People Detection from Overhead Cameras

A study of impact of occlusion on performance

by

Lu Liu

in partial fulfillment of the requirements for the degree of

Master of Science

at the Delft University of Technology,
to be defended publicly on Friday August 31, 2018 at 01:00 PM.

Student number:	4621832	
Thesis committee:	Dr. Hayley Hung (supervisor)	EEMCS
	Laura Cabrera-Quiros (mentor)	EEMCS
	Prof. Marcel Reinders,	EEMCS
	Dr. Julian Kooij,	3ME

Acknowledgments

I would like to thank my supervisor Dr. Hayley Hung first for her patient and instruction on this work. I am very grateful to her encouragement when I feel uncertain and dispirited in the beginning of my thesis. It is so lucky to work on this project under the supervision of her. I have the opportunity to know what is research question, what a hypothesis is like, and how to start a research. This knowledge is helpful not only for this project but also for my career. My thesis is an independent work, and she gave me space to think of every step independently rather than just a goal. I can feel that she is care about my own thoughts and my critical thinking.

Then I would like to thank Laura, my mentor. She is so helpful when I have problem about the detail in this research. She gave examples and told me her research experience. This let me feel more confident and patient on this work. Also she supports me on the writing of this thesis, and helps me out of my unclear expression. I know she is at the last stage of her PhD now. It is so kind of her to proofread my report.

Also I want to say thank you to Stephanie, the PhD student who cooperate with me on the head annotation collection. Since I will continue doing research as a PhD student in University of Twente, this is a nice and helpful experience to work with a PhD student.

Finally, I have to thank my family and my sweet friends. They gave me accompany and support when I was working on this thesis. Without them, I can not get rid of anxiety and disturbance at the difficult stage.

Lu Liu
Delft, August 2018

Contents

List of Tables

List of Figures

1 Introduction

People are always the most important object in the research of social scenes, and the interactions between people is one of the most interesting and potentially useful challenges for modern engineering [4]. Therefore, detection of people is a significant topic of research, and machine vision is always the material for research.

The aim of detecting people can be summarized as: *given an image or video sequence, localize all people.* Typically, researchers use rectangular bounding boxes to show the location of people. During the last decades, people detection has received great attention in computer vision and pattern recognition because of its various applications. Though there are thousands of papers that provide approaches for people detection, most of them focus on datasets from side view. According to related surveys[4][1], popular datasets available publicly are separated by application including image retrieval, video surveillance, and driving assistance. And most of these datasets have no data from the top view, but this is a common view for indoor surveillance. Most of people detectors from side view fail on detecting people from the top view. The possible reason could be the self-occlusion and changed body shape of people. In this viewpoint, the camera can not capture every body part of a person, especially the lower part.

This thesis is based on **MatchNMingle**[5], a multi-sensor dataset for the analysis of social interactions and group dynamics in-the-wild during free-standing conversations and speed dates. The MatchNMingle dataset have annotated video for people in crowded scenes from the top view. The deep network, **Overfeat-GoogLeNet**[6] is selected as the training system for this research due to its effectiveness and the similarity between its training data and MatchNMingle. It is a end-to-end people detection approach aiming at crowded scene.

In the first chapter, I explain the research objective, motivate its importance, and summarize contributions briefly. In the end of this chapter, I give the structure of this report.

1.1. Background

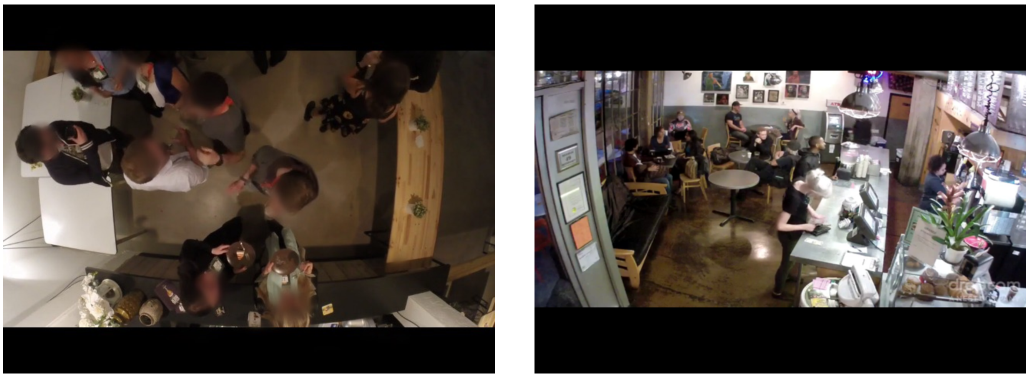
MatchNMingle dataset

As mentioned above, this study is based on **MatchNMingle**[5]. This dataset was created to contribute to help analyzing the challenges of social signals and natural interactions. There are two portions in this event, which are speed dates and mingle. In this research, we only focus on the mingle part, since only its annotations is completed. There was multiple forms of social interactions, and one of the common forms are free-standing conversational groups, which are small groups of two or more conversing people are emerging. This kind of spatial formations vary due to the purpose and desires of each participant in the group. These unstable social situations contain abundant information but are challenging. MatchNMingle contains about two hours of uninterrupted recordings for 92 participants, and cases of conversations in the free-standing groups and sitting dyads.

Multiple sensors were collected in this dataset, including video, audio, personality surveys, frontal pictures, speed-date responses, etc. during 3 real speed date events, each followed by a mingle party. One of the most important data is the video for the entire event as well as manual annotations for social actions and position for 30 minutes at 20fps for each day. Since in the social formations, position data is one of the important factors for analyzing form, merging, and dissolution of a formation, in our research, we are interested in this videos and the position information of every participant.

The video of event area was captured using 9 different GoPro Hero 3+ cameras from top-view. The resolution of video is 1920 x 1080 (16:9), and the sample rate is 30 fps. In this portion, participants were fenced in a rectangular space created by tables. There are some overlap between the 5 cameras that recorded the mingle for this area. In this situation, there is different illuminations, shadows, occlusions and a crowded environment in the video, which causes difficulties for analysis. Because of financial limitations, not all cameras were annotated. A 30-minute segment of the mingle party was selected randomly for each day. To maximize the people density and number of social actions in the whole scene, the cameras with the highest concentration of people were selected for annotation(both position and social actions). The manual annotations of position and social actions was made using the Vatic tool proposed by Vondrick et al[7]. This tool was designed to annotated crowd-sourcing in Amazon‘s Mechanical Turk (MTurk).

The 30-minute segment was divided into smaller 2-minute tasks(or HITs). An example frame from MatchNMingle videos is shown in Figure 1.1(a) and a body-annotated frame from Vatic tool is shown in Figure 1.2(a).



(a) A frame from MatchNMingle videos[5] (b) A frame from Brainwash videos [6]

Figure 1.1: Frames from MatchNMingle and Brainwash

There is great interest for using computer vision and machine learning algorithms on people detection from overhead cameras in crowded social scenes. So far there are hundreds of algorithms and systems for people detection. However, for the situation in MatchNMingle, there is almost no system aiming to solve or analysis people detection from the overhead cameras. One of the challenges is the occlusion in this viewpoint. Since the cameras are located on the top of the ceiling, people in the cameras are out of common human shape and with heavy occlusion. Due to occlusion, there are multiple algorithms aim to improve the detection precision for occluded people. Thus we are interested in the influence caused by occlusion in datasets.

Overfeat-GoogLeNet Network

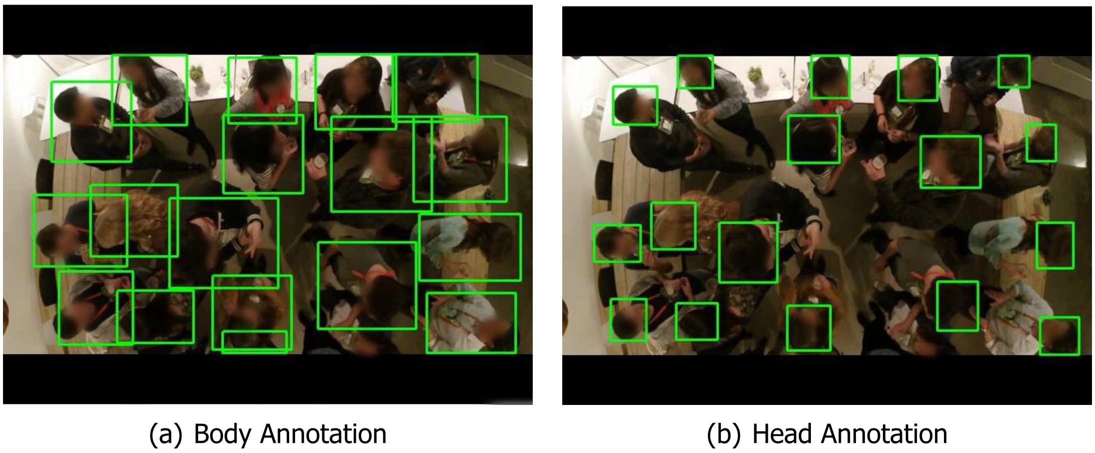
To analyze this, based on literature research, I selected a deep network called **Overfeat-GoogLeNet** from Russell Stewart et. al[6] as the research system. This system takes an image as input and directly generates a set of people bounding boxes as output. Different from original Overfeat model, as the name of the systems tells, the image representations of the system are trained with GoogLeNet. There are three reasons for experiment on this network:

- Compared to other detectors, this network shows effectiveness on the detecting people in crowded scenes especially for people with occlusion.
- Stewart tested their models on the Brainwash dataset, which consists of 11917 images with 91146 labeled people. In Brainwash, the viewpoint of cameras is not exactly the same as our dataset as shown in Figure 1.1(b). But it is similar, as the camera is located at the corner of ceiling. This is also helpful when comparing the performance of detectors trained with Brainwash and MatchNMingle.
- The implementation of this network is completed and publicly available in Github.

The detail explanation of this network is illustrated in Chapter 2 and Chapter 3.

1.2. Research Objective

The main goal of this work is to study the impact of occlusion on performance of detectors. Based on the study, we aim to improve the performance.


(a) Body Annotation

(b) Head Annotation

Figure 1.2: Example frame of body and head annotations

The common-used annotation for people is body bounding box(eg. Figure 1.2(a)). But when the viewpoint is not proper to view the whole body of people or it is too crowded to see the whole body, the body annotation can lead to unexpected overlapping bounding boxes. Several datasets avoid the body annotation rather using head annotation(eg. Figure 1.2(b)). For example, in the Brainwash collected by Stewart[6], the lower part of the body are always invisible and head bounding boxes generates less overlapping boxes compared to body annotation. In our dataset, similar to Brainwash, the lower body are always occluded. It is clear that with different annotation, the occlusion level varies. This can make extremely difference on performance of detectors. Thus, it is possible that training with head annotated data can promote the performance of detectors.

However, collection of proper head annotation for such a big dataset takes both time and money. Thus it is meaningful to improve the performance of body detectors. Based on our experiments results, we are trying to give a instruction on how to modify a training dataset and lead to a better performance.

1.3. Research Questions and Hypotheses

There are three research questions in my study related to occlusion level and annotation type separately. According to these research questions, I have four hypotheses. Below the research questions, hypotheses, and experiments are listed.

For people with different occlusion level, the visible parts and ratio of visibility for body varies. Referring to previous evaluation in Stewart's experiments [6], their system can detect some people with occlusion. However, there is not much information about when the system can not detect people with occlusion and how the occlusion level in training data influence the performance. Thus I have the first research question:

- **Research Question 1: How does the network learn with different occlusion levels in training data?**
  - Hypothesis 1: Detectors make more mistakes with increase of occlusion level.
  - Experiment 1: Train detectors and analyze their performance on people with different occlusion level.
  - Hypothesis 2: Detectors trained with subsets with low occlusion levels make more mistakes for people with high occlusion levels.
  - Experiment 2: Train detectors with high-occlusion level and low-occlusion level, and analyze their performance on people with different occlusion.

  Second research question is an attempt at head annotation applying in our dataset.

- **Research Question 2: How does the trained model change with head annotation?**
  - Hypothesis 3: Detectors trained with head annotation perform better than detector trained with body annotation.
  - Experiment 3: Train detectors with head annotation and body annotation separately and analyze their performance.

  With the analysis in Experiment 1 and Experiment 2, we can select more appropriate training data for people detection from overhead cameras, which may inspire other researchers on frame selection.

- **Research Question 3: How can we improve the performance of body detectors using the existing data?**
  - Hypothesis 4: Body detectors trained with less images with low occlusion level perform better than the initial detectors in Experiment 1.

  - Experiment 4: Randomly reduce half of the images with occlusion level lower than the average image occlusion level in training data, and train detectors again. Compare the performance of newly trained detectors and initial detector.

All the results from experiments are analyzed by the evaluation measures, average precision, count error, equal error rate and F1 measure. According to these experiments, in this study we discuss the relation between model performance and occlusion level as well as annotation type.

1.4. Contribution

There are three main contributions in this work.

- We study the impact of occlusion levels in the detection performance.
- We show the effectiveness of head annotation.
- A training data selection strategy on body-annotated data: randomly reduce half of the images with occlusion level lower than the average image occlusion level. Experiments shows that detectors trained with this strategy predicted less bounding boxes that are not people.

1.5. Thesis Organization

In this chapter, we introduce the dataset MatchNMingle which my research is relied on, the research objective and the contributions including my research questions and hypotheses. The other chapters of this thesis is structured according to my work-flow as follows:

- Chapter 2 reviews the formulation of people detection, including its application, popular publicly available datasets, general framework. Based on the framework, this chapter also gives some methods for feature representation and popular classifiers. Because of the powerful performance of deep network, we introduce several deep model as well as occlusion handling methods.
- Chapter 3 explains the structure and performance of Stewart‘s detector(the GoogleNet-Overfeat Network) as well as theory of every parts. Then I will give the evaluation methods including the definition and significance of the evaluation parameters (recall, average precision, recall-1-precision curve, F1 measure, equal error rate etc.) used in this research.
- Chapter 4 covers the details of every experiments. Firstly it introduces the quantitative analysis for datasets for frame number, people density, bounding box occlusion level, etc. The explanation and analysis of experiments setup and experiments results from occlusion level and annotation

type experiments will be given according to the research questions and hypotheses.

- Chapter 5 concludes this work, and recommends the future work.

2 Related Work

People detection has been arguably addressed as a special topic beyond general object detection. In decades, various methods were proposed based on different features. In the early research of people detection, low-level features and handcrafted features are extracted from people samples as the training features for classifiers. In recent years, compared with detectors trained with handcrafted features, deep learning detectors have shown better performance for people detection.

In this chapter, I searched on people detection surveys, methods, and related datasets. From these material, I summarized the general architecture of people detector. And based on the architecture, I talked about detectors based on the handcrafted features and detectors based on deep models. Since our research model is based on GoogLeNet and Overfeat architecture, these two designs are illustrated in this part. Apart from these, to deal with people detection from top view with the MatchNMingle dataset, I investigated some methods to handle occlusion of people.

2.1. Brief Introduction to People Detection

Assume we have some images or video sequence, our goal is to develop a system that can detect all the instances that are people and return their location. People detectors typically returns a list of rectangular bounding boxes. There are thousands of papers that provides and analyze approaches for people detection. Due to different applications, various datasets are publicly available.

Common applications includes image retrieval, video surveillance, and driving assistance. The definition of these application and their corresponding common-used datasets are listed below.

- **Image retrieval** is to searching and retrieving images from a large digital picture dataset [8].

  - Datasets: MIT[9], INRIA[10], PASCAL VOC[11]

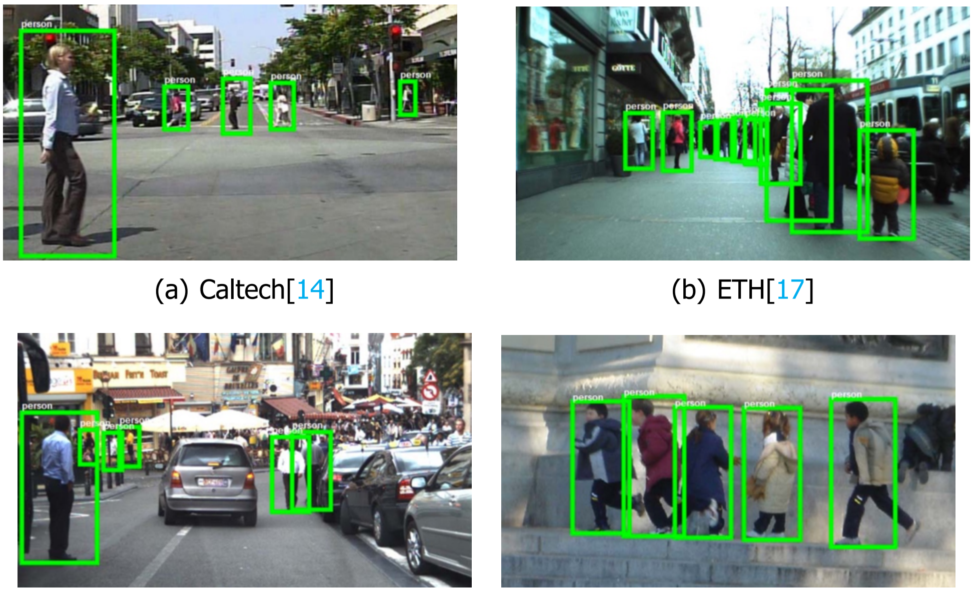
- For **video surveillance**, systems always analyze the images from video surveillance cameras in order to recognize human. These images are always in restricted areas within the camera's view like a fenced off area, and a parking lot.

  - Datasets: USC-B[12], CAVIAR[13]

- For **driving assistance**, cameras are always equipped on vehicle to obtain images on road. So different from video surveillance, systems are to analyze the images from public place especially streets.

  - Datasets: Caltech[14], TUD[15], CVC[16], ETH[17]

Some example images from these datasets are shown in Figure 2.1. It is clear that these datasets are always from a side view of people, which we can recognize every body parts of a person except occluded people.



(a) Caltech[14]

(b) ETH[17]

(c) TUD[15]

(d) INRIA[10]

Figure 2.1: Example images (cropped) and annotations from four pedestrian detection datasets

2.2. General Framework

Nguyen and his colleagues[1] gives a general framework for human detection as shown in Figure 2.2. It has the following sequential steps: extracting candidate regions that are potentially covered by human objects, describing the extracted regions, classifying/verifying the regions as human or non-human, and post-processing. Even for deep network like Overfeat-GoogLeNet, this framework is commonly applied.

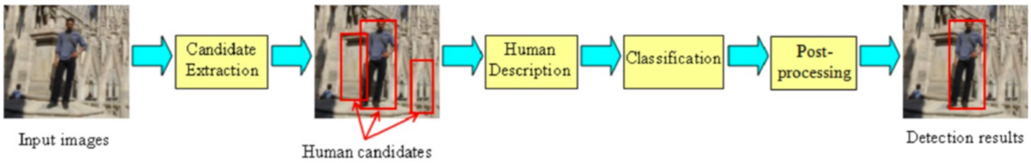


Figure 2.2: A general human detection framework.[1] Note that preprocessing, e.g. image filtering, may be applied before candidate extraction to enhance the quality of the input image/video sequence.

The common and ordinary approach for candidate extraction is to assume every person can be enclosed by a detection "window". Without prior knowledge of size and location of people, windows with various scales and positions are extracted from the images. Merging of some nearby windows is necessary to obtain the final candidate human objects. The common-used method for merging multiple windows is the non-maximal suppression (NMS). NMS can be positively formulated as local maximum search, where a local maximum is greater than all its neighbours (excluding itself)[18]. Then NMS can help to suppress all the other values (by setting them to 0) except the local maxima. When the input data is video, the most common technique, background subtraction can extract human candidates.

2.3. Feature Representation

To construct human descriptors, typically, selected features are organized in a structure. They are expected to describe human objects with various poses from various viewpoints. Features can be generated through various methods. Hand-crafted features are developed from low-level information (e.g. edge, colour, and motion)[1]. Thus, they can be classified as *shape features*, *appearance features*, *motion features*, and a *combination*.

2.3.1. Shape features

Shape features are to describe the shape of people, edge-based features. The location, orientation and magnitude of edge pixels are always considered for extracting edge-based features from edge maps or gradient images. The well-known example of shape features is histogram of oriented gradients (HOG), which counts occurrence of gradient orientation in localized portions of an image, proposed by Dalal and Triggs[10].

2.3.2. Appearance features

Appearance features are mainly to capture the colour or texture information from local image regions. An example of simple appearance features is image intensity[19]. Also, there are some commonly used appearance features like Haar feature and local binary pattern (LBP)[20] that was originally proposed for texture classification. Haar-like features that are based on Haar wavelets consider adjacent rectangular regions at a specific location in a detection window, sums up the pixel intensities in each region and calculates the difference between these sums. And LBP labels the pixels of an image by thresholding the neighbourhood of each pixel and considers the result as binary number.

2.3.3. Motion features

Motion information can be used to discriminate one object from another if they have different motion patterns. Thus, it is efficient for describing objects, especially for non-rigid objects such as people who always perform cyclic movements. Motion features are always extracted from temporal difference[21] or optical flows[22]. Temporal difference (TD) refers to the change or differences of future values of a given signal. Generally, it is used in predictions over successive time steps. And optical flow or optic flow is the pattern of apparent motion of objects, surfaces, and the edges in a visual scene caused by the relative motion between an observer and a scene[23]. In Viola's work[21], rectangular features are calculated on the difference images between consecutive frames to encode the temporal difference in the motion of people. Similarly, Nguyen[24] applied the NR-LBP that calculated on the difference images as the motion feature. In Dalal's work[22], histogram of flows is computed in a similar manner with the HOG using optical flows. It can be used to describe the boundary motion as well as internal motion.

2.3.4. Features extracted through deep network

Recently, Deep Neural Networks are also applied in people detection. They use different models from the discriminative classifiers based on handcrafted features. For example, in Ouyang's work[25], sub tasks of human detection such as features selection, object description, occlusion handling are organized into different layers of a deep convolutional neural network and the parameters are jointly learned through the network. But due to the structure of deep network, the feature extraction and classification are always associated together. More information about deep learning for people detection is in Section 2.5.

Compared with early work, there are several algorithms that have respectively good performance on the publicly available datasets mentioned above, but there are still lack of research on the challenge difficulty, people occlusion, on people detection. Generally, they can be classified into two groups by features extracted from raw images. The first group contains classifiers based on handcrafted features, while the other group is classifiers based on features extracted by deep model.

2.4. State-of-the-art Classifiers

The structure that human descriptors extracted from handcrafted features followed by a discriminative classifier like support vector machine (SVM) or AdaBoost is the most commonly used model before 2014[4][1]. During research, I found that the most advanced approaches for people detection are mostly based on deep model. However, there are still several methods based on handcrafted features that have convincing performance and contributions on publicly available benchmarks.

In machine learning, SVM is a supervised learning model with associated learning algorithms that analyze data used for classification and regression analysis. A joint person detector based on deformable part models (DPM)[26] is proposed in [27] aiming at detecting occluded people. It combines both single and double-person detectors into a single model that is jointly trained. The single and double-person detectors are represented as different components of the DPM. The performance of the joint detector strongly depends on its ability to distinguish between single and double-person hypotheses. In their related work[28], they address this issue by reformulating joint detector using structural SVM framework and modifying the loss function to penalize detection of single people and double-person components and vice versa.

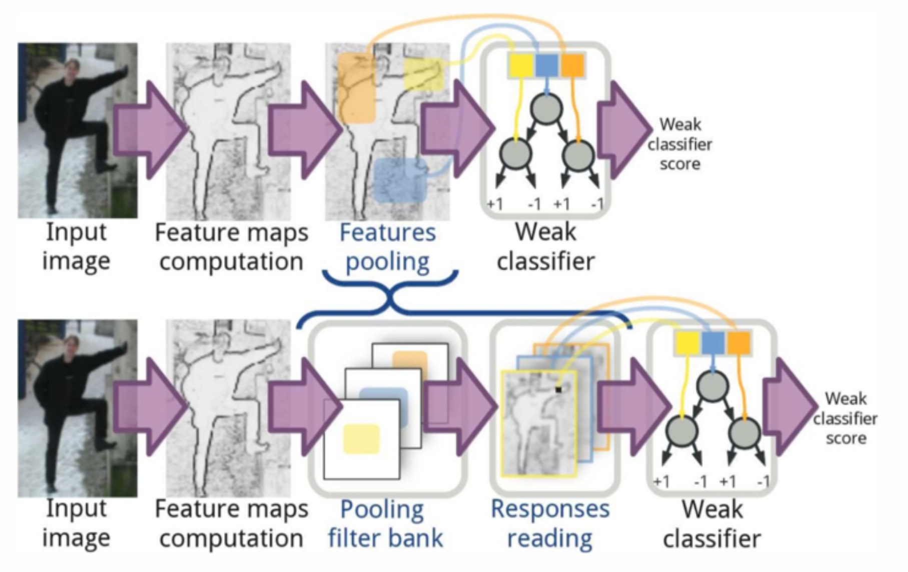
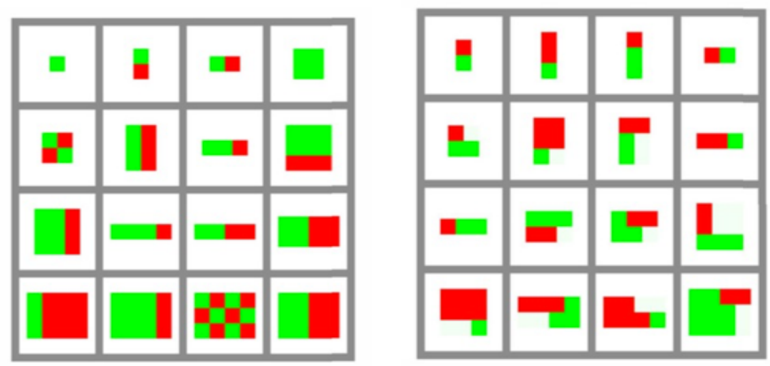


Figure 2.3: Filtered feature channels illustration, for a single weak classifier reading over a single feature channel [2]

Another well-performing model based on only low-level features is provided in [2]. In this paper, authors summarized the general architecture of their model as Figure 2.3. They applied an intermediate layer between low-level feature maps and the classifier, which is a linear transformation implemented as convolution with a filter bank. Their experiments show that, with the proper filter bank, filtered channel features reach top detection quality before 2015. The top method in the research used solely HOG+LUV features and Checkerboards+SDt as filter bank. Checkerboards is a naive set of filters that covers the same sizes (in number of cells) as InformedFilters[29] and for each size defines: a uniform square, all horizontal and vertical gradient detectors, and all possible checkerboard patterns as shown in Figure 2.4. SDt refers to the difference of frames from weakly stabilized video[30], and is used as an add-on for optical flow information of 2 channels (not filtered).



(a) Checkerboards Filter (b) InformedFilters

Figure 2.4: Illustration of filter banks. red, white, green indicate {-1, 0, +1}

2.5. Deep Learning for People Detection

In recent years, deep learning has been applied to pedestrian detection and achieved promising results[31][32][33][34][35][36][37][38]. Compared with using handcrafted features, it can automatically learn features in an unsupervised or supervised fashion. More advanced than the early deep models, some novel detectors like switchable deep network (SDN) [38] can learn hierarchical representations with semantic meanings (such as the body parts of head-shoulder, upper-body, and lower-body). However, Deep Neural Network (DNN) models are known to be very slow, especially when used as sliding-window classifiers. Some methods like DeepCascade [33] focus on the trade-off between accuracy and speed. And some methods (e.g. [39][35]) tried to combine handcrafted features and deep model to optimise people detection.

For most of the deep models, *pre-training* is used. Pre-training has been demonstrated in an object detection method, R-CNN[40]. With it, the weights are initialized from the weights of a network that has been trained on ImageNet. ImageNet is an image database organized according to the WordNet (a large lexical database of English) hierarchy in which each node of the hierarchy is depicted by hundreds and thousands of images. [40] shows that fine-tuning a pre-trained Convolutional Neural Network (CNN) on ImageNet classification task on object detection and segmentation data can significantly improve the performance. There are three popular deep models and three pre-training strategies as below[32]. Three deep models are AlexNet [41], Clarifai[42], and GoogLeNet[3](deeper than the first two models), which are the best-performing models of the ImageNet classification challenge in the past several years. Apart from these models, there are some other models like VGG-16[43], and CifarNet[34]. Three pre-training strategies exist: (1) no pre-training, (2) pre-training the deep models by using the ImageNet training data with image-level annotations of 1000 classes, and (3) pre-training the deep models by using ImageNet training data with object-level annotations. All the models mentioned above are firstly applied on object classification, while later some researchers employed them on people detection. For example, [32][34] applied AlexNet in their detectors; [6] applied GoogLeNet in their models; and [39] used VGG-16 in their method.

Another way to find an approach for people detection is to search on object detection methods. Zhang and colleges[36] got inspired from a general object detection method, Faster R-CNN[44]. Faster R-CNN consists of two components: a fully convolutional Region Proposal Network (RPN) for extracting candidate regions, followed by a downstream Fast R-CNN[43] classifier.

Different from hybrid methods (e.g. [34]) that combines traditional handcrafted features and deep convolution features, the Faster R-CNN system is a purely CNN-based method without using handcrafted features. [36] illustrated two reasons for the unsatisfactory accuracy: (1) insufficient resolution of feature maps for handling small instances, and (2) lack of any bootstrapping strategy for mining hard negative examples. To overcome these weakness, they proposed a detector using a RPN followed by boosted forests (BF) shared, high-resolution convolutional feature maps.

Since our research is based on Overfeat-GoogLeNet deep neural network, the information about Overfeat and GoogLeNet is given below.

2.5.1. GoogLeNet

One of the popular deep concolutional neural network architecture, GoogLeNet, is designed for the classification and detection in the ImageNet Large-Scale Visual Recognition Challenge 2014(ILSVRC14)[3]. It is a 22 layer deep network based on the Hebbian principle and the intuition of multi-scale processing. Increasing the size of deep neural network can straightly improve its performance. In this situation, the depth(the number of levels) and the width(the number of units at each level) increase. To train models with higher quality safely, a large amount of labeled training data is necessary. However, this solution can lead to two main drawbacks. Bigger network with more parameters can make the network larger, but tends to overfit, especially when the training data is limited. Apart from this, the use of computational resources is increased sharply with bigger network.

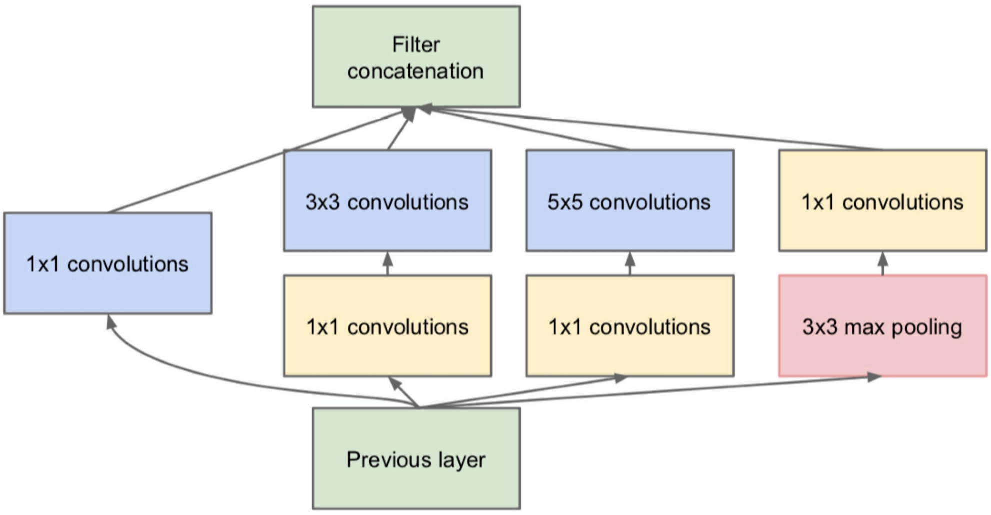
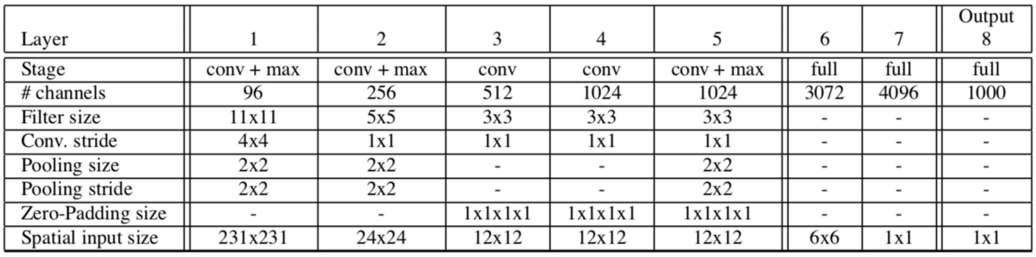


Figure 2.5: Inception module[3]

To overcome these challenges, the researchers proposed the Inception architecture. Inception module with dimension reductions is shown in Figure 2.5. Inception networks are composed of inception modules stacked on each other, with occasional max-pooling layers. With this architecture, the computational complexity can be in control when the number of unites at each layer increasing dramatically. In addition, this design aligns with the intuition that visual information should be processed at various scales and then aggregated so that the next stage can extract features from different scales simultaneously.

With inception architecture, the computation for bigger network with increasing width and depth is possible and without getting into computational difficulties. GoogLeNet refers to the incarnation of the Inception architecture. The schematic view of GoogLeNet is depicted in Figure A.1.

2.5.2. OverFeat



Layer	1	2	3	4	5	6	7	Output 8
Stage	conv + max	conv + max	conv	conv	conv + max	full	full	full
# channels	96	256	512	1024	1024	3072	4096	1000
Filter size	11x11	5x5	3x3	3x3	3x3	-	-	-
Conv. stride	4x4	1x1	1x1	1x1	1x1	-	-	-
Pooling size	2x2	2x2	-	-	2x2	-	-	-
Pooling stride	2x2	2x2	-	-	2x2	-	-	-
Zero-Padding size	-	-	1x1x1x1	1x1x1x1	1x1x1x1	-	-	-
Spatial input size	231x231	24x24	12x12	12x12	12x12	6x6	1x1	1x1

Figure 2.6: Architecture specifics for OverFeat model

OverFeat is an integrated framework for using convolutional networks for classification, localization and detection [45]. In this model, predicted bounding boxes are accumulated for object localization and detection. By combining multiple localization predictions, it is possible to avoid training on background samples for detection. This means the time-consuming and complicated bootstrapping training passes can be avoided. Thus the network focus only on positive classes.

The architecture sizes are in Figure 2.6. The spatial size of the feature maps depends on the input image size, which varies during inference step. In this model, the entire image is explored by densely running the network at each location and at multiple scales. To circumvent that the network window aligned the object improperly when sliding window approach applied in the case of ConvNets, resolution augmentation and last subsampling operation is performed.

This network can be used as a classification-trained network. And it changes to a localization-trained network when the classifier layers are replaced to a regression network. The feature extraction is initially trained with the classification task. The network generates many bounding box predictions first. Then the individual predictions are combined through a merge strategy applying to the regressor bounding boxes. Training the network in a spatial manner and the model is for detection. Note that it is necessary to predict a background class when no object is present in the image.

2.6. Occlusion Handling

One of the most difficult challenges in people detection is occlusion[1]. In general, occlusion refers to as a phenomenon in which an object of interest is not fully visible. In practice, occlusion can be categorized into two types: (1) inter-object occlusion: a human object is blocked by another object called occlude, and (2) intra-object occlusion/self-occlusion: a human object may not be fully observed due to his/her pose and/or the camera's viewpoint. In my project, we are mainly faced with the second type occlusion as shown in Figure 1.1(a). However, existing approaches mainly focus on the first type of occlusion.

The current occlusion handling methods can be categorized as detection-based occlusion handling or inference-based occlusion handling. Detection-based approaches determine the occlusion of people by using only the information of that object and its parts (e.g. detection scores, geometric information of parts). As for inference-based approaches, the occlusion is inferred based on the mutual relationship between that object and other objects. For example, in [27][28], a double-person detector was applied together with the typical single-person detector to detect two-person patterns for inter-object occlusion.

Methods (e.g.[32][31][6]) based on DNN shows some improvements in occlusion handling. Using this model, the classification of body parts is verified via different combinations of that part with other parts. [32] proposed DeepParts that is robust against occlusion. They constructed a part pool where different complementary parts can be automatically selected in data driven manner. The selected parts can be adopted to different scenarios or different datasets. And from their research, a single part detector can achieve convincing performance. The positive samples are computed from the visible map of each ground truth box and extracted from the corresponding region only if the part template is fully covered by the visible map. Meanwhile, the negative samples are extracted from corresponding regions within negative proposals.

3

Research Methodology

In Chapter 2, the architecture of GoogLeNet and Overfeat is described. The original version of Overfeat provided by [45] relied on image representation trained with AlexNet[41]. In GoogLeNet-Overfeat network, the GoogLeNet architecture is substituted into the Overfeat model. To describe the training process in our research, the implementation details of this network are given below in Section 3.1 as well as the information of training and initialization.

Since this work is to study the impact of occlusion, we give the definition of occlusion levels in Section 3.2.

With the trained models and detection results, reasonable evaluation metrics are critical for performance analysis. Different metrics reveal a model from various aspects. Thus, to have a coherent understanding of detectors, multiple evaluation measures are needed. Besides common performance evaluation, occlusion level is one of the study target. Thus, we selected commonly-used evaluation measures used in related work and went one step further to analyze these measures with occlusion level. Below these selected measures are listed with definition with and example in Section 3.3.

3.1. GoogleNet-Overfeat Network

3.1.1. Architecture

The architecture of the system is shown in Figure 3.1. The model encode an image into a 15x20 grid of 1024-dimensional top level GoogLeNet features[6]. The 1024 dimensional vector extracted the contents of the region and contains rich information referring to the position of the objects. Each cell in the grid is of size 139x139 receptively. They are trained to generate the list of bounding boxes intersecting the central 64x64 region. The size of 64x64 was considered as a large enough region to capture people with local occlusion in the images.

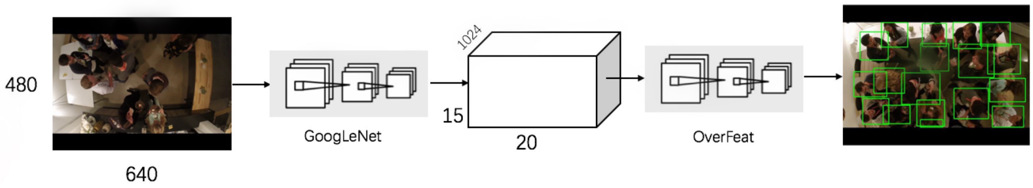


Figure 3.1: GoogLeNet-Overfeat first encodes an image into a block of high level features. Overfeat acts as a controller, decoding this information into a set of detection.

This region size is chosen according to the occlusion interactions in the dataset. Thus it can be changed larger or smaller if necessary.

3.1.2. Implementation

Initially, the models in Stewart et. al[6] are trained and evaluated using the Caffee open source deep learning framework[46]. Their code(named TensorBox[47]) is available for Tensorflow[48], hence our models are also trained and evaluated with the deep learning framework Tensorflow. The models are trained with learning rate $\epsilon = 0.2$, and every 100000 iterations, the learning rate is decreased by multiplying 0.8[6]. GoogLeNet weights are initialized with the weights parameter pre-trained on ImageNet[49]. According to Stewart‵s experiments, fine-tuning of GoogLeNet features is critical to meet the demands of the decoder, and it influence the precision.

3.1.3. Model Training

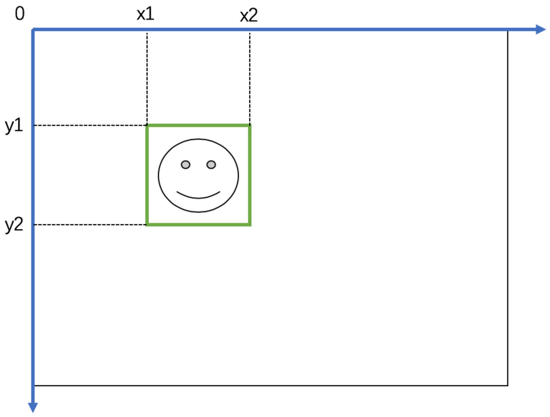


Figure 3.2: Example of bounding box description

Training requires a "json" file containing a list of images and the bounding boxes in each image. There are several data annotation formats allowed for TensorBox. The most simple format, json-file, is recommended as the same format used in Stewart's researches. Each annotation is an object with two properties: `image_path`(string) and `rects`(list). `rects` contains information of all bounding boxes which present on the current image. The format of bounding box description consists of four integer properties which mean the main diagonal of the rectangle $(x1, y1) - (x2, y2)$ as shown in Figure 3.2. TensorBox reading procedure expects that $x1 < x2$ and $y1 < y2$.

3.2. Occlusion Levels

To study how occlusion level influence the performance of trained models, there are two types of occlusion levels defined here, which are average image occlusion level, and bounding box occlusion level.

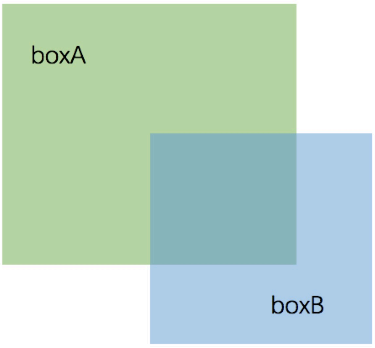


Figure 3.3: Example of overlapped bounding boxes

Bounding box occlusion level of a bounding box A refers to the sum of overlap ratio between all the other bounding boxes in the same image. To calculate this, we use the definition of overlap ratio(OR) between two bounding boxes(see Figure 3.3), which is calculated with the formula below:

$$OR(boxA, boxB) = \frac{A \cap B}{A \cup B}$$

Thus the bounding box occlusion level(BBOL) of box A can be calculated:

$$BBOL(boxA) = \sum_{boxI} OR(boxA, boxI) \tag{3.1}$$

$boxI$ refers to all the other bounding boxes in the same images as $boxA$.

And the average image occlusion level(AIOL) is defined as the average bounding box occlusion level in a certain image.

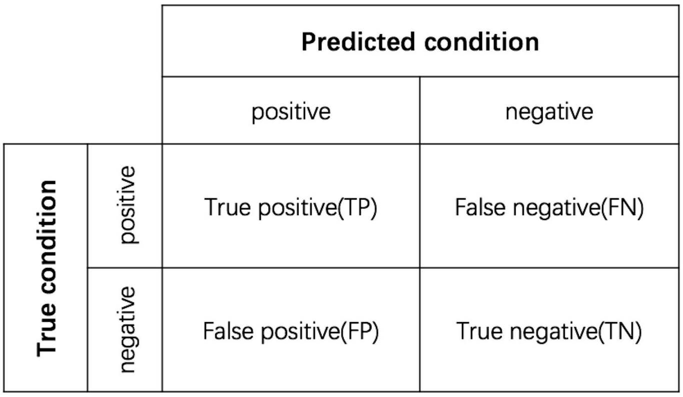
$$AIOL(imageA) = \frac{\sum_{boxK} BBOL(boxK)}{numberofBB(imageA)} \tag{3.2}$$

Due to the fact that the Overfeat-GoogLeNet network takes an image as a input, we consider the average image occlusion level as a quality of training data. In fact, the bounding box occlusion level is more related to the occlusion situation of every person. Thus, when analyzing performance of people with different occlusion level, we consider the bounding box occlusion level.

3.3. Evaluation Methods

There are many measures used for evaluation in pattern recognition, information retrieval and binary classification. And much effort and research has gone into solving the problem of evaluation of people detectors. The most commonly used measures are Recall, Precision, and F-measure. These measures are biased sometimes and should not be used without understanding of the biases. To evaluate the models comprehensively, using multiple measures is more reliable. To make it easy to compare our results with the initial model by [6], I considered recall, average precision, recall-1-precision curve, equal error rate, and counting ability. Apart from these measures, F1-Measure is applied to measure the effectiveness with respect to both recall and precision.

Below, the measures used in this research are explained in detail both in definition and understanding with example. To understand the metrics for evaluation, let us define an experiment. The terms *true positive*, *true negative*, *false positive*, and *false negative* compare the performance under test with trusted ground truth. The terms *positive* and *negative* refer to the detector‘s prediction, and the *true* and *false* refer to the ground truth corresponding to the observation. The outcomes can be formulated in a $2x2$ table as in Figure 3.4.



		Predicted condition	
		positive	negative
True condition	positive	True positive(TP)	False negative(FN)
	negative	False positive(FP)	True negative(TN)

Figure 3.4: Contingency table for experiments

To understand the terms in this table, we give their explanation in our research.

- **True positive** is the items that are people in all the detected items.

- **False negative** is the items that are people in the ground truth but not detected by the detector.
- **False positive** is the detected items that are not people.
- **True negative** should be the non-human items that are not detected by the detector. Since there is no information about non-human items in our research, it is not considered.

3.3.1. Recall

Recall is defined as:

$$Recall = \frac{TP}{TP + FN} \tag{3.3}$$

Recall is the ratio of the correct predictions and the total number of correct items in the set. It is the percentage of the total correct items correctly predicted by the model. In the problem of object detection, it also referred to as the true positive rate or sensitivity. Thus it indicates how good is the model detect the correct people.

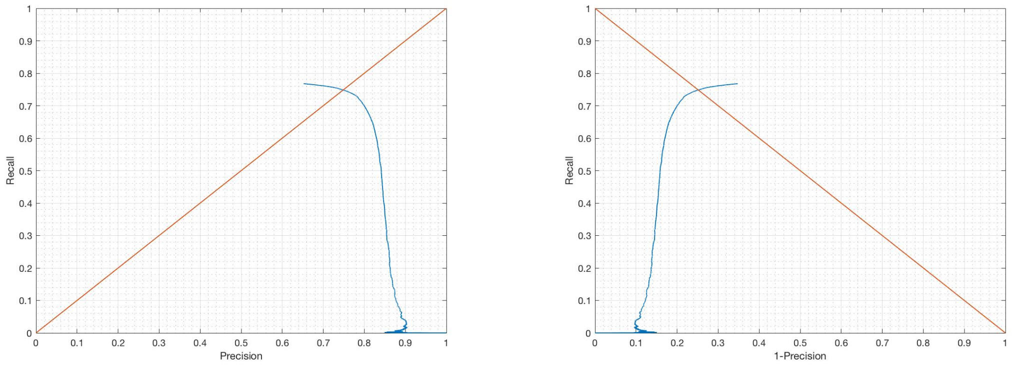
3.3.2. Precision

Precision is defined as:

$$Precision = \frac{TP}{TP + FP} \tag{3.4}$$

Precision is the ratio between the correct predictions and the total predictions, in other word, it is the positive predictive value. Thus it indicates how much confidence does the detector have for all the detected objects.

3.3.3. 1-Precision-Recall Curve



(a) An example of Precision-Recall Curve

(b) An example of 1-Precision-Recall Curve

Figure 3.5: Example frame of body and head annotations

Precision-Recall curve is a useful measure of success of prediction. Since in Stewart's paper[6] they use 1-Precision-Recall curve as one of the evaluation metrics, in our research, we also use this curve(example in Figure 3.5(b)). This curve is a horizontal flip version of Precision-Recall curve(eg. in Figure 3.5(a)) , and it shows the tradeoff between precision and recall for different threshold. A large area under the curve represents both high recall and high precision. High values for both precision and recall indicate that the detector returns good prediction, as well as a majority of all positive items. Note that there is no linear relationship between precision and recall, and precision may not decrease with recall.

A model with high recall but low precision returns more predicted items most of which are predicted incorrectly when compared with the ground truth. In the opposite, a model with high precision but low recall returns less predicted items, but most of them are predicted correctly. Thus an ideal detector with both high precision and high recall should return more predicted people that labeled correctly.

3.3.4. False Positive Ratio

False positive ratio also known as false discovery rate(FDR), is the probability of falsely detecting people for tests. It is calculated as the ratio between the number of negative items wrongly categorized as positive and the total number of all items categorized as positive.

$$FPR = \frac{FP}{TP + FP} \tag{3.5}$$

It tell the ratio of wrong predictions in people detection.

3.3.5. COUNT

COUNT represents he count error by computing the average absolute difference between the number of predicted and ground truth detection in test set images.

$$COUNT = \frac{abs((TP + FP) - (TP + FN))}{TP + FN} \tag{3.6}$$

A smaller COUNT indicates that the number of predicted people is more close to the number of people in ground truth.

3.3.6. Average Precision

Average precision(AP) summarizes the curve as the weighed mean of precision at each threshold, with the increasing recall from the previous the threshold used as weight:

$$AP = \sum_n (R_n - R_n - 1)P_n \tag{3.7}$$

where P_n and R_n are the precision and recall at the nth threshold. A pair (R_k, P_k) is referred to as an operation point.

3.3.7. Equal Error Rate

Equal error rate(EER) point is the point where precision equals recall. In the 1-Precision-Recall curve, the EER point is the cross point of the blue curve and red line in Figure 3.5(b). Some people refer it to as a ”natural” operation point.

3.3.8. F1 Measure

F-measure, also known as F-score, is a measure that combines precision and recall is the harmonic mean of precision and recall. This measure is approximately the average of the two when they are close, and is more generally the harmonic mean, which, for the case of two numbers, coincides with the square of the geometric mean divided by the arithmetic mean. The general $F_\beta measure$ for non-negative real values of β is:

$$F_\beta = (1+\beta^2)\frac{precision \times recall}{\beta^2 \times precision + recall}$$

When the recall and precision are evenly weighted, it is known as F_1 measure:

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{3.8}$$

3.3.9. Recall,Precision,F1 Measure-Occlusion Level Curve

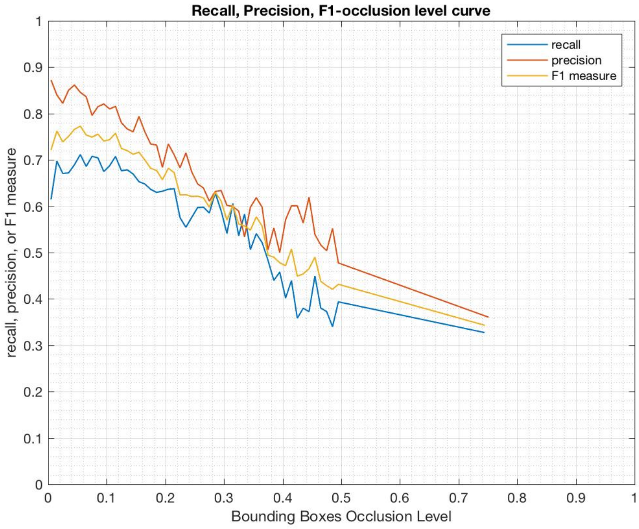


Figure 3.6: Example of Recall,Precision,F1 Measure-Occlusion Level Curve

Since this research studies the occlusion in people detectors, it is crucial to analyze the performance at different occlusion levels. Thus we plot the recall, precision, and F1 measure for bounding boxes with different occlusion level as shown in Figure 3.6. To extract this information from the predicted data, we calculate the occlusion level for every bounding box with Equation 3.1 in Section 1.2 and bin them with $binwidth = 0.01$ for all bounding boxes with occlusion level less than 0.5. The bin width is selected by the `histogram` function in Matlab to use a larger bin width corresponding to the the maximum number of bins. If we bin all bounding boxes with $binwidth = 0.01$, there is not enough bounding boxes with occlusion level larger than 0.5 to deliver convincing performance metrics. Thus all the bounding boxes with occlusion level larger than 0.5 are bin into a group. When plotting, the performance measure points are drawn at the median of the bounding box occlusion level in the corresponding bin. For example, in Figure 3.6, the last points of curves are drawn at around 0.75, which is the median of the bin over 0.5.

4 Experiments

In this chapter, we evaluate and analyze the performance of the models trained using the Overfeat-GoogLeNet system experimentally. All the training data and test data are from the video and annotation in MatchNMingle dataset.

MatchNMingle has the position of bounding box for every people in frames, so we can use this information to analysis quality of this dataset, such as number of images, number of bounding box, bounding box density, and occlusion level. The evaluation metrics used in this chapter are selected from related papers, and described in Chapter 3.3.

Apart from the details of dataset in Section 4.1, the experiments and results are described in the sequence of the hypotheses mentioned in Chapter 1.3. Finally, in Section 4.4 we provide modified training data selection strategy for this deep network as well as the performance evaluation of models trained with the selected data.

4.1. Statistic of Datasets

The annotations for positions in MatchNMingle contain the location of people by body bounding boxes. To study the occlusion in this situation, the quantitative analysis is necessary for the dataset.

Later the head annotation was added. Thus, the quantitative analysis was done for data with different annotation. There are four metrics for the quantitative analysis, of which are number of images in the dataset, total bounding box number(BB number), bounding box density(BB density), and average image occlusion level. Bounding box density is calculated by the equation:

$$BBdensity = \frac{BBnumber}{number\ of\ images} \tag{4.1}$$

The average image occlusion level is calculated by the Equation 3.2.

Data	Number of images	BB number*	BB density*	Average image occlusion level**
Day1	4228	43933	10.1973	1.194
Day2	4438	40789	9.1909	0.7213
Day3	4472	43729	9.7784	1.444
Day1&2	8664	83883	9.6818	0.9519
Day1&3	8698	86823	9.9819	1.3225
Day2&3	8910	84518	9.4858	1.084
Day1&2less	4188	37712	9.0048	0.4619
Day1&2more	4476	46171	10.3152	1.4103
Day1&3less	3705	36882	9.9547	0.9174
Day1&3more	4993	49941	10.0022	1.6232
Day2&3less	4915	44050	8.9624	0.7657
Day2&3more	3995	40468	10.1297	1.4756

* ”BB” refers to bounding box

** It is calculated by the Equation 3.2

Table 4.1: Quality of datasets with body annotation in MatchNMingle

Firstly, the data with body annotation is for the experiments referring to the first two hypotheses. The videos in MatchNMingle are taken in three days for the three-day speed date event. Since the participants in different days are different, the people in the videos of different days have different hair style and clothes. The most visible parts of people are head and shoulders.

The training data and validation set for every model is composed of data from two days, and the remaining data from the other day is used as test data for this model. This is called a leave-one-day-out cross-validation. In Table 4.1, we have different subsets. The subsets of Day1, Day2, and Day3 are used as test sets, while the subsets of Day1&2, Day1&3, Day2&3 are used as training sets and validation sets. It can be seen from the measures in the table that there is a different average image occlusion level for different days. The bounding box density and average image occlusion level of Day2 is much lower than that of Day1 and Day3.

In addition, as we need experiments with different occlusion levels, all the subsets for training are separated into two groups(”less” and ”more”) by the average occlusion level per image. We do this for the experiments of second hypothesis. The ”less” group contains frames with occlusion level lower than the median of average image occlusion level of all available for each subset: Day1&2, Day1&3, and Day2&3. The remaining images are classified into the ”more” sets. The median of the image average occlusion level among all data, 0.8, is used as the threshold for these two groups. Thus, all the images from subsets in ”less” group have the image occlusion level lower or equal to 0.8. With this separation, the subsets in ”less” group have much lower bounding box density and average image occlusion level than subsets in ”more” group. This means that in the ”less” group, there are less bounding boxes and bounding boxes have lower possibility to overlap with each other.

Data	Number of images	BB number*	BB density*	Average image occlusion level**
HDay1	4228	43933	10.391	0.0218
HDay2	4479	39198	8.7515	0.0505
HDay3	4473	41902	9.3678	0.0348
HDay1&2	8707	83131	9.5476	0.0366
HDay1&3	8701	85835	9.8649	0.0285
HDay2&3	8952	81100	9.0594	0.0426

* ”BB” refers to bounding box

** It is calculated by the Equation 3.2

Table 4.2: Quality of datasets with head annotation in MatchNMingle

For our third hypothesis, we present the statistics of the head annotation in Table 4.2. To distinguish head annotation from body annotation, the letter ”H” is for the head-annotated data. Same as the body-annotated images, all the data is separated into different subsets for testing or training. The subsets of Day1, Day2, and Day3 are used for testing, while the subsets of Day1&2, Day1&3, and Day2&3 are used for training. Since we annotated the same frames as body annotation, the number of images in corresponding subsets for head annotation are the same as the number for body annotation.

The head annotation information is collected separately from the body annotation with different collection ways. There are three 30-minute videos of each day, which are nine videos in total. Annotating one frame per second, we have 16200 frames. Due to the budget and time limitation, it is impossible to annotate all 16200 frames for MatchNMingle dataset. Thus, we select the periods around the time point of form, merging, and dissolution of conversation formations. With this selection, the time periods when most of the participants are static are eliminated. The head annotation was done manually with the Amazon�much's Mechanical Turk(MTurk). With head annotation, the average occlusion level per image(lower than 0.1) is much lower than that of datasets with body annotation(around 1).

To see the concrete information of bounding box occlusion level in MatchNMingle, we calculate the bounding box occlusion level of every bounding box and sort them ascendingly in a vector. In Figure 4.1, we show the vector of ascending bounding box occlusion level for data with body annotation(blue curve) and head annotation(red curve). The bounding box occlusion level is calculated by the Equation 3.1, and it represents the occlusion ratio for each bounding box. For body-annotated data, most of the bounding boxes have low occlusion levels of at most 0.4. The extreme cases with very high occlusion level are rare. For head-annotated data, most of the bounding boxes have no occlusion. Thus, by using of head annotation, we can avoid most of occluded bounding boxes in body-annotated data.

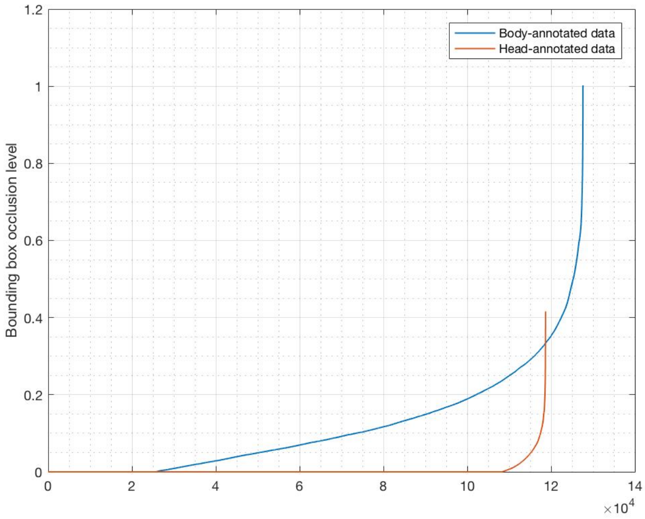


Figure 4.1: Bounding Box Occlusion levels ascending curve

4.2. Body Occlusion Experiments

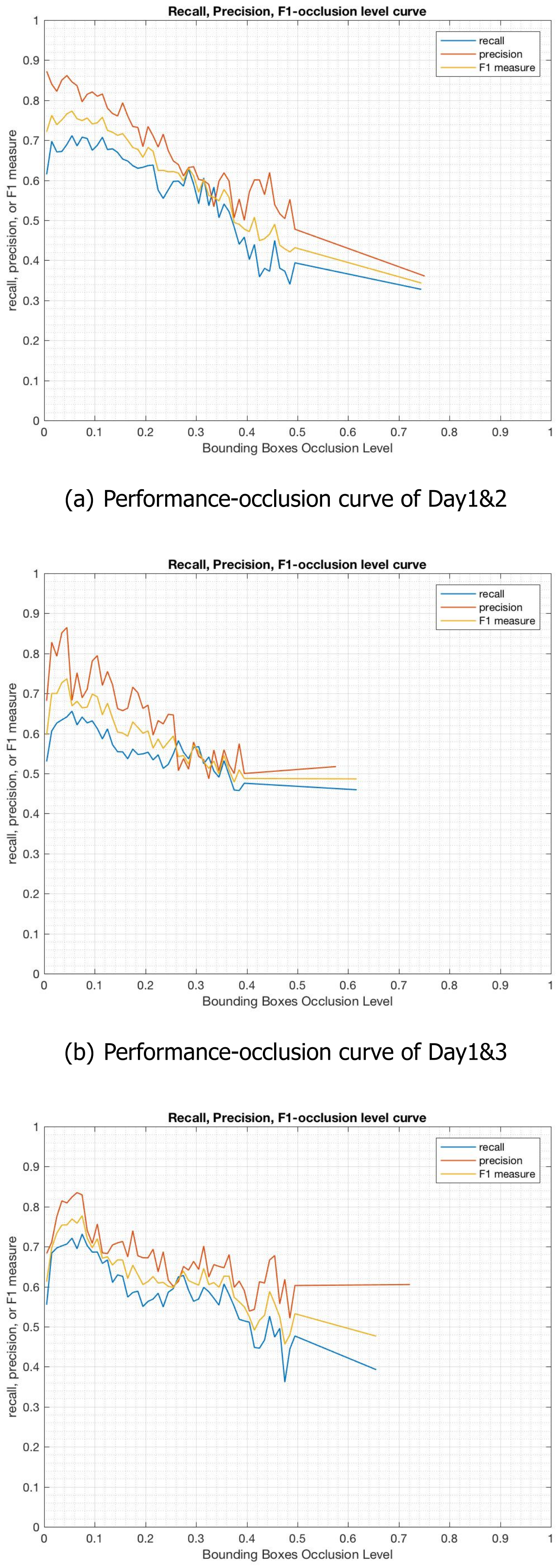
In this section, we describe the experiments and performance evaluation of research question 1. Two hypotheses based on this research question, are discussed.

4.2.1. Hypothesis 1: Detectors make more mistakes with increase of occlusion level.

Training data	Test data	AP	Recall	EER	COUNT	F1 measure
Day1&2	Day3	0.6296	0.7555	0.71	0.2	0.6868
Day1&3	Day2	0.5697	0.7122	0.68	0.2501	0.633
Day2&3	Day1	0.6114	0.7126	0.68	0.1655	0.6581
Average		0.6036	0.7267	0.69	0.2052	0.6593

Table 4.3: Performance evaluation of body annotated data in MatchNMingle

Three detectors are separately trained with Day1&2, Day1&3, and Day2&3, and tested with Day3, Day2, and Day1, respectively. Table 4.3 shows the results for average precision(AP, calculated with Equation 3.7), recall(calculated with Equation 3.3), equal error rate(EER), count(calculated with Equation 3.6), and F1 measure(calculated with Equation 3.8) for the three detectors. For cross-validation, each detector represents one fold and the average value of these metrics is given as well. From the results of average precision and recall, the value of recall is respectively lower than that of average precision. Thus, these detectors tend to return more predicted items, and many of them are predicted incorrectly.



(a) Performance-occlusion curve of Day1&2

(b) Performance-occlusion curve of Day1&3

(c) Performance-occlusion curve of Day2&3

Figure 4.2: Performance-occlusion curves of body annotated data in MatchNMingle

To see the performance of detectors with different levels of occlusion, we plot the recall, precision, and F1 measure with increasing occlusion levels in Figure 4.2, as mentioned in Section 3.3.9. Figures of 1-Precision-Recall curve are listed in Appendix B. It can be seen in Figure 4.2 that the recall, precision and F1 measure show a fluctuate declining as the bounding boxes occlusion level increases. But this tendency becomes fluctuant when the occlusion level becomes too large. This unstable condition may be the result of lack of heavy occluded bounding boxes in the test data. One unexpected case is that there is a small increase in the beginning of these curves, which means that the detection performance for bounding boxes without occlusion is slightly lower than that with a little occlusion. A possible explanation for it is that the detectors maybe overfit at the bounding boxes with no occlusion due to the much larger number of non-overlapped bounding boxes in training data.

From the performance evaluation results, as we expected, the value of recall, precision, and F1 measure decrease with increasing of bounding box occlusion level. It suggests that bounding boxes with higher occlusion level are more difficult to detected.

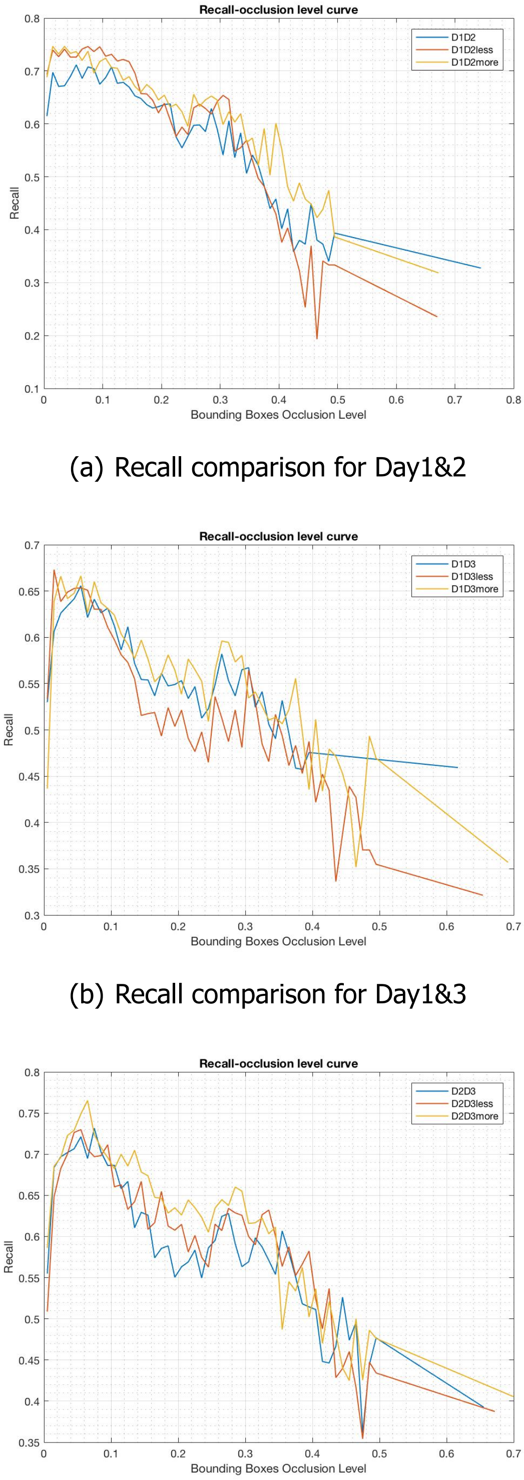
4.2.2. Hypothesis 2: Detectors trained with subsets with low occlusion levels make more mistakes for people with high occlusion levels.

Training data	Test data	AP	Recall	EER	COUNT	F1 measure
Day1&2less	Day3	0.6843	0.7667	0.74	0.1203	0.7232
Day1&3less	Day2	0.5635	0.7221	0.68	0.2813	0.633
Day2&3less	Day1	0.6174	0.7378	0.69	0.195	0.6723
Average		0.6191 (+0.0155)*	0.7422 (+0.0155)*	0.7033 (0.0133)*	0.1989 (-0.0063)*	0.6762 (+0.0169)*
Day1&2more	Day3	0.6785	0.7632	0.72	0.1249	0.7183
Day1&3more	Day2	0.564	0.6894	0.66	0.2223	0.6205
Day2&3more	Day1	0.6497	0.7121	0.68	0.0961	0.6795
Average		0.6307 (+0.0271)*	0.7216 (-0.0146)*	0.6867 (-0.0033)*	0.1478 (-0.0574)*	0.6728 (+0.0202)*

* compared with average value in Table 4.3

Table 4.4: Performance evaluation of body annotated data with different occlusion level

We trained three detectors with subsets in ”less” group, and three detectors with subsets in ”more” group. When testing, we applied images with ”less” occlusion and ”more” occlusion together. Table 4.4 shows the performance evaluation results for ”less” models and ”more” models. The 1-Precision-Recall curve and Recall, precision, F1-occlusion level curve for these models are listed in Appendix C.



(a) Recall comparison for Day1&2

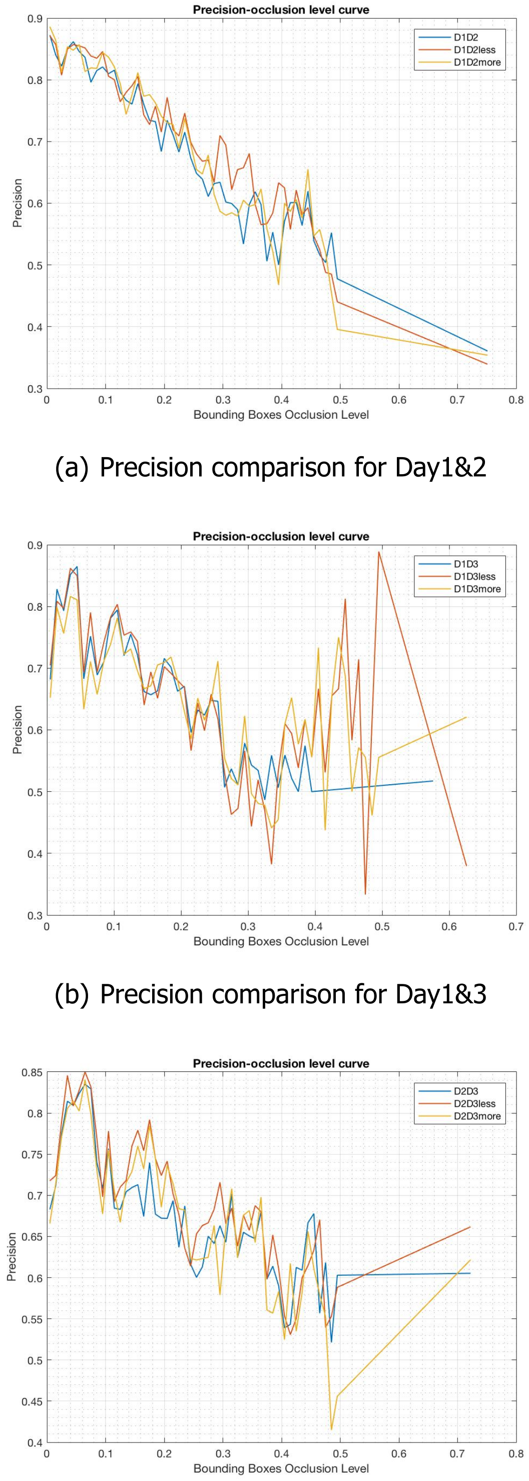
(b) Recall comparison for Day1&3

(c) Recall comparison for Day2&3

Figure 4.3: Recall-occlusion curve comparison

Comparing the values of evaluation metrics for ”less” and ”more” models to the general models trained in Table 4.3, it suggests that for detectors in both groups, the average precision increases. It indicates that the detectors return less detected items. For detectors trained with ”less”, there are more items predicted correctly in the detected items as the average recall is 0.0155 higher. For detectors trained with ”more”, there are less items predicted correctly in the detected items as the average recall is 0.0146 lower.

We show the Recall, precision, F1-occlusion level curve for the six models and try to compare them with the general models in Section 4.2.1. Since there are too many curves for these models, to make the comparison simple and easy to understand, these curves are compared separately. In Figure 4.3, Figure 4.4, and Figure 4.5, the recall, precision, and F1 measure curves for ”less”, ”more” models and corresponding general model are displayed in different sub-figures.



(a) Precision comparison for Day1&2

(b) Precision comparison for Day1&3

(c) Precision comparison for Day2&3

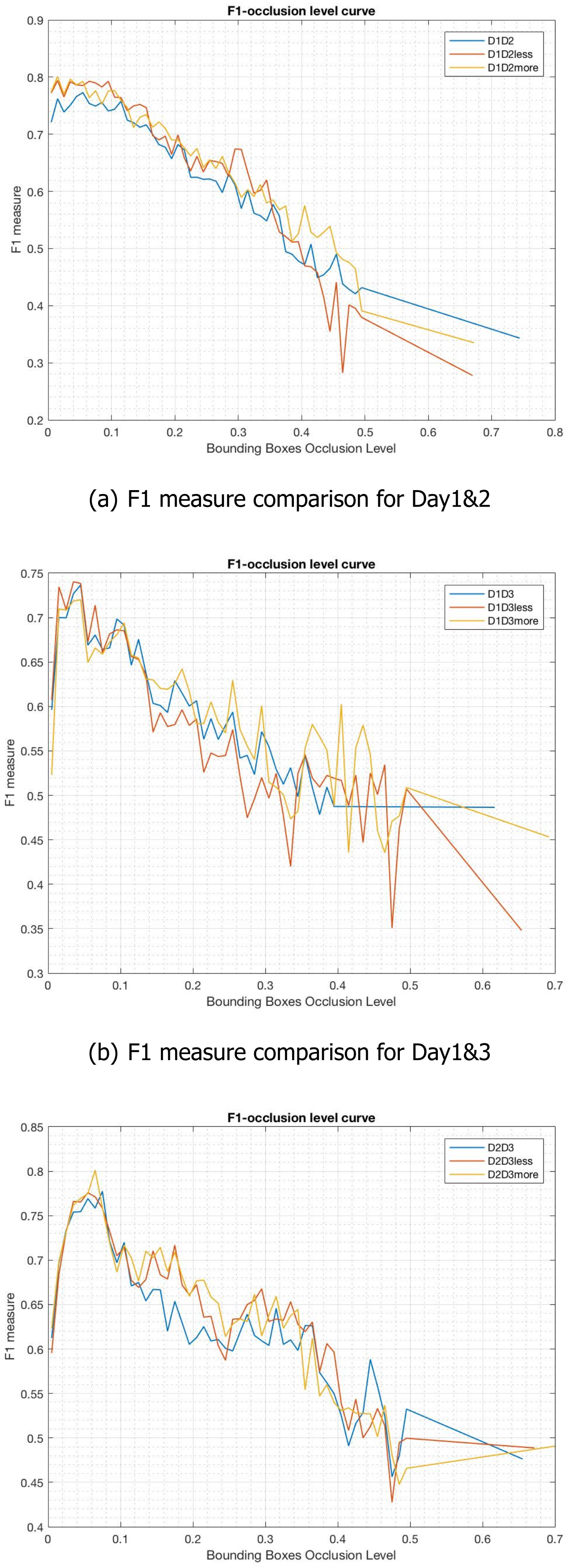
Figure 4.4: Precision-occlusion curve comparison

Compared to the general models, the value of recall for most of ”less” and ”more” models improved slightly. And the recall of ”more” models at higher bounding box occlusion level is higher than that of ”less” models. The recall of ”less” models at lower bounding box occlusion level is not always lower than that of ”more” models.

As for the precision, compared to the general model, only the average precision of ”less” and ”more” models trained with Day1&2 data is better, while the other models trained with Day1&3, and Day2&3 perform similarly to the general models. And for the precision curves, the precision decreases with fluctuation, and with not much difference from the general models.

Since F1 measure combines the recall and precision together, it has similar tendency to recall.

From the analysis of recall, precision and F1 measure, in general, the detectors trained with subsets with low occlusion perform worse than detectors trained with higher occlusion level. The performance at low occlusion level of all our models is similar. It indicates that for every training subset there is enough or even too many samples with low occlusion level. And the higher ratio of samples with higher occlusion level shows the possibility to improve the correct predictions.

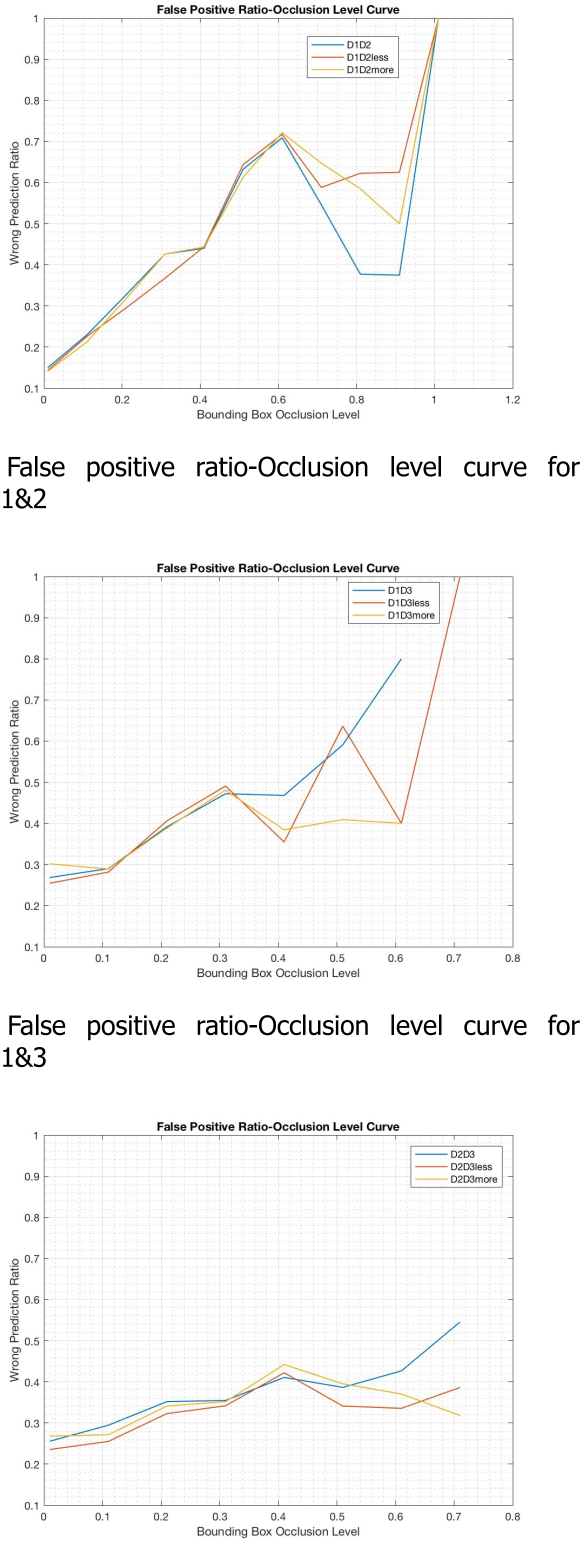


(a) F1 measure comparison for Day1&2

(b) F1 measure comparison for Day1&3

(c) F1 measure comparison for Day2&3

Figure 4.5: F1 measure-occlusion curve comparison



(a) False positive ratio-Occlusion level curve for Day1&2

(b) False positive ratio-Occlusion level curve for Day1&3

(c) False positive ratio-Occlusion level curve for Day2&3

Figure 4.6: False positive ratio-Occlusion level curve

It is helpful to understand all the trained models if we can find where the wrong predictions are. Here, the false positive ratio defined by Equation 3.5 can tell the ratio of wrong predictions at different occlusion levels. Figure 4.6 below shows the wrong predictions ratio on different occlusion levels for models trained with ”less”, ”more”, and overall data. In these figures, we can see the models trained with “less” make more wrong predictions at higher occlusion levels. The models trained with “more” make less wrong predictions at higher occlusion levels but make similar wrong predictions at lower occlusion levels compared with “less” models.

Most of the wrong predictions happen on bounding boxes with low occlusion levels especially bounding boxes with almost no occlusion. This is because in the test set, most of the bounding boxes have very low occlusion levels. Even the ratio of wrong predictions is very low, the number of wrong prediction is still high compared to other occlusion levels.

According to the wrong prediction ratio analysis, there are two ways to improve the detection performance for our existing models. First, even though the wrong prediction ratio becomes higher with occlusion level increasing, it makes sense to improve the performance at low occlusion levels. In our condition, most of bounding boxes have low occlusion and our detectors show unexpected increase at the beginning of the recall, precision-occlusion level curve. Thus, it is necessary to overcome this problem. Second, in training data, the number of samples with high occlusion levels is apparently less than that with low occlusion levels. This means there is not enough samples with high occlusion levels for training. Due to the common learning curve for deep models, it is possible to enhance the performance at high occlusion levels by training with more occluded samples.

4.3. Head Annotation Experiments

In this section, the experiments and performance evaluation for the third hypothesis are described.

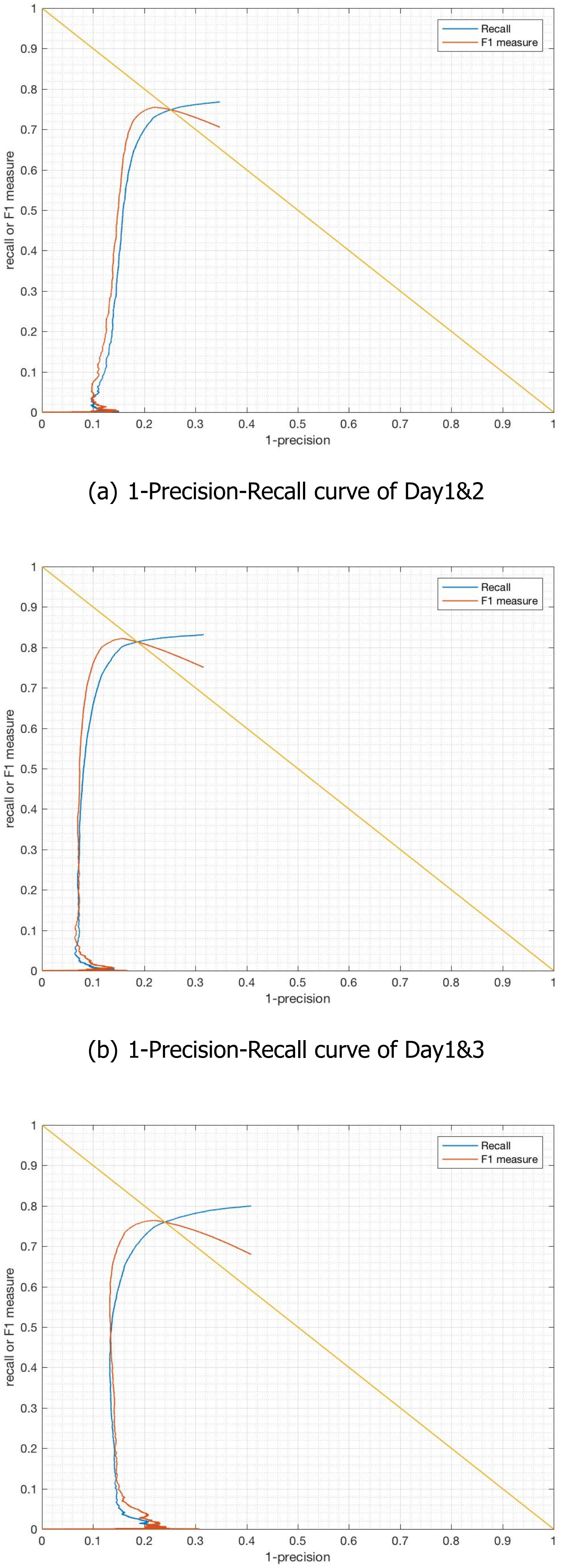
4.3.1. Hypothesis 3: Detector trained with head annotation performs better than detector trained with body annotation.

Training data	Test data	AP	Recall	EER	COUNT	F1 measure
HDay1&2	HDay3	0.6583	0.7589	0.74	0.1529	0.705
HDay1&3	HDay2	0.6839	0.8313	0.81	0.2156	0.7597
HDay2&3	HDay1	0.591	0.7967	0.77	0.3535	0.6797
Average		0.6444 (+0.0408)*	0.7967 (+0.07)*	0.77 (+0.08)*	0.2407 (+0.0355)*	0.7117 (+0.0522)*

* compared with average value in Table 4.3

Table 4.5: Performance evaluation of head annotated data in MatchNMingle

We trained and tested three detectors with head-annotated data. To compare the performance of models trained with head annotation and body annotation, the models trained in Section 4.2.1 use the same frames as the models trained in this section. The performance evaluation metrics are shown in Table 4.5, and the 1-Precision-Recall curves are shown in Figure 4.7. From the performance table, the average value of AP, recall, EER, and F1 measure increase, which means the head detectors return less detected items and more of them are predicted correctly comparing to the body detectors. Unexpectedly, the precision of model trained with HDay2&3 is lower than that of model trained with body annotation, while the recall is higher. This suggests the detector returns more predicted items and most of them are predicted correctly. The possible reason for this case can be the quantitative analysis of head-annotated data. From Table 4.2, the bounding box density of HDay1 is much higher than that of HDay2 and HDay3. This means the difficulty of detecting people in HDay1 is higher. However, in general, as expected, head detectors tend to be a better solution when detecting people from the overhead cameras.



(a) 1-Precision-Recall curve of Day1&2

(b) 1-Precision-Recall curve of Day1&3

(c) 1-Precision-Recall curve of Day2&3

Figure 4.7: 1-Precision-Recall Curves for head annotation data

4.4. Modified Body-annotated Training Data Experiments

4.4.1. Hypothesis 4: Body detectors trained with less images with low occlusion level perform better than the initial detectors in Experiment 1.

Data	Number of images	BB number	BB density	Average occlusion level(per image)
MDay1&2	5753	59616	10.3626	1.1923
MDay1&3	5805	61150	10.5340	1.6596
MDay2&3	5696	58382	10.2496	1.4265

Table 4.6: Quality of modified body annotated data in MatchNMingle

Training data	Test data	AP	Recall	EER	COUNT	F1 measure
MDay1&2	Day3	0.7027	0.7352	0.72	0.0461	0.7186
MDay1&3	Day2	0.6117	0.7112	0.69	0.1626	0.6577
MDay2&3	Day1	0.6731	0.7283	0.7	0.0821	0.6996
Average		0.6625 (+0.0589)* (+0.0181)**	0.7249 (-0.0018)* (-0.0718)**	0.7033 (+0.0133)* (-0.0667)**	0.0969 (-0.1089)* (-0.1438)**	0.6920 (+0.0327)* (-0.0197)**

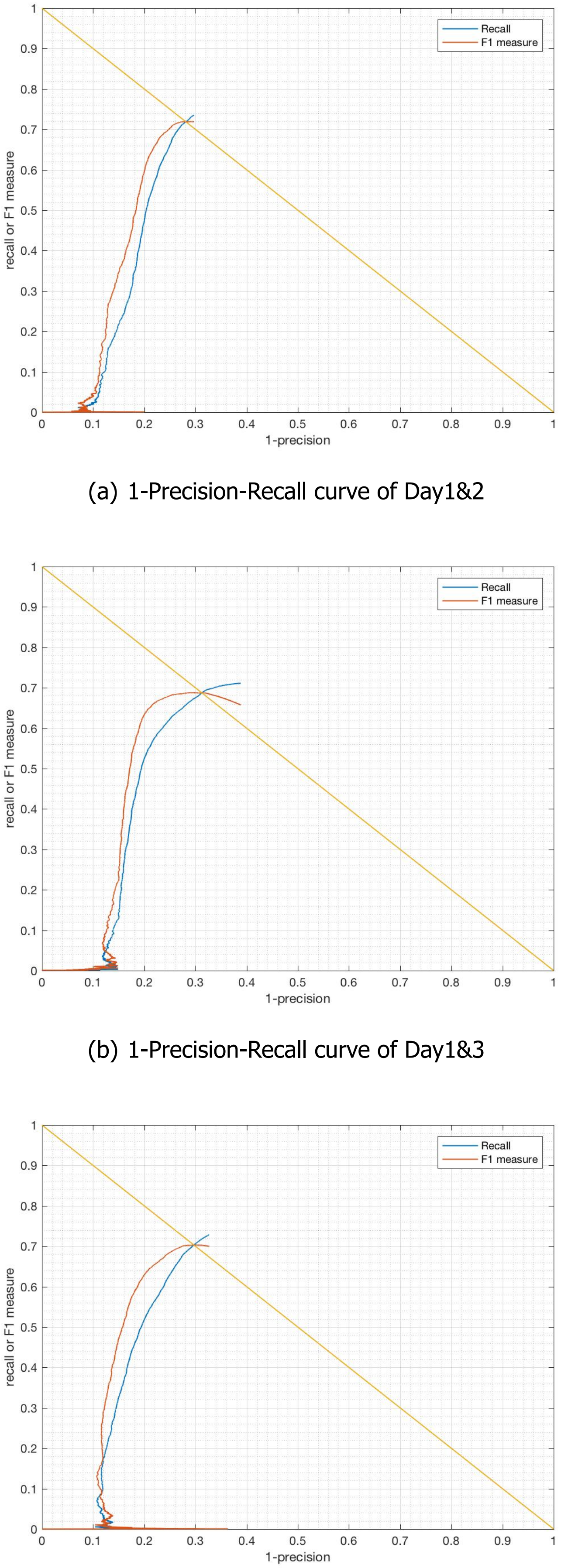
* compared with average value in Table 4.3
** compared with average value of head detectors in Table 4.5

Table 4.7: Performance evaluation of modified models

With the analysis of results from above experiments, I propose modified training subsets based on the training subsets used in Section 4.2.1. The idea of modification is to reduce the number of samples with low occlusion level and to increase the ratio of samples with high occlusion level. Thus, we randomly remove half of the images with occlusion level lower than the average image occlusion level, and trained detectors again with these modified training subsets. Before training, the quantitative analysis is done for the data, and the result is shown in Table 4.6. It is clear that after the modification the number of images and bounding boxes is lower, while the bounding box density and average image occlusion level becomes higher.

According to the performance measure in Table 4.7 and Figure 4.8, the average precision increases, while the value of recall almost stays static compared to the initial models. And this leads to the improvement of EER and F1 measure. Thus, the modified models return less predicted items and most of them are predicted correctly.

Even though there is an improvement for models trained with modified training data, the performance of body detectors still can not reach that of head



(a) 1-Precision-Recall curve of Day1&2

(b) 1-Precision-Recall curve of Day1&3

(c) 1-Precision-Recall curve of Day2&3

Figure 4.8: 1-Precision-Recall Curves for Modified models

detector. Compared to head detectors, the average precision of modified body detectors is slightly better, while there is large distance between the value of recall. The average recall of head detectors is about 10% better, and this leads to larger EER and F1 measure. However, the modified body detectors seem to have better ability for counting people, since the their average count is much lower than that of initial body detectors and head detectors.

5 Conclusion and Future Work

5.1. Conclusion

This thesis work study the impact of occlusion on performance of models trained by Overfeat-GoogLeNet network with subsets of different occlusion levels. With MatchNMingle dataset, our research goes through four hypotheses and corresponding experiments.

For Hypothesis 1, we trained three detectors, and these detectors tend to returns more predicted items and most of them are predicted incorrectly when compared to the ground truth. With the increase of occlusion level, the value of recall reduces from around 0.7 to of at most 0.5; the value of precision reduce from around 0.8 to of at most 0.6; the value of F1 measure reduces from around 0.7 to less than 0.5. Their performance at different occlusion levels suggests that the bounding boxes with higher occlusion level are more difficult to detect.

For Hypothesis 2, we trained models with less occlusion and more occlusion separately. We found that the models trained with higher occlusion levels return better predictions at higher occlusion level. In the original training data, it seems that the ratio of images with low occlusion levels is too high, and this may lead to overfit. In addition, it seems that detectors trained with less frames have higher average precision. It means there detectors return less detected items. And the detectors have better counting ability as COUNT reduces.

The third hypothesis about head annotation is more like an attempt and verification that head detector is supposed to be a better solution in crowded scenes. With experiments and evaluation, the average precision, recall and EER improve. The head detectors reach 0.0522 larger F1 measure comparing to the body detectors. Thus, head detectors returns less prediction items and most of them are correctly predicted. These models are closer to the ideal detector with high precision and high recall.

For Hypothesis 4, we proposed modification of training data for body-annotated data. The modification strategy is to reduce the ratio of images with low occlusion level and improve the average occlusion level of the dataset. With the modified training data, we obtained three body detectors. There is a remarkable improvement on the precision of these models, which means our modification reduces the number of detected items and results in better balance between precision and recall. However, this improvement is segmentary. It seems that these models have better counting ability, but the ratio of undetected people does not reduce. Comparing to the comprehensive improvement on both precision and recall of head detectors, the body detectors still have critical limitation. Therefore, in the situation of MatchNMingle, head detector is a better solution for people detection.

5.2. Future Work

Firstly, according to the occlusion level ascending curves (Figure 4.1), we are lack of instances with high occlusion level. This limitation influence on the body occlusion experiments. If more samples with high occlusion level are collected, we can separate them into more groups for training and testing rather than only ”less” and ”more” groups in Section 4.2.2. In this way, the analysis of relation between training set occlusion level and performance at different occlusion level can become more detailed and effective.

Secondly, it is meaningful if the accuracy and uniformity of annotation on our dataset can be improved by manually checking. There are some inappropriate bounding boxes when people is moving in the camera. And the head annotation are collected through MTurk, where the workers are not trained to do annotation. This results in bounding boxes of different styles. Some bounding boxes contain more background, while some contain not the complete head. Some workers annotated the partical visible people on the edge, while some did not. And some workers only annotated the visible part when people are overlapping, while some guessed the position of the whole head and annotated. This divergence of annotation could be misleading when training, especially the case of occlusion.

Thirdly, the variation of our dataset is low comparing to the images in Brainwash. When extracting images from video, they ensured a fixed interval of 100 seconds to obtain a dataset with large variation. Variation in images is important for training and evaluation.

Lastly, validation and analysis on single dataset is not enough for a research, so it is better to do more experiments with other dataset for people detection. There is not much dataset contains people from the overhead cameras, but occlusion is a common challenge for people detection. Thus, study of occlusion is not limited in our dataset.

A

GoogLeNet

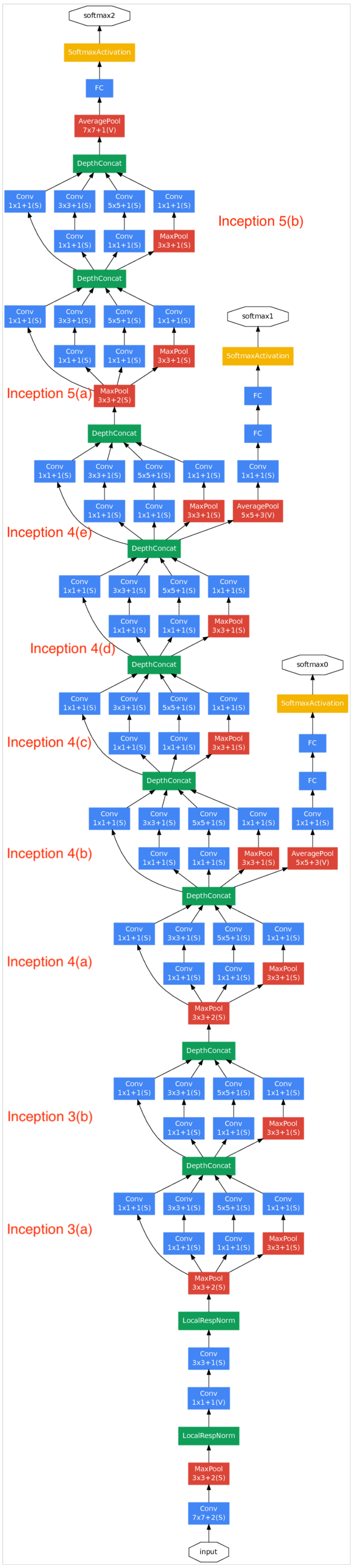
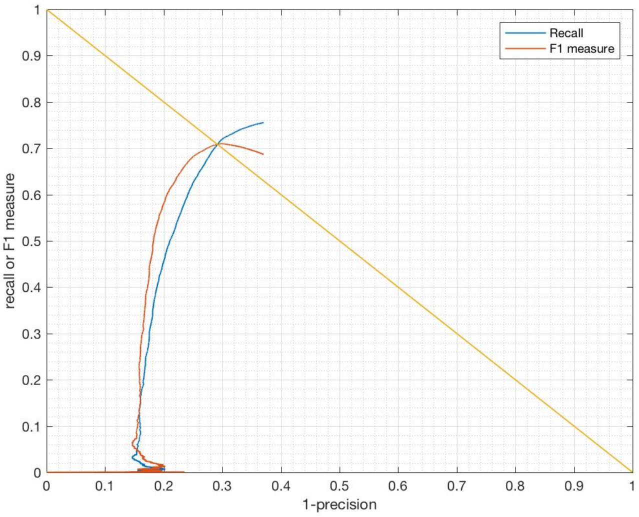


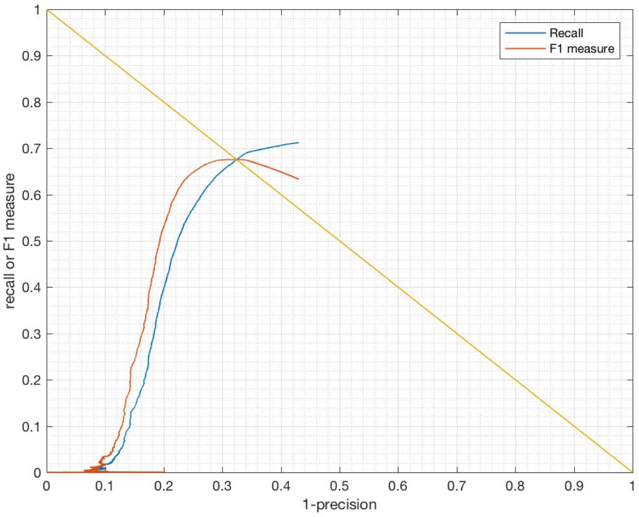
Figure A.1: GoogLeNet network with all the bells and whistles

B

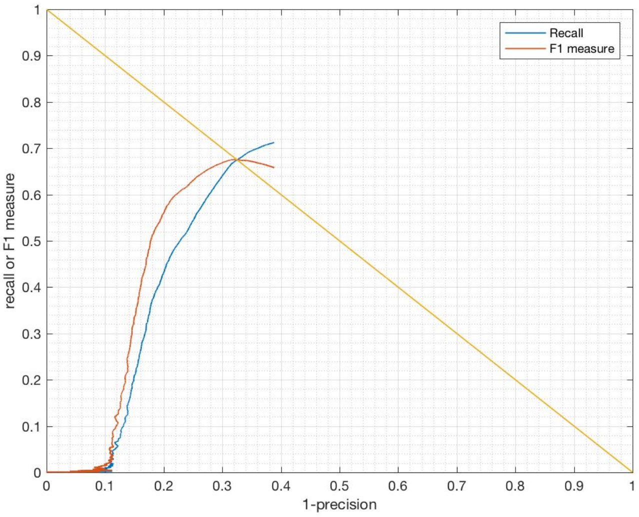
Curves for Hypothesis 1



(a) 1-Precision-Recall curve of Day1&2



(b) 1-Precision-Recall curve of Day1&3

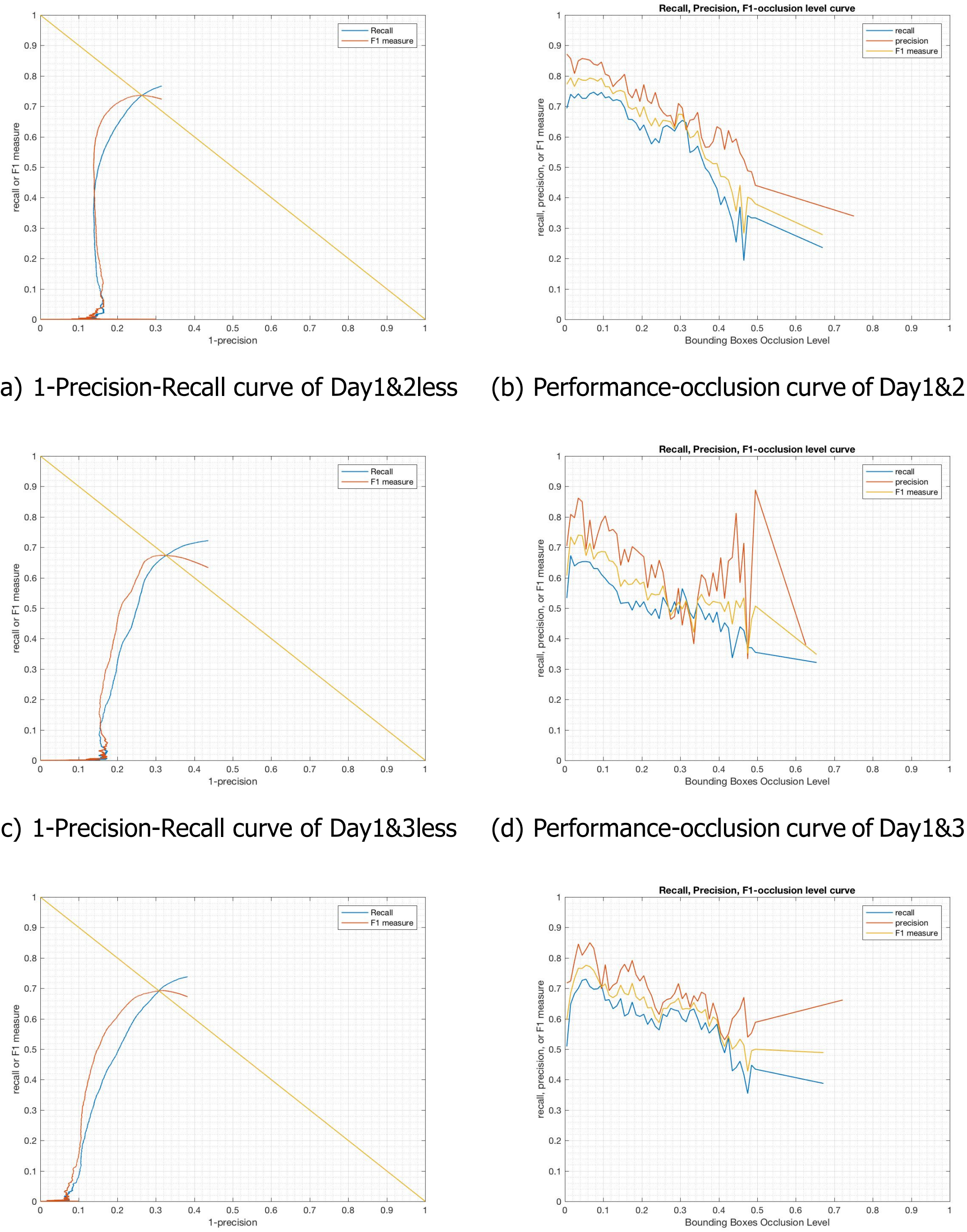


(c) 1-Precision-Recall curve of Day2&3

Figure B.1: 1-Precision-Recall curves of body annotated data in MatchNMingle

C

Curves for Hypothesis 2



(a) 1-Precision-Recall curve of Day1&2less

(b) Performance-occlusion curve of Day1&2less

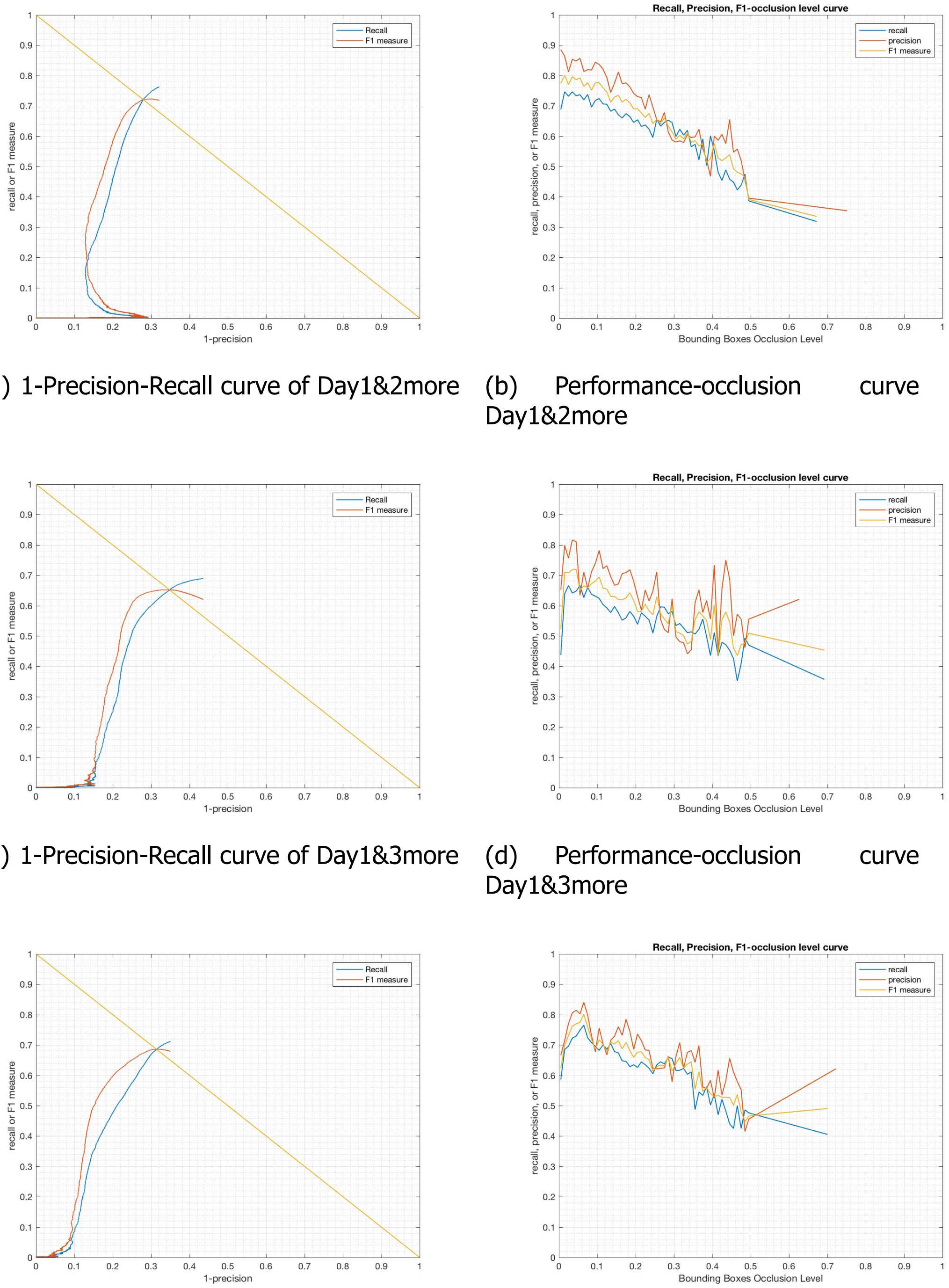
(c) 1-Precision-Recall curve of Day1&3less

(d) Performance-occlusion curve of Day1&3less

(e) 1-Precision-Recall curve of Day2&3less

(f) Performance-occlusion curve of Day2&3less

Figure C.1: Performance evaluation of body annotated data with lower occlusion level



(a) 1-Precision-Recall curve of Day1&2more

(b) Performance-occlusion curve of Day1&2more

(c) 1-Precision-Recall curve of Day1&3more

(d) Performance-occlusion curve of Day1&3more

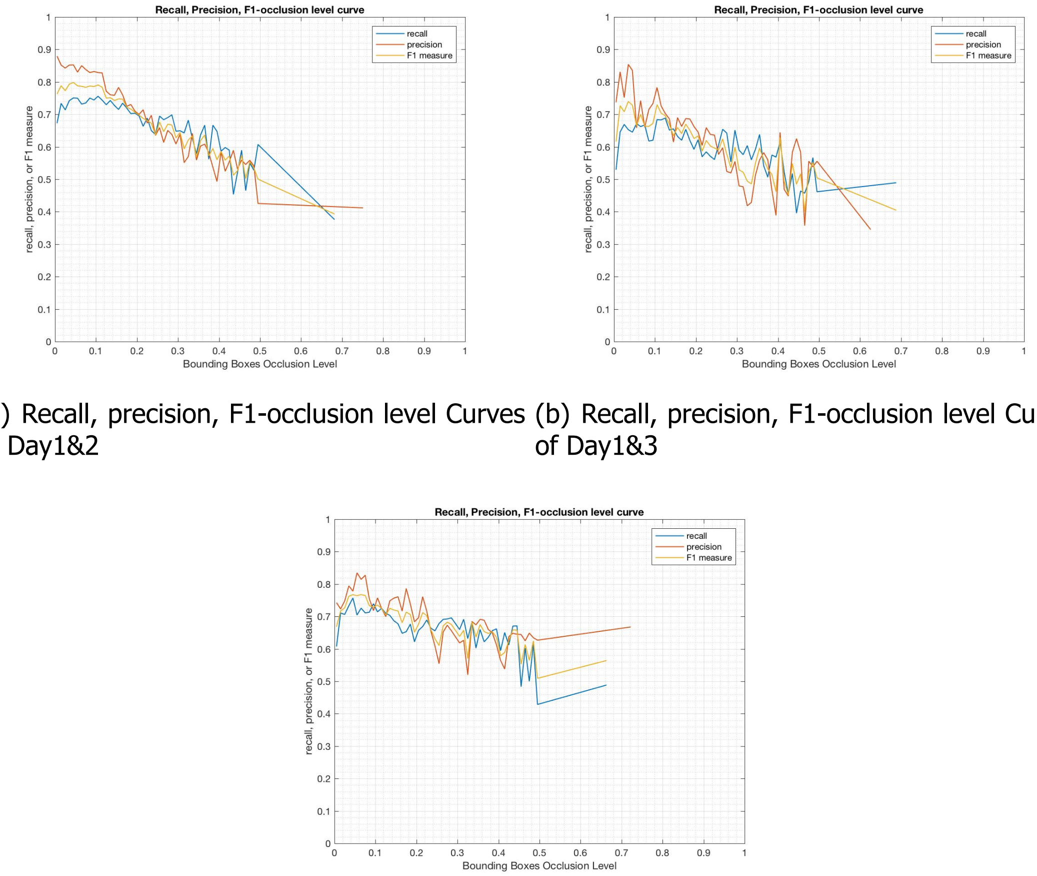
(e) 1-Precision-Recall curve of Day2&3more

(f) Performance-occlusion curve of Day2&3more

Figure C.2: Performance evaluation of body annotated data with higher occlusion level

D

Curves for Modified models



(a) Recall, precision, F1-occlusion level Curves of Day1&2

(b) Recall, precision, F1-occlusion level Curves of Day1&3

(c) Recall, precision, F1-occlusion level Curves of Day2&3

Figure D.1: Recall, precision, F1-occlusion level Curves for Modified models

Bibliography

[1] D. T. Nguyen, *Human detection from images and videos,* Doctor of Philosophy thesis (2012).

[2] S. Zhang, R. Benenson, B. Schiele, *et al.*, *Filtered channel features for pedestrian detection.* in *CVPR*, Vol. 1 (2015) p. 4.

[3] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, *Going deeper with convolutions,* in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2015) pp. 1–9.

[4] P. Dollar, C. Wojek, B. Schiele, and P. Perona, *Pedestrian detection: An evaluation of the state of the art,* IEEE transactions on pattern analysis and machine intelligence **34**, 743 (2012).

[5] E. G. L. v. d. M. Laura Cabrera-Quiros, Andrew Demetriou and H. Hung, *The matchnmingle dataset: a novel multi-sensor resourse for the analysis of social interactions and group dynamics in-the-wild during free-standing conversations and speed dates,* (2018).

[6] R. Stewart, M. Andriluka, and A. Y. Ng, *End-to-end people detection in crowded scenes,* in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016) pp. 2325–2333.

[7] C. Vondrick, D. Patterson, and D. Ramanan, *Efficiently scaling up crowdsourced video annotation,* International Journal of Computer Vision **101**, 184 (2013).

[8] R. Datta, D. Joshi, J. Li, and J. Z. Wang, *Image retrieval: Ideas, influences, and trends of the new age,* ACM Computing Surveys (Csur) **40**, 5 (2008).

[9] C. Papageorgiou and T. Poggio, *A trainable system for object detection,* International Journal of Computer Vision **38**, 15 (2000).

[10] N. Dalal and B. Triggs, *Histograms of oriented gradients for human detection,* in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, Vol. 1 (IEEE, 2005) pp. 886–893.

[11] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, *The pascal visual object classes (voc) challenge,* International journal of computer vision **88**, 303 (2010).

[12] B. Wu and R. Nevatia, *Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors,* International Journal of Computer Vision **75**, 247 (2007).

[13] Z. Lin, L. S. Davis, D. Doermann, and D. DeMenthon, *Hierarchical part-template matching for human detection and segmentation,* in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on* (IEEE, 2007) pp. 1–8.

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

[15] C. Wojek, S. Walk, and B. Schiele, *Multi-cue onboard pedestrian detection,* (2009).

[16] D. Geronimo, A. M. Lopez, A. D. Sappa, and T. Graf, *Survey of pedestrian detection for advanced driver assistance systems,* IEEE transactions on pattern analysis and machine intelligence **32**, 1239 (2010).

[17] A. Ess, B. Leibe, and L. Van Gool, *Depth and appearance for mobile scene analysis,* in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on* (IEEE, 2007) pp. 1–8.

[18] A. Neubeck and L. Van Gool, *Efficient non-maximum suppression,* in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, Vol. 3 (IEEE, 2006) pp. 850–855.

[19] B. Leibe, E. Seemann, and B. Schiele, *Pedestrian detection in crowded scenes,* in *null* (IEEE, 2005) pp. 878–885.

[20] Y. Mu, S. Yan, Y. Liu, T. Huang, and B. Zhou, *Discriminative local binary patterns for human detection in personal album,* in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on* (IEEE, 2008) pp. 1–8.

[21] P. Viola, M. J. Jones, and D. Snow, *Detecting pedestrians using patterns of motion and appearance,* International Journal of Computer Vision **63**, 153 (2005).

[22] N. Dalal, B. Triggs, and C. Schmid, *Human detection using oriented histograms of flow and appearance,* in *European conference on computer vision* (Springer, 2006) pp. 428–441.

[23] E. M. Markman, *Thinking in perspective: Critical essays in the study of thought processes,* Psyccritiques **24**, 719 (1979).

[24] D. T. Nguyen, P. Ogunbona, and W. Li, *Human detection with contour-based local motion binary patterns,* in *Image Processing (ICIP), 2011 18th IEEE International Conference on* (IEEE, 2011) pp. 3609–3612.

[25] W. Ouyang and X. Wang, *Joint deep learning for pedestrian detection,* in *Proceedings of the IEEE International Conference on Computer Vision* (2013) pp. 2056–2063.

[26] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, *Object detection with discriminatively trained part-based models,* IEEE transactions on pattern analysis and machine intelligence **32**, 1627 (2010).

[27] S. Tang, M. Andriluka, and B. Schiele, *Detection and tracking of occluded people,* International Journal of Computer Vision **110**, 58 (2014).

[28] S. Tang, M. Andriluka, A. Milan, K. Schindler, S. Roth, and B. Schiele, *Learning people detectors for tracking in crowded scenes,* in *Proceedings of the IEEE international conference on computer vision* (2013) pp. 1049–1056.

[29] S. Zhang, C. Bauckhage, and A. B. Cremers, *Informed haar-like features improve pedestrian detection,* in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2014) pp. 947–954.

[30] D. Park, C. L. Zitnick, D. Ramanan, and P. Dollár, *Exploring weak stabilization for motion feature extraction,* in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2013) pp. 2882–2889.

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

[32] Y. Tian, P. Luo, X. Wang, and X. Tang, *Deep learning strong parts for pedestrian detection,* in *Proceedings of the IEEE international conference on computer vision* (2015) pp. 1904–1912.

[33] A. Angelova, A. Krizhevsky, V. Vanhoucke, A. S. Ogale, and D. Ferguson, *Real-time pedestrian detection with deep network cascades.* in *BMVC*, Vol. 2 (2015) p. 4.

[34] J. Hosang, M. Omran, R. Benenson, and B. Schiele, *Taking a deeper look at pedestrians,* in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2015) pp. 4073–4082.

[35] Y. Tian, P. Luo, X. Wang, and X. Tang, *Pedestrian detection aided by deep learning semantic tasks,* in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2015) pp. 5079–5087.

[36] L. Zhang, L. Lin, X. Liang, and K. He, *Is faster r-cnn doing well for pedestrian detection?* in *European Conference on Computer Vision* (Springer, 2016) pp. 443–457.

[37] Z. Cai, M. Saberian, and N. Vasconcelos, *Learning complexity-aware cascades for deep pedestrian detection,* in *Proceedings of the IEEE International Conference on Computer Vision* (2015) pp. 3361–3369.

[38] P. Luo, Y. Tian, X. Wang, and X. Tang, *Switchable deep network for pedestrian detection,* in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2014) pp. 899–906.

[39] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele, *How far are we from solving pedestrian detection?* in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016) pp. 1259–1267.

[40] R. Girshick, J. Donahue, T. Darrell, and J. Malik, *Rich feature hierarchies for accurate object detection and semantic segmentation,* in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2014) pp. 580–587.

[41] A. Krizhevsky, I. Sutskever, and G. E. Hinton, *Imagenet classification with deep convolutional neural networks,* in *Advances in neural information processing systems* (2012) pp. 1097–1105.

[42] M. D. Zeiler and R. Fergus, *Visualizing and understanding convolutional networks,* in *European conference on computer vision* (Springer, 2014) pp. 818–833.

[43] R. Girshick, *Fast r-cnn,* in *Proceedings of the IEEE international conference on computer vision* (2015) pp. 1440–1448.

[44] S. Ren, K. He, R. Girshick, and J. Sun, *Faster r-cnn: Towards real-time object detection with region proposal networks,* in *Advances in neural information processing systems* (2015) pp. 91–99.

[45] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, *Overfeat: Integrated recognition, localization and detection using convolutional networks,* arXiv preprint arXiv:1312.6229 (2013).

[46] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, *Caffe: Convolutional architecture for fast feature embedding,* in *Proceedings of the 22nd ACM international conference on Multimedia* (ACM, 2014) pp. 675–678.

[47] R. Stewart, *Tensorbox: A fast object detection framework in tensorflow,* (2016).

[48] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, *et al.*, *Tensorflow: a system for large-scale machine learning.* in *OSDI*, Vol. 16 (2016) pp. 265–283.

[49] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, *Imagenet: A large-scale hierarchical image database,* in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on* (Ieee, 2009) pp. 248–255.