-	41.4	-	Semi-supervised
RR [49]	-	-	-	61.6	-	77.1	Ranking refinement

Finally, a slightly different approach is considered in [49], where focus is channeled onto refining an existing ranking result. Assuming one retrieves a list of k-nearest neighbors of a given probe (query), ranked by a distance metric. The underlying idea here is that if the probe itself is present among the k-nearest neighbors of the ranked samples (in the feature space), thus being k-reciprocal neighbors, then those samples are more likely to be positive matches with the probe.

Table 2.4.2 creates a unified summary of the performance of the above described approaches (in terms of Rank-1 scores) and enables direct comparison between them.

2.5. Person re-ID in ACANTO

Investigating most recent work in person re-ID made possible for us to gain a solid perspective over the topic and familiarize with best practices, common benchmarks and performance measures. However, when it comes to deployment, we are relying on the previously developed approach that computes human bodyprints for re-identification. While actively searching for a promising approach suitable for ACANTO, our current implementation uses a standard RGB sensor to extract bodyprints based on the detected bounding boxes (BBs) from our pedestrian detector.

We use the BBs generated by our pedestrian detector to calculate the approximate positions of the head and feet, respectively. Note that, unlike in previous approach, where we were performing head detection first by using video motion detection (VMD) and segmentation, followed by computing the feet positions using a foot correlator, we are now able to compute these positions at frame level. Similarly to computing heat maps, for every BB we consider the top and bottom most middle points as proxy for head and feet positions, respectively.

Once the head and feet locations are determined, we take all the pixels contained on the line that connects these two locations and accumulate these pixels over a time interval. We therefore form a temporal bodyprint of the person, which is further adjusted so that the waist line is horizontal along the time axis. We correct the waist line by first fitting a polynomial curve to the bodyprint and then applying the inverse transformation. Few examples showing this process can be seen in Fig. 2.5.2.

The corrected bodyprints serve as a basis for extracting visual features (color cues, saturation, contrast), which are then used to match a query print to a gallery of existing potential matches. We use simple euclidean distance to match feature vectors coming from bodyprints.

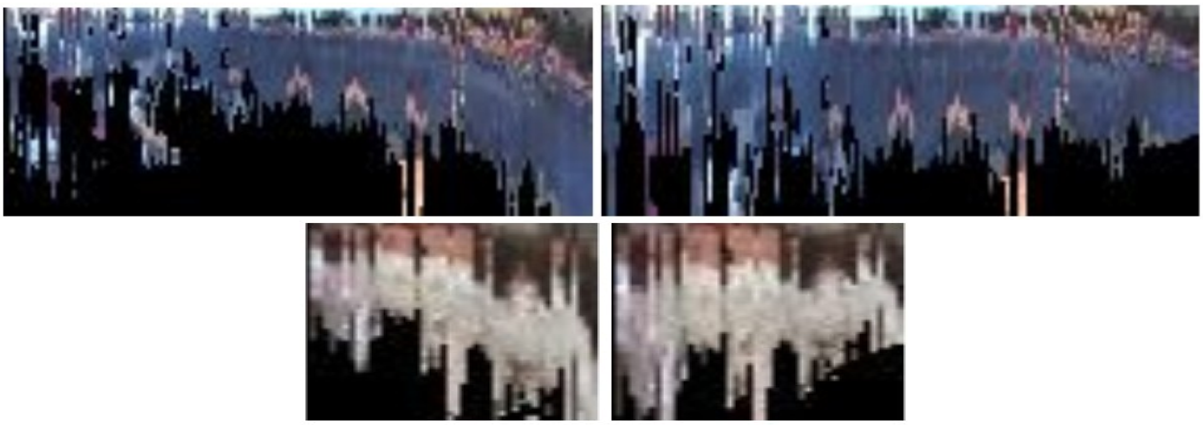


Fig. 2.5.2. Fitting a polynomial curve to the waist line allows us to correct the original bodyprint (left). Result on the right.

3. Detecting anomalies in videos

As surveillance becomes ubiquitous, the amount of data to be processed grows exponentially alongside with the demand for human intervention to interpret the data. A key goal of intelligent surveillance is to detect behaviors that can be considered anomalous. As a result, an extensive body of research in automated surveillance has been developed, often with the goal of automatic detection of anomalies. We clearly see some added value of such system in ACANTO, especially in managing security and safety. Regarding the benefits for the motion planner, we argue that to some extent, an anomaly is a concept that can be defined by the user. In other words, we can ask an anomaly detection system to detect whatever we tell it an abnormal situation looks like, as long as we have the means (*i.e.* the measurements) to dissociate the two situations. In this line of thoughts, we can describe crowded places as abnormal events, by looking at them as deviations from a situation of motion calm, which is what one would expect to normally see. In what follows, we describe a general recipe to automatically detect abnormal situation in videos, by combining appearance and motion cues in a deep learning framework.

Most of current approaches for automatic analysis of complex video scenes typically rely on hand-crafted appearance and motion features, which is clearly suboptimal, as it is more desirable to learn descriptors specific to the scenes of interest. To address this need, we developed a deep learning framework that automatically learns feature representations by combining appearance and motion from videos. We call this approach AMDN. In order to exploit the complementary information of both appearance and motion patterns, we introduce a novel double fusion framework, combining the benefits of traditional early fusion and late fusion strategies. Specifically, stacked denoising autoencoders (SDAEs) are used to separately learn both appearance and motion features as well as a joint representation (counting as early fusion). Then, based on the learned features, multiple one-class SVM models are used to predict the anomaly scores of each input. Finally, a late fusion strategy is employed to combine the computed scores and detect abnormal events. An overview of the whole system is painted in Fig. 3.1.

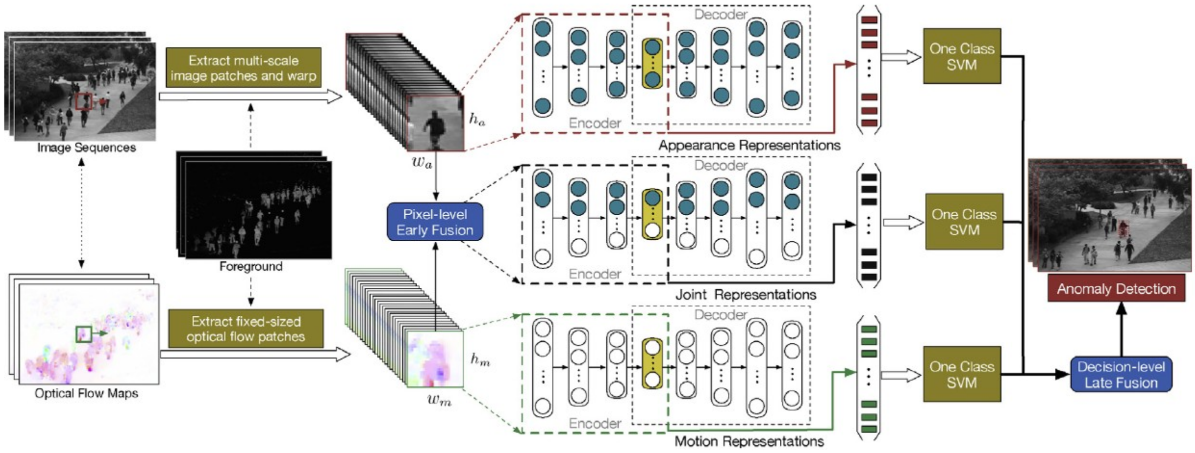


Fig. 3.1. Overview of the AMDN system: we use both appearance and motion information from videos to first train a set of stacked denoising autoencoders (SDAEs) in order to get a compact representation for every modality as well as for the fusion of the two. Consequently, we train one-class SVM models and perform late decision fusion for predicting anomalous behavior in videos.

Stacked Denoising Autoencoders

In the first stage of the system we use denoising autoencoders (DAEs) for extracting a compact and meaningful representation from the video input. Introduced in [50], the idea behind a DAE is to force the hidden layer to discover more robust features and prevent it from simply learning the identity function. This is achieved by training the autoencoder to reconstruct the input from a corrupted version of it. The structure of a DAE can be broken down into two parts: the encoder and the decoder, connected by a single compressed hidden layer. We use this hidden layer as feature representation of much larger data coming from video volumes. Training DAEs involves minimizing the average error between the reconstructed (corrupted) training samples and the original (uncorrupted) ones. Optimization is typically done by means of Stochastic Gradient Descent.

In order to extract mid-level appearance representations from the original image pixels, we apply a multi-scale sliding-window approach with a stride *d.* The resulting dense image patches are then warped into an equal size of $w_a \times h_a \times c_a$, where w_a, h_a are the width and height of each patch and c_a is the number of the channels ($c_a = 1$ for gray images). The warped patches are used for training. All the patches are linearly normalized into the range [0, 1]. We stack 4 encoding layers with $v_a \times w_a \times h_a \times c_a$ neurons in the first layer, where $v_a > 1$ is an amplification factor for constructing an over-complete set of filters. The use of over-complete representations in combination with sparsity terms has been shown to be effective in learning meaningful compressed representations in previous work [51], [52].

Moving to the motion features, we compute dense optical flow and we use a sliding window approach with windows of fixed size $w_m \times h_m \times c_m$ ($c_m = 2$ for optical flow magnitude along x and y axes) for motion representation learning. Similar to the appearance feature pipeline, the patches are normalized into [0,1] for each channel and 4 encoding layers are used. The number of neurons of the first layer is set to $v_m \times w_m \times h_m \times c_m$, where $v_m > 1$.

Finally, to account for the correlations between motion and appearance, we perform a pixel-level early fusion of the gray image patches and the corresponding optical flow patches. As mentioned earlier, each autoencoder is trained separately using SGD by minimizing the reconstruction loss regularized by a sparsity-inducing term.

SVM for abnormal event detection

We formulate our anomaly detection problem as a patch-based binary classification, *i.e.,* given a test frame, we adopt a sliding window approach and classify each patch as corresponding to a normal or an abnormal event. Specifically, given the t^{th} test patch, we compute the associated deep features representations $s_t^k, k \in \{A, M, J\}$. Then, we rely on three one-class SVM models to calculate a set of anomaly scores $A(s_t^k)$. Finally, the scores are linearly combined to obtain the global anomaly score:

$$S(s_t^k) = \sum_{k \in \{A,M,J\}} \alpha^k A(s_t^k) \tag{3.1}$$

We use SVMs with RBF kernel to compute the predictions of the three modalities. $A(s_t^k)$ Are the prediction scores at the output of the SVMs. Once computed, the predictions corresponding to the three modalities (for a given test patch) need to be combined in order to get the final anomaly score. Formally, we need to set the values of the vector $\alpha = [\alpha^A, \alpha^M, \alpha^J\ ]$ used in Eq. 1. There are many approaches in finding candidate fusion weights. Here we propose to solve the following optimization problem:

$$\min_{\mathbf{P}^k \in P, \alpha^k \geqslant 0} -\sum \alpha^k \mathbf{tr}\left(\mathbf{P}^k\mathbf{S}^k(\mathbf{P}^k\mathbf{S}^k)^T\right) + \lambda_s \|\alpha\|_p^p \tag{3.2}$$

where $P = \{\mathbf{P} : \mathbf{P}\mathbf{P}^T = \mathbf{I}\}$. Similarly to PCA, the matrix $\mathbf{P}^k \in \mathbb{R}^{m \times M}, m << M$ maps the samples $s_i^k \in R^M$ associated to the k^{th} modality into a new subspace in order to maximize the variance of the first m components, subject to orthogonality constraints. The matrix $\mathbf{P}^k\mathbf{S}^k(\mathbf{P}^k\mathbf{S}^k)^T$ represents the covariance of the k^{th} feature type in the new subspace and measures the spread of the projected samples for each modality. Setting the weights α^k by solving the optimization problem (2), we favor feature types associated with data sets with smaller variance: our intuition is that scattered data sets correspond to noisy features which must be deemphasized.

In the proposed optimization problem (2) we also introduce an l_p-norm term, which, compared with traditional l_2-norm and l_1-norm terms, guarantees an enhanced flexibility, by allowing to tune for p [53], [54]. Intuitively, l_1-norm imposes sparsity on the learned weights, while l_2 norms produces an "averaging" effect. Setting a priori one of the two may be suboptimal in term of performance. Moreover, the complexity of solving the problem (2) with l_p-norm is the same as for l_2-norm [55]. Therefore, in our experiments, we tune the parameter p in the interval $[1.1, ...2.5]$ with a step of 0.1. We also set the parameter $m = 100$, as it empirically provides the best performance. Equation (2) describes a convex optimization problem, whose solution can be found using an alternating minimization algorithm.

Datasets

The UCSD pedestrian dataset [56] includes two subsets: Ped1 and Ped2. The video sequences capture different crowded scenes and anomalies including the presence of bicycles, vehicles, skateboarders and wheelchairs. In some frames the anomalies occur at multiple locations. Ped1 has 34 training and 16 test image sequences with about 3400 anomalous and 5500 normal frames, and the image resolution is 238 × 158 pixels. Ped2 has 16 training and 12 test image sequences with about 1652 anomalous and 346 normal frames. The image resolution is 360 × 240 pixels.

The Subway dataset [57] was collected using CCTV cameras and consists of two video streams corresponding to two different subway station scenarios (an entrance and an exit gate). The length of the videos is 96 min and 43 min, respectively. In the entrance subset, there are 66 abnormal events include ing people moving in a wrong direction, unusual gesture interactions between people and sudden stopping or running. In the exit subset 19 abnormal events are included, such as people moving in a wrong direction and loitering near the exit gate. The image resolution is 512 × 384 pixels.

Experimental results

To improve the computational speed of our framework in the test phase, we use a foreground detection approach based on background subtraction. This is motivated by the fact that abnormal events are typically found in correspondence with moving pixels. An example depicting this process can be seen in Fig. 3.2. For an input test image, the probability map of the foreground pixels is estimated with a background subtraction algorithm and binarized. The foreground regions are detected by identifying the patches which contain more than a certain number of foreground pixels (10% in our experiments).

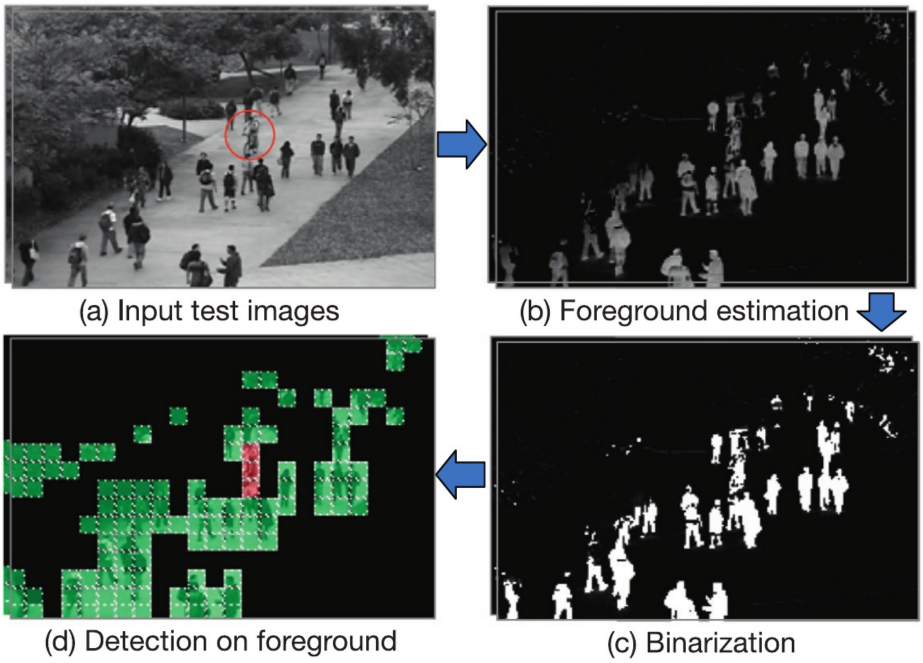


Fig. 3.2. Example of foreground detection preprocessing: original image (a) is subject to background subtraction (b) followed by a binarization step (c). We focus only on patches that contain at least a certain number of foreground pixels (10% in our experiments) - (d)

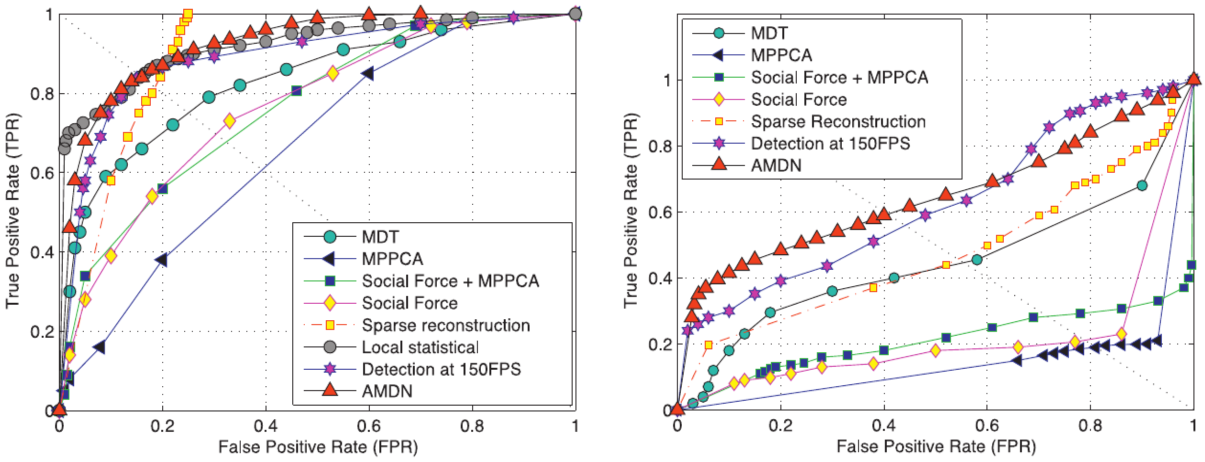
We first evaluate our approach on UCSD. UCSD is the dataset closest to our scenarios in ACANTO, since it contains recordings of crowded scenes, much like in shopping centers/museums. We use a sliding window at different scales (15×15, 18×18 and 20×20 pixels) to generate training patches, which are then resized at $w_a \times h_a = 15 \times 15$ pixels. From the motion space we explore the scene using only patches of $w_m \times h_m = 15 \times 15$ pixels. Concerning the SDAEs, we set the architecture of the encoders to $1024 \rightarrow 512 \rightarrow 256 \rightarrow 128$ neurons for appearance and motion modalities and $2048 \rightarrow 1024 \rightarrow 512 \rightarrow 256$ for the joint modality. Decoders have the architecture mirrored *w.r.t.* the encoders. We use samples corrupted with Gaussian noise to train the SDAEs, while the SVMs are trained using LibSVM.

To perform a quantitative evaluation, we use both a frame-level ground truth and a pixel-level ground truth on UCSD. The frame-level ground truth indicates whether one or more anomalies occur in a test frame. The pixel-level ground truth is used to assess the anomaly localization performance. If the detected anomaly region overlaps more than 40% with the annotated region, it is considered a true detection. We carry out a frame-level evaluation on both Ped1 and Ped2. Ped1 also provides 10 test image sequences with pixel-level ground truth. The pixel-level evaluation is performed on these sequences.

The proposed approach is compared with several state of the art methods. Specifically, we consider the Mixture of Probabilistic Principal Component Analyzers (MPPCA) approach in [58], the social force model in [59] and its extension in [56], the sparse reconstruction method in [60], mixture of dynamic texture (MDT) [56], Local Statistical Aggregates [61] and detection at 150 FPS [62]. Numerical results depicting AUC are shown in Tab. 3.1, while the associated ROC curves are plotted in Fig. 3.3.

Table 3.1: UCSD dataset: comparison (AUC) with state of the art methods

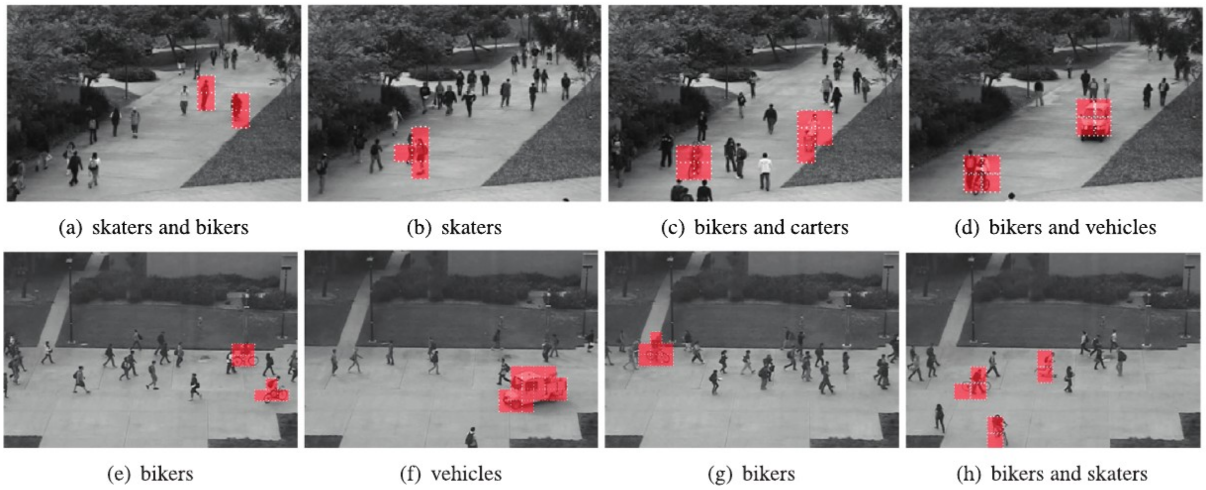
Model	Ped1 (frame)	Ped1 (pixel)	Ped2
MPPCA	59.0%	20.5%	69.3%
Social Force	67.5%	19.7%	55.6%
Social Force + MPPCA	66.8%	21.3%	61.3%
Sparse Reconstruction	-	45.3%	-
Mixture dynamic texture	81.8%	44.1%	82.9%
Local Statistical Aggregates	**92.7%**	-	-
Detection at 150 FPS	91.8%	63.8%	-
AMDM	92.1%	**67.2%**	**90.8%**



a) b)

Fig. 3.3. Receiver Operating Curves obtained on UCSD dataset – comparison between various approaches at frame level (a) and pixel level (b)

While on par with recent methods concerning frame-level anomaly detection in videos, we score better at localizing the anomalies (pixel-level), which is an important advantage in ACANTO, since we would like to understand not only if there's an anomaly in a certain video volume, but also understand where it is spatially localized. Figure 3.4 shows qualitative results on the same UCSD dataset.



(a) skaters and bikers (b) skaters (c) bikers and carters (d) bikers and vehicles

(e) bikers (f) vehicles (g) bikers (h) bikers and skaters

Fig. 3.4. Examples of anomaly detection results on Ped1 (top) and Ped2 (bottom) sequences.

On Subway dataset, we use the same preprocessing approach as for UCSD, except that now all the video frames are resized down to 320 × 240 pixels for computational efficiency and we no longer apply a multi-scale patch sampling, but rather keep the patch dimension fixed (to 15 × 15 pixels) for both appearance and motion cues. We follow the dataset evaluation protocol as in [41],[45] and compare with the following methods: Spatio-Temporal Composition (STC) [63], MPPCA [58], Spatio-Temporal Oriented Energy (STOE) [64], Dynamic Sparse Coding (DSC) [65], Sparse Reconstruction [60] and Local Optical Flow [57]. We report the number of abnormal events detected, as well as the number of false alarm cases in Tab. 3.2.

Table 3.2: Comparison of different methods on the Subway dataset

Model	Abnormal Events		False Alarms	
	entrance	exit	entrance	exit
STC	60/66	19/19	4	2
MPPCA	57/66	19/19	6	3
DSC	60/66	-	5	-
Sparse reconstruction	27/31	19/19	4	3
Local optical flow	27/31	19/19	4	3
AMDN	61/66	19/19	4	1

Figure 3.5 shows qualitative examples of using AMDN on Subway dataset. Detected abnormal events include: exiting through the entrance gate, entering through the exit gate and entering through the entrance gate without paying.

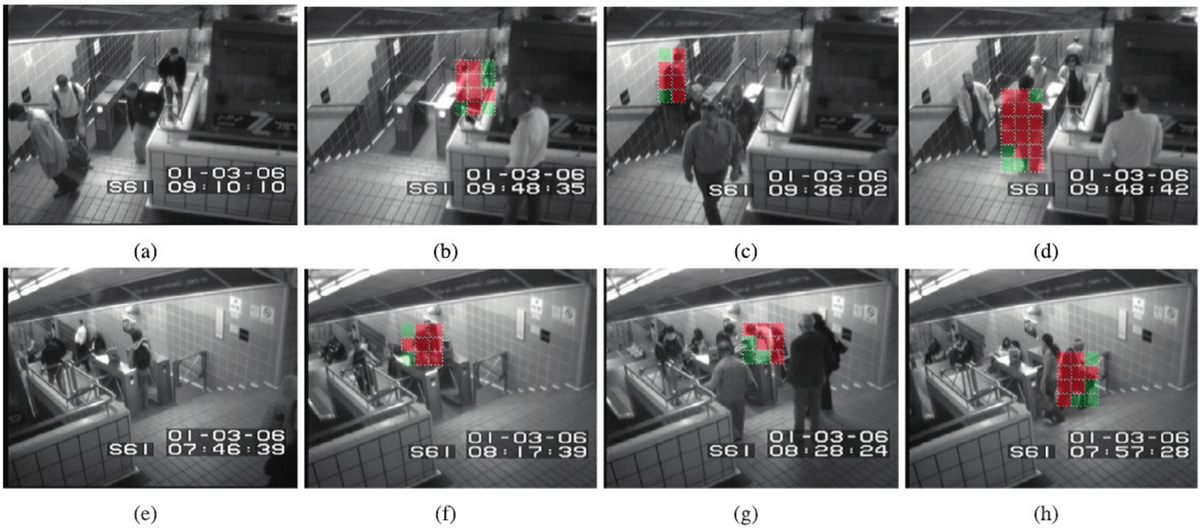


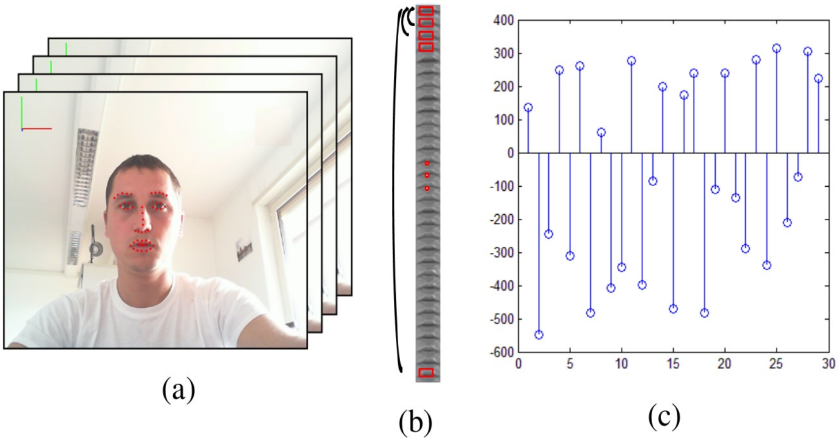
Fig. 3.5. Examples of anomaly detection results on the Subway exit (top) and entrance (bottom) datasets. The regions with abnormal events are marked with red color. (a) and (e) show examples of normal frames of the exit and entrance scenarios. The detected anomalies in the examples include: people entering through the exit gate shown in (b), (c) and (d); people entering without payment shown in (f) and (g); people exiting through the entrance gate shown in (h).

4. Detecting face-to-face interactions from the platform point-of-view

So far, in ACANTO we have developed approaches to model social interactions from a surveillance system's point of view (D3.5). We are able to detect when groups of people are engaged in social interactions by assuming an F1-formation, using a multi-modal head and body pose estimation framework. However, face-to-face interactions are more difficult to spot and model from a surveillance perspective. In order to address this task, we propose to shift the viewpoint from surveillance to on-board and install an additional RGB camera mounted on the same vertical pole as the one analyzing the user of the walker and facing ahead. Under this setup, one is able to track and analyze human faces “targeting” our walker subject, potentially engaged in a verbal communication. Inspired by previous research work [66], [67], the ingredients used for modeling social interactions in this scenario are the attention gaze (expressed in terms of head orientation) and the speech pattern. Since head orientation has been previously developed for modeling the user state, we focus here on developing a method for recognizing speech activity from a visual perspective.

4.1. Speech activity detection in face videos

The proposed solution models speech activity detection as a binary classification problem by analyzing consecutive frames inside a predefined temporal window. Given a video sequence containing speech activity (*e.g.* a potential face-to-face interaction) we assume there is a face model tracking the user from which mouth crops are extracted and concatenated. In this study we used the facial landmark detection system from [68] for simplicity and speed. Once the key-features belonging to the mouth region are localized within the model, the cropped images are scaled to the same size and prepared for feature extraction. We use generalized Haar-like features (*i.e.* differences of regions of pixels) exploiting the gray-scale information and sample each temporal window with 200 random regions. For each such rectangular region (spatially enclosed inside a mouth crop), we compute differences between all the crops inside the window and a reference one (the first). In this way we are able to model changes in appearance by looking at how differences evolve over time. From each distance signal (of the predefined 200) we compute statistical coefficients like mean, standard deviation and mean over the first order derivative, hence forming the final feature vector. The processing pipeline is outlined in Fig. 4.1.1.



(a)

(b)

(c)

Fig. 4.1.1. Processing pipeline for speech activity detection: (a) - initial video frames, (b) - extracted mouth crops with superimposed rectangular regions, (c) - computed distances

between consecutive frames and the first one. Statistical coefficients (e.g. mean, std, mean of first order derivative) are determined from each distance vector.

Speech Dataset
The training set was generated from the Map Task Corpus, an extension branch of the Self-Presentations Corpus [69], containing video recordings of 89 users engaged in spontaneous dialogues and interactions, elicited by finding the solution to a specific task. The videos were captured with a standard webcam at a resolution of 320 × 240 pixels and a frame rate of 15 fps. Each user recorded 4 video sequences each 5 minutes long in a frontal view setting. Few examples can be found in Fig. 4.1.2. We first process each video in a facial landmark detection stage in order to obtain the mouth crops on a frame-by-frame basis. Valid frames are then explored using a temporal window and grouped together to form individual samples. We set the window size experimentally to 15 frames, corresponding to 1 second of speech activity. Our training set added up over 130.000 labeled samples.

Fig. 4.1.2. Example frames from the Map Task Corpus [69]

Experimental results
We Random Forest classifier was trained following a 10-fold cross validation scheme, repeated 50 times. Each trial left aside one 10th of the whole data, randomly generated, while the rest was used for testing. The random partitions were generated in compact structures, *i.e.* continuous pieces of data, since uniform sampling transfers distribution properties between training and testing sets. For training the Random Forest, the following settings were used: the number of trees was set to 20, maximum depth fixed at 10, each tree from the forest was grown using 80% of the training-set as *in-bag* data, while the choice of the best binary test out of all 50 randomly generated for each split was made based on information gain. The mean and standard deviation of the accuracy obtained using these parameters is 83.2% ± 3.8%, which agrees with the findings of [70][71].

We compared the forest to another machine learning algorithm that relies on randomization to achieve good performance. Random Ferns were initially introduced by Ozuysal *et al.* in [72] as a competitive alternative to the Random Forests. Ferns are non-hierarchical structures, implementing the same binary test approach as the forests, but without any information gain based optimization. Instead, each fern applies all the tests at once and constructs a probability distribution for each possible output, for each class. The underlying assumption here is a semi-Naive Bayesian rule, according to which groups of features are considered to be independent, thus allowing to simply multiply probabilities between ferns in the final decision computation. Even if the assumption is not always valid, it has been proven [73] to work remarkably well in practice. As a consequence of building a probability matrix for each fern, the memory requirements grow linearly with the number of ferns and exponentially with the depth of the ferns, for a given number of classes. This is a serious limitation when forced to increase depth for increased discrimination power, but in practice there are ways to control/reduce the amount of working memory (*e.g.* working with single precision structures).

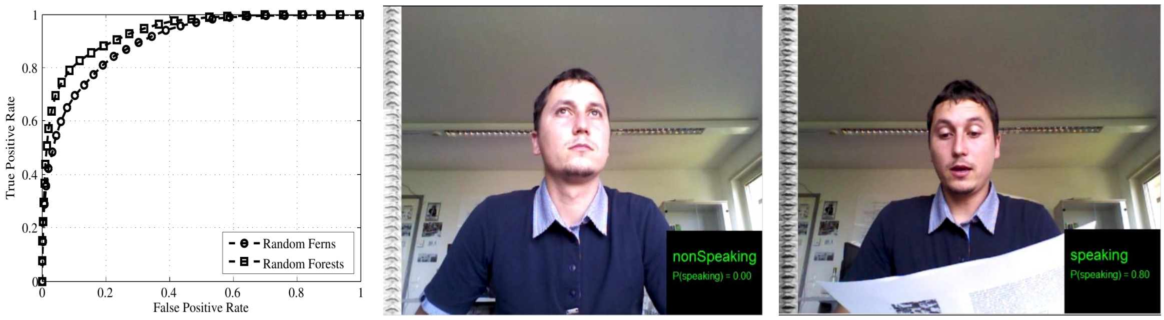


Fig. 4.1.3. ROC plots comparing Random Forests to Random Ferns at visual speech recognition (left). Examples of correctly classified non-speaking (middle) and speaking (right) sessions along with probability estimates. Note the mouth area crop stack at the left of each snapshot, corresponding to a temporal window of 1 second.

Although similar size Random Ferns were proven to yield similar performance values as the Random Forests on keypoint detection tasks [73], we discovered the ferns to be inferior for our speech activity detection. In fact, increasing the number of ferns to 200 and the depth to 12 still did not achieve the same detection rate as the Random Forest based approach. Figure 4.1.3. plots the ROC curve for this comparison, along with two examples of negative and positive speech activity, correctly classified.

4.2. Building a model for face-to-face interactions

In order to build a face-to-face interaction model, we take inspiration from recent work in human social interaction modeling [66], [67] and combine our measurements for focus of attention with those of speech activity. To this goal, we define a head pose activation flag, that sets itself to 1 every time a potential face is detected to “target” the user of the walker. In terms of head pose, this is evaluated as a logical check that the *yaw* and *tilt* are within a narrow range of angle values (we empirically set the thresholds of the angles to 15° in all directions). A similar threshold-based mechanism is used to define a speech flag, that switches to 1 every time speaking activity is detected by the RF classifier. For a full likelihood of a potential social interaction, we require that both speech and head pose flags are active at the same time. We apply temporal smoothing (in the order of a second) on all flags to avoid flickering behavior, which is a good trick in practice. This means that, even if the head pose is at the border of the 15° for some time (*i.e.* not entirely stable) , the head pose activation flag will experience at least a one second inertia window.

Figure 4.2.1. shows qualitative results of the system on Disney dataset [74], a highly challenging publicly available data collection showing social interactions captured under a first-person setting. The dataset contains recordings of 8 subjects wearing head-mounted cameras at a theme park, summing up to more than 42 hours of real world video. We focus on a very narrow subset of this corpus, namely the dialog windows, since we aim at capturing face-to-face interactions. Figure 4.2.1 captures an entire temporal window approximately 15 seconds long and depicts the contribution of the head pose (in the first part of the window) and later on the speech component to the overall likelihood of the face-to-face social interaction.

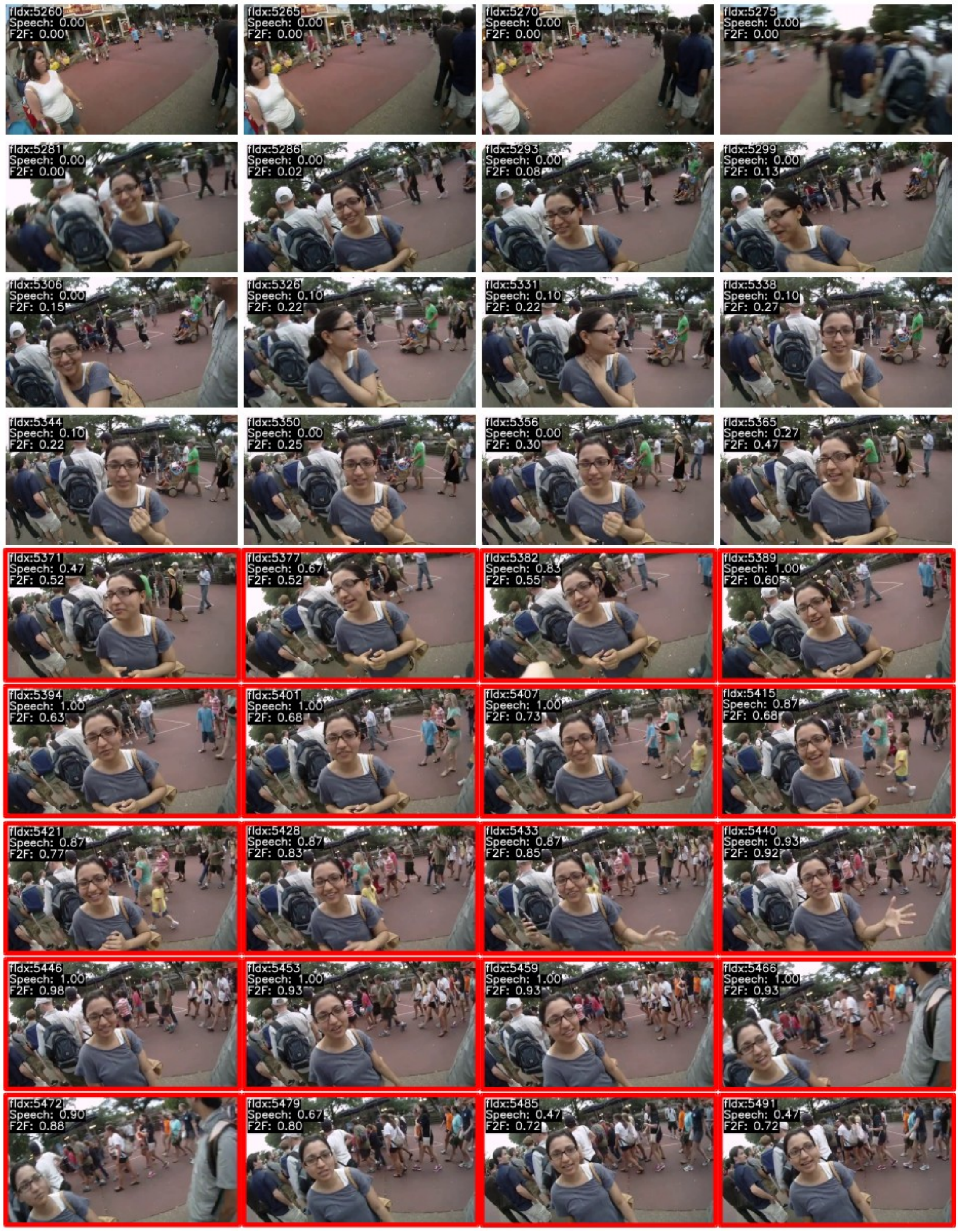


Fig. 4.2.1. Example frames capturing a temporal window from Disney dataset [74], showing the outcome of the face-to-face interaction detection component.

Note the speech output probability as well as the face-to-face likelihood (F2F) written on every frame. We color code the frames for which the probability of a face-to-face interaction exceeds 0.5. For a better understanding of how the system behaves, we invite the reader to access the full video[2] from which Fig. 4.2.1 was generated.

As with the other face analysis components (*e.g.* pain detection, emotional valence/arousal classification), we implemented the speech activity detector in C++, as a standalone service connecting directly to the webcam facing ahead of the walker (see Fig. 4.2.2 – the right-most stream), processing the frames grabbed from the video and sending the result (in the form of a probability measure encoding the likelihood of an undergoing face-to-face interaction) as a message on a dedicated channel using ZeroMQ. As in Fig. 4.2.2, we assume an outside existing client and a set of outside subscribers (above and below the gray dotted lines). The client will pass request (req) commands to the Launcher (implemented as a server), which will respond with reply (rep) messages to the client. The launcher himself controls a set of components (the black arrows), such as the camera frame publisher or the various frame processors (Sub 1, Sub 2, ...). The control commands are limited to start/stop signals and some parameters (such as internal communication ports). The subscribers Sub 1, Sub 2, … take the form of face analysis components (*e.g.* pain detection) and are all connected to the camera streaming publisher (the camera facing the user of the walker). The *f2f* module on the other hand is connected directly to the second camera (Cam2) facing ahead of the walker and implements the face-to-face component.

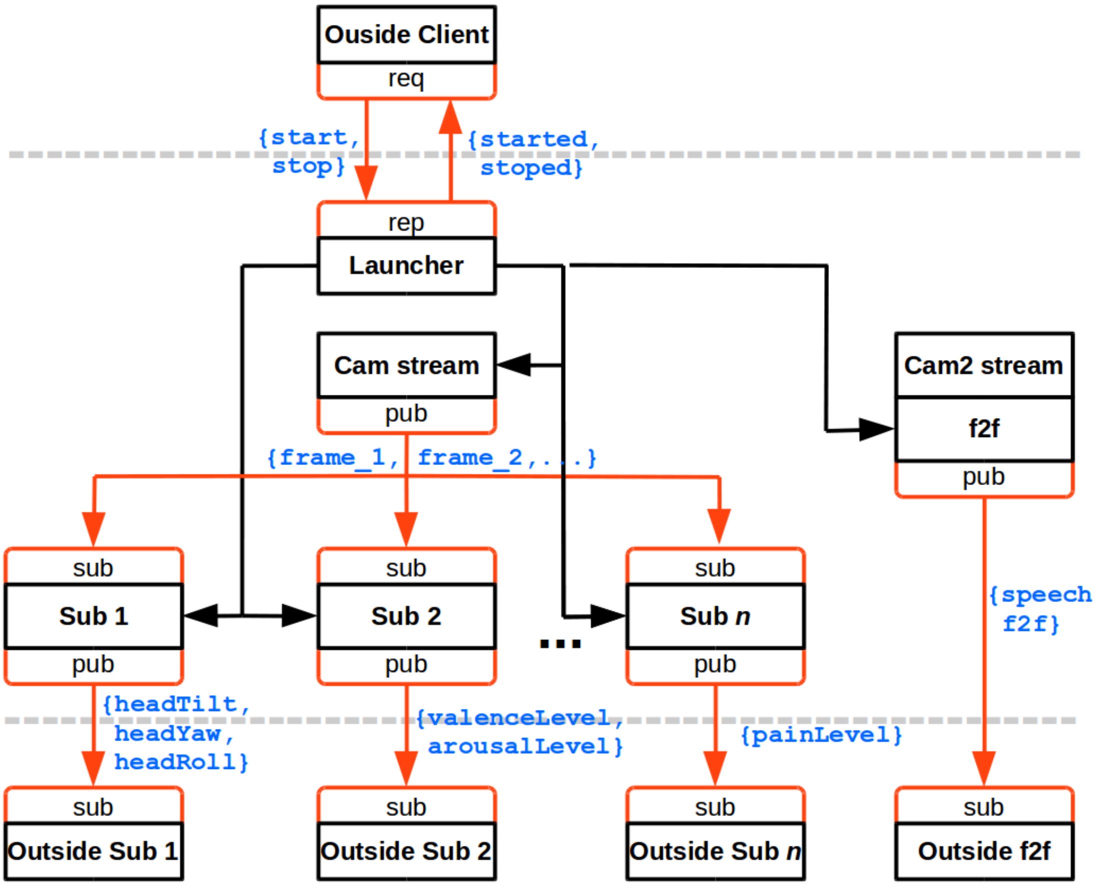


Fig. 4.2.2. Overall picture of the face analysis module in the context of external clients/subscribers. Blue keywords are message keys used to communicate between blocks.

[2] https://youtu.be/-u1EVBb3x48

5. Conclusions

In this deliverable we have addressed some of the main aspects in social context modeling, from both platform's point of view, as well as from the perspective of surveillance cameras. We have posed and answered three questions meant to strengthen security in surveillance settings, as well as provide important cues relating human social interactions.

In the platform world view, we have developed and implemented a solution that models face-to-face interactions, using close-range facial cues. The resulting probabilistic F2F model combines head pose information with speech activity provided by a dedicated component that processes visual crops of the mouth area inside a temporal window.

Security concerns have been addressed in a surveillance setting by developing approaches for person detection and tracking in complex, real world scenarios, as well as identifying trends and good practices for people re-identification. By leveraging pedestrian data coming from different modalities (color and thermal), our CMT-CNN model has shown state-of-the-art results in pedestrian detection using a cross-modal deep representation. In a similar fashion, by combining appearance and motion information from surveillance footage, we have developed a deep learning based approach to detecting abnormal behavior in videos. We show its effectiveness in several public benchmarks.

Overall, both D3.5 and D3.6 bring important contributions to describing the social context in ACANTO. Solutions to model various social aspects, ranging from anomaly detection to spotting different types of social interaction (F-formations, face-to-face interactions), have been presented. All the information provided by the developed components is meant to form the building blocks for a high level of social awareness, which in turn should provide context for the planning unit (WP5). In particular, we expect the reactive planner to benefit most from modeling the social context, as it is now possible to take into account the “social” nature of human motion and coordination, much needed to refine the SMC (statistical model checking) engine (WP5).

Finally, since some of the components described in this document are based on state-of-the-art deep learning modeling (which needs support from dedicated hardware units), we expect this deliverable to be of significant relevance to WP7, in charge of managing cloud resources.

References

[1] L. Zheng, Y. Yang, and A. G. Hauptmann, “Person Re-identification: Past, Present and Future,” vol. 14, no. 8, pp. 1–20, 2016.

[2] R. Benenson, M. Omran, J. Hosang, and B. Schiele, “Ten Years of Pedestrian Detection, What Have We Learned?,” in *ECCV*, 2014.

[3] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes (VOC) challenge,” *IJCV*, vol. 88, no. 2, pp. 303–338, 2010.

[4] M. Everingham, L. Van-Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes challenge 2012 (voc2012) Results.” 2012.

[5] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, “ImageNet Large Scale Visual Recognition Challenge,” *IJCV*, vol. 115, no. 3, pp. 211–252, 2015.

[6] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, “Microsoft COCO: Common Objects in Context,” in *ECCV*, 2014.

[7] J. Hosang, M. Omran, R. Benenson, and B. Schiele, “Taking a Deeper Look at Pedestrians,” in *CVPR*, 2015, pp. 4073–4082.

[8] Y. Tian, P. Luo, X. Wang, and X. Tang, “Pedestrian detection aided by deep learning semantic tasks,” in *CVPR*, 2015, vol. 07–12–June, pp. 5079–5087.

[9] J. Li, X. Liang, S. Shen, T. Xu, J. Feng, and S. Yan, “Scale-aware Fast R-CNN for Pedestrian Detection,” *arXiv*. 27-Oct-2015.

[10] L. Zhang, L. Lin, X. Liang, and K. He, “Is Faster R-CNN Doing Well for Pedestrian Detection?,” in *ECCV*, 2016, pp. 443–457.

[11] N. Dalal and B. Triggs, “Histograms of Oriented Gradients for Human Detection,” in *CVPR*, 2005, pp. 886–893.

[12] P. Dollar, C. Wojek, B. Schiele, and P. Perona, “Pedestrian Detection: An Evaluation of the State of the Art,” *PAMI*, vol. 34, no. 4, pp. 743–761, 2012.

[13] P. Dollár, C. Wojek, B. Schiele, and P. Perona, “Pedestrian Detection: A Benchmark,” in *CVPR*, 2009, pp. 304–311.

[14] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Fu, and A. C. Berg, “SSD : Single Shot MultiBox Detector,” *arXiv*. 2016.

[15] S. Zhang, R. Benenson, and B. Schiele, “Filtered channel features for pedestrian detection,” in *CVPR*, 2015, pp. 1751–1760.

[16] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele, “How Far are We from Solving Pedestrian Detection?,” in *CVPR*, 2016, pp. 1259–1267.

[17] A. Angelova, A. Krizhevsky, V. Vanhoucke, A. Ogale, and D. Ferguson, “Real-Time Pedestrian Detection With Deep Network Cascades,” in *BMVC*, 2015, pp. 1–12.

[18] S. Hwang, J. Park, N. Kim, Y. Choi, and I. S. Kweon, “Multispectral Pedestrian Detection: Benchmark Dataset and Baseline,” in *CVPR*, 2015, pp. 1037–1045.

[19] C. Premebida, J. Carreira, J. Batista, and U. Nunes, “Pedestrian Detection Combining RGB and Dense LIDAR Data,” in *IEEE RSJ*, 2014, pp. 4112–4117.

[20] G. Wang and Q. Liu, “Far-Infrared Based Pedestrian Detection for Driver-Assistance Systems Based on Candidate Filters, Gradient-Based Feature and Multi-Frame Approval Matching,” *Sensors*, vol. 15, no. 12, pp. 32188–32212, Dec. 2015.

[21] S. Gupta, J. Hoffman, and J. Malik, “Cross Modal Distillation for Supervision Transfer,” in *CVPR*, 2016, pp. 2827–2836.

[22] J. Hoffman, S. Gupta, and T. Darrell, “Learning with Side Information through Modality Hallucination,” in *CVPR*, 2016, pp. 826–834.

[23] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv*. 2015.

[24] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, “Caffe: Convolutional architecture for fast feature embedding,” *arXiv*. 2014.

[25] P. Dolí, R. Appel, S. Belongie, and P. Perona, “Fast Feature Pyramids for Object Detection,” *PAMI*, vol. 36, no. 8, pp. 1532–1545, 2015.

[26] P. Viola and M. J. Jones, “Robust Real-Time Face Detection,” *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.

[27] W. Nam, P. Dollár, and J. Hee Han, “Local Decorrelation for Improved Pedestrian Detection,” in *NIPS*, 2014, pp. 424–432.

[28] S. Paisitkriangkrai, C. Shen, and A. van den Hengel, “Pedestrian Detection with Spatially Pooled Features and Structured Ensemble Learning,” *arXiv*. 18-Sep-2014.

[29] Y. Yang, Z. Wang, and F. Wu, “Exploring Prior Knowledge for Pedestrian Detection,” in *BMVC*, 2015, pp. 1–12.

[30] B. Yang, J. Yan, Z. Lei, and S. Z. Li, “Convolutional Channel Features,” in *ICCV*, 2015, pp. 82–90.

[31] Y. Tian, P. Luo, X. Wang, and X. Tang, “Deep Learning Strong Parts for Pedestrian Detection,” in *ICCV*, 2015, pp. 1904–1912.

[32] Z. Cai, M. Saberian, and N. Vasconcelos, “Learning Complexity-Aware Cascades for Deep Pedestrian Detection,” in *ICCV*, 2015, pp. 3361–3369.

[33] K. Chen, S. Gong, T. Xiang, and C. Change Loy, “Cumulative Attribute Space for Age and Crowd Density Estimation,” in *CVPR*, 2013, pp. 2467–2474.

[34] D. Yi, Z. Lei, S. Liao, and S. Z. Li, “Deep Metric Learning for Person Re-identification,” in *ICPR*, 2014, pp. 34–39.

[35] W. Li, R. Zhao, T. Xiao, and X. Wang, “Deepreid: Deep filter pairing neural network for person re-identification,” in *CVPR*, 2014, pp. 152–159.

[36] D. Gray, S. Brennan, and H. Tao, “Evaluating appearance models for recognition, reacquisition, and tracking,” *PETS*, vol. 3, no. 5, pp. 1–7, 2007.

[37] W. Zheng, S. Gong, and T. Xiang, “Associating Groups of People,” in *BMVC*, 2009, p. 23.1-23.11.

[38] P. M. Roth, M. Hirzer, M. Köstinger, C. Beleznai, and H. Bischof, “Mahalanobis Distance Learning for Person Re-identification,” in *ACVPR*, 2014, pp. 247–267.

[39] W. Li, R. Zhao, and X. Wang, “Human Reidentification with Transferred Metric Learning,” in *ACCV*, 2012, pp. 31–44.

[40] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, “Scalable Person Re-Identification : A Benchmark,” in *ICCV*, 2015, pp. 1–7.

[41] K. H. Dangwei Li, Xiaotang Chen, Zhang Zhang, “Learning Deep Context-Aware Features Over Body and Latent Parts for Person Re-Identification,” in *CVPR*, 2017, pp. 384–393.

[42] X. T. Haiyu Zhao, Maoqing Tian, Shuyang Sun, Jing Shao, Junjie Yan, Shuai Yi, Xiaogang Wang, “Spindle Net: Person Re-Identification With Human Body Region Guided Feature Decomposition and Fusion,” *CVPR*, vol. 1, pp. 1077–1085, 2017.

[43] J. Z. Ji Lin, Liangliang Ren, Jiwen Lu, Jianjiang Feng, “Consistent-Aware Deep Learning for Person Re-Identification in a Camera Network,” in *CVPR*, 2017, pp. 5771–5780.

[44] W. Chen, X. Chen, J. Zhang, and K. Huang, “Beyond triplet loss: a deep quadruplet network for person re-identification,” in *CVPR*, 2017, pp. 403–412.

[45] N. Z. Sanping Zhou, Jinjun Wang, Jiayun Wang, Yihong Gong, “Point to Set Similarity Based Deep Feature Learning for Person Re-Identification,” *CVPR*, pp. 3741–3750, 2017.

[46] S. Bai, X. Bai, and Q. Tian, “Scalable Person Re-identification on Supervised Smoothed Manifold,” in *CVPR*, 2017, pp. 2530–2539.

[47] L. S. Jiaxin Chen, Yunhong Wang, Jie Qin, Li Liu, “Fast Person Re-Identification via Cross-Camera Semantic Binary Transformation,” in *CVPR*, 2017, pp. 3873–3882.

[48] P. C. Slawomir BÄ…k, “One-Shot Metric Learning for Person Re-Identification,” in *CVPR*, 2017, pp. 2990–2999.

[49] Z. Zhong, L. Zheng, D. Cao, and S. Li, “Re-ranking Person Re-identification with k -reciprocal Encoding,” in *CVPR*, 2017, pp. 1318–1327.

[50] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, “Extracting and Composing Robust Features with Denoising Autoencoders - LISA - Publications - Aigaion 2.0,” in *ICML*, 2008, pp. 1096--1103.

[51] M. A. Ranzato, C. Poultney, S. Chopra, and Y. Lecun, “Efficient Learning of Sparse Representations with an Energy-Based Model,” in *NIPS*, 2007, pp. 1137–1144.

[52] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Mangazol, “Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion,” *JMLR*, vol. 11, pp. 3371–3408, 2010.

[53] P. Jawanpuria, M. Varma, and S. Nath, “On p-norm Path Following in Multiple Kernel Learning for Non-linear Feature Selection,” *PMLR*, pp. 118–126, Jan. 2014.

[54] F. Orabona, L. Jie, and B. Caputo, “Multi Kernel Learning with Online-Batch Optimization *,” *JMLR*, vol. 13, pp. 227–253, 2012.

[55] F. Nie, H. Wang, H. Huang, and C. Ding, “Joint Schatten p-norm and lp-norm robust matrix completion for missing value recovery,” *Knowl Inf Syst*, vol. 42, pp. 525–544, 2015.

[56] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, “Anomaly Detection in Crowded Scenes,” in *CVPR*, 2010, pp. 1975–1981.

[57] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz, “Robust Real-Time Unusual Event Detection using Multiple Fixed-Location Monitors,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 3, pp. 555–560, Mar. 2008.

[58] J. Kim and K. Grauman, “Observe Locally, Infer Globally: a Space-Time MRF for Detecting Abnormal Activities with Incremental Updates,” in *CVPR*, 2009, pp. 2921–2928.

[59] R. Mehran, A. Oyama, and M. Shah, “Abnormal Crowd Behavior Detection using Social Force Model,” in *CVPR*, 2009, pp. 935–942.

[60] Y. Cong, J. Yuan, and J. Liu, “Sparse Reconstruction Cost for Abnormal Event Detection,” in *CVPR*, 2011, pp. 3449–3456.

[61] V. Saligrama and Z. Chen, “Video Anomaly Detection Based on Local Statistical Aggregates *,” in *CVPR*, 2012, pp. 2112–2119.

[62] C. Lu, J. Shi, and J. Jia, “Abnormal Event Detection at 150 FPS in MATLAB,” in *ICCV*, 2013, pp. 2720–2727.

[63] M. J. Roshtkhari and M. D. Levine, “Online Dominant and Anomalous Behavior Detection in Videos,” in *CVPR*, 2013, pp. 2611–2618.

[64] A. Zaharescu and R. Wildes, “Anomalous Behaviour Detection Using Spatiotemporal Oriented Energies, Subset Inclusion Histogram Comparison and Event-Driven Processing,” in *ECCV*, 2010, pp. 563–576.

[65] B. Zhou, L. Fei-Fei, E. P. Xing, and B. Zhao, “Online Detection of Unusual Events in Videos via Dynamic Sparse Coding,” in *CVPR*, 2011, pp. 3313–3320.

[66] D. A. Salter, A. Tamrakar, B. Siddiquie, M. R. Amer, A. DIvakaran, B. Lande, and D. Mehri, “The Tower Game Dataset: A multimodal dataset for analyzing social interaction predicates,” *ACII*, pp. 656–662, 2015.

[67] S. Grover, M. R. Amer, B. Siddiquie, A. Tamrakar, D. A. Salter, B. Lande, D. Mehri, and A. Divakaran, “Human Social Interaction Modeling Using Temporal Deep Networks,” *arXiv*. 2015.

[68] T. Baltrušaitis, P. Robinson, and L. P. Morency, “OpenFace: An open source facial behavior analysis toolkit,” in *WACV*, 2016, pp. 1–10.

[69] L. Batrinca, N. Mana, B. Lepri, and F. Pianesi, “Please, tell me about yourself: automatic personality assessment using short self-presentations,” in *ICMI*, 2011, pp. 255–262.

[70] H. Yehia, P. Rubin, and E. Vatikiotis-Bateson, “Quantitative association of vocal-tract and facial behavior,” *Speech Commun.*, vol. 26, no. 1–2, pp. 23–43, Oct. 1998.

[71] L. Wang, X. Wang, and J. Xu, “Lip Detection and Tracking Using Variance Based Haar-Like Features and Kalman filter,” in *FCST*, 2010, pp. 608–612.

[72] M. Özuysal, P. Fua, and V. Lepetit, “Fast keypoint recognition in ten lines of code,” in *CVPR*, 2007, pp. 1–8.

[73] M. Ozuysal, M. Calonder, V. Lepetit, and P. Fua, “Fast Keypoint Recognition Using Random Ferns,” *PAMI*, vol. 32, no. 3, pp. 448–461, 2010.

[74] A. Fathi, J. K. Hodgins, and J. M. Rehg, “Social interactions: A first-person perspective,” in *CVPR*, 2012, pp. 1226–1233.