EXPLOITING MULTISPECTRAL AND CONTEXTUAL INFORMATION TO IMPROVE HUMAN DETECTION

by

JINGJING LIU

A dissertation submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Computer Science

Written under the direction of

Dimitris N. Metaxas

And approved by

_____

_____

_____

_____

New Brunswick, New Jersey

OCTOBER, 2017

ABSTRACT OF THE DISSERTATION

Exploiting Multispectral and Contextual Information to Improve Human Detection

By JINGJING LIU

Dissertation Director:

Dimitris N. Metaxas

Human detection has various applications, e.g., autonomous driving car, surveillance system, and retail. In this dissertation, we first exploit multispectral images (i.e., RGB and thermal images) for human detection. We extensively analyze Faster R-CNN for the detection task and then model multispectral human detection into a fusion problem of convolutional networks (ConvNets). We design four distinct ConvNet fusion architectures that integrate two-branch ConvNets on different stages of neural networks, all of which yield better performance compared with the baseline detector. In the second part of this dissertation, we leverage instance-level contextual information in crowded scenes to boost performance of human detection. Based on a context graph that incorporates both geometric and social contextual patterns from crowds, we apply progressive potential propagation algorithm to discover weak detections that are contextually compatible with true detections while suppressing irrelevant false alarms. The method significantly improves the performance of any shallow human detectors, obtaining comparable results to deep learning based methods.

Acknowledgements

First of all, I would like to express my deepest gratitude to my advisor, Prof. Dimitris N. Metaxas, for his generous support and guidance throughout the past six years. Prof. Metaxas not only continuously encourages me to solve challenging and practical problems but also provides me with the excellent research environment that I have sufficient research freedom and lots of opportunities to collaborate with world-class researchers. He always gives me beneficial pieces of advice whenever I feel confused in research or life. His great passion in research, remarkable leadership, wealthy knowledge, and unexhausted energy set him an extraordinary example for me in both life and career.

I am also very grateful to the rest of my dissertation committee: Prof. Kostas Bekris, Prof. Jingjin Yu, Dr. Nalini K. Ratha (IBM T.J. Watson Research Center), for their time in reviewing my dissertation and their valuable suggestions. I would also like to express my gratitude to Prof. Konstantinos Michmizos and Prof. Santosh Nagarakatte who served on my oral qualifying exam committee, for their insightful comments on my research from various aspects.

I also appreciate my co-authors, Prof. Shaoting Zhang (University of North Carolina at Charlotte), Dr. Quanfu Fan (IBM T.J. Watson Research Center), Prof. Chao Chen (City University of New York), Prof. Carol Neidle (Boston University), Dr. Wei Liu (Tencent AI Lab), for their patience in guidance and generous help on my research.

Besides, I am also deeply grateful to Dr. Sharath U. Pankanti (IBM T.J. Watson Research Center) and Dr. Feng Tang (Apple Inc.) who provided me the great opportunities of working as their intern.

My sincere thanks also go to the folks of Computational Biomedicine Imaging and Modeling Center (CBIM), for our insightful discussions, collaborations on projects, and all kinds of

help and friendship. In addition, I am very grateful to the staff of Computer Science Department, Carol DiFrancesco, Maryann Holtsclaw, Ginger Olszewski, and Charles McGrew, who have given me lots of help.

Last but not least, I want to thank my parents for their love, understanding, support, and encouragement. I am proud to be their son.

Dedication

To my parents

Table of Contents

List of Tables

List of Figures

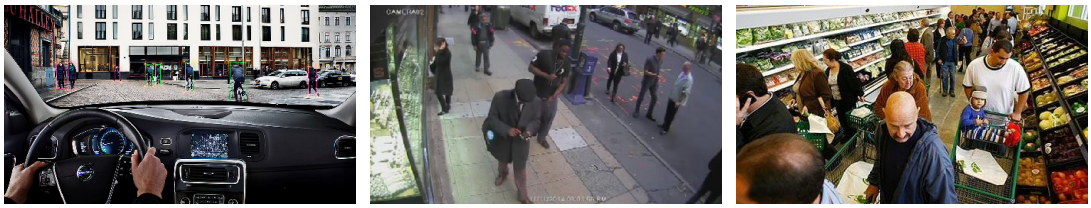
Chapter 1

Introduction

1.1 Background

The goal of human detection is to localize human in the physical world. There exists a bunch of techniques using different kinds of signals to localize people. Some signal modalities, as audio [1] and Wi-Fi signals [2], are out of the scope of our work. In this dissertation, we are focusing on the vision-based human detection that aims at predicting image locations of human objects by using computer vision techniques: given images (visible or invisible), the task of human detection is to generate bounding boxes (each one is represented by 4 image coordinates) for human hypotheses and their corresponding detection confidences. To be noticed, some researchers categorize human detection into people detection and pedestrian detection which have different assumptions and technical focuses. In general, people detection places extra emphasis on static images with human instances of various poses and layouts, while pedestrian detection assumes human objects are upright (walking or standing). Besides, motion patterns are frequently exploited in pedestrian detection to improve the performance [3]. We do not distinguish people detection and pedestrian detection in this dissertation.

Human detection has various applications in real-life, such as autonomous driving, video surveillance and retail (in Figure. 1.1). For instance, it is a principle component of autonomous driving systems which require the functionality of avoiding human hitting. It is very useful especially in urban areas where people are walking around or crossing streets. The localizations of people in surrounding areas could help the autopilot in planning safe driving paths. Besides, based on detected people, surveillance systems could automatically discover the occurrence of abnormal events, e.g., abandoned object and street fighting. By looking at the appearance of detected people, the intelligent surveillance system could also find out suspects according to



(a) Autonomous driving (b) Surveillance (c) Retail

Figure 1.1: Applications of human detection techniques in real-life applications.

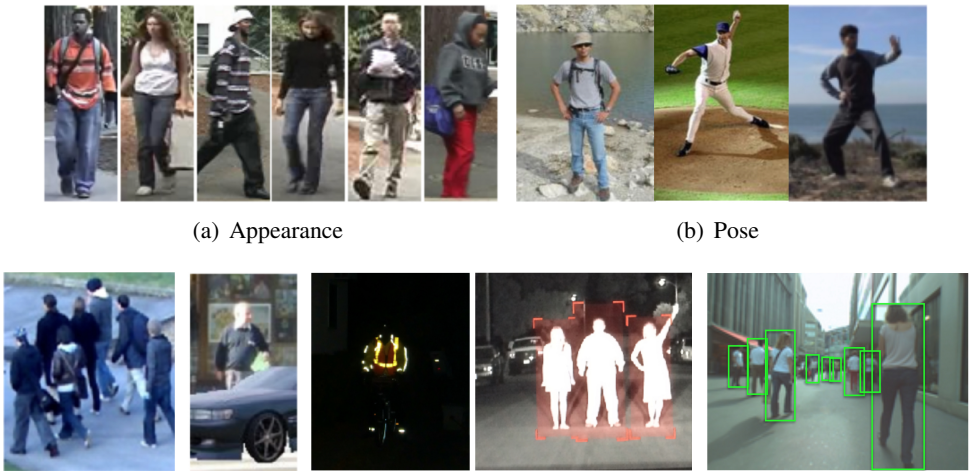
visual descriptions. In checkout-free stores, such as Amazon Go [4], results of human detection would be fundamental inputs for action recognition module. By recognizing the actions of a specific custom, a back-end system can automatically and simultaneously check out the products picked up by that person. Besides, the visual attributes of detected customs could also help in recognizing their identities, which enables the system to charge money from the right person. Moreover, potential customs and their shopping interest could be discovered by investigating how long they stay by a specific product.

1.1.1 Challenges

Despite various applications, many challenges have prevented current artificial vision systems from approaching human-level perception ability on identifying human objects in images [5] (as shown in Figure. 1.2), which include:

Appearance. Because of different clothing and carry-ons, people have highly diverse appearances. A good human detector should be capable of capturing these visual complexities, in order to discriminate human instances against a cluttered background. As a result, effective image representations and powerful classifiers are required.

Pose. Different from rigid objects, the human object is more articulated. They would have different poses and layouts in images, due to actions they are performing. Different camera views would also lead to people of various poses in images. Generally speaking, one unified and rigid human detector cannot cover these pose variants. For instance, a human detector trained with upright human samples apt to fail in detecting a person lying on a floor.



(a) Appearance (b) Pose

(c) Occlusion (d) Lighting (e) Scale

Figure 1.2: Challenges of vision-based human detection.

Occlusion. In many application scenarios of human detection, *e.g.*, in stations, streets, and grocery stores, we could observe lots of occlusions happening between people and people, or between people and other background objects. It is likely that the lower body or the left shoulder of a person is occluded in the image. A human detector should be robust to occlusions and output high detection confidence for a human instance even if it is partially invisible.

Scale. It is also called image resolution problem. First of all, human instances of small image scales have very low image resolutions. In other words, it is difficulty to extract informative visual features from these small image areas to represent human objects. Additionally, image features from different scales would have distinct visual statistics, which can bring challenges to a single detector in validating human instances of various scales, even if feature transformations are applied before the final classification.

Lighting. As we introduced above, human detection is a key technique for many around-the-clock applications, *e.g.*, self-driving car and video surveillance. Currently, most of the human detectors are using color images in detecting people. These detectors work well on daytime images captured under good lighting. Nevertheless, on night time images with human instances of bad visibility, these detectors would possibly be incompetent. Even on daytime images, traditional human detector would mistake people standing

Figure 1.3: The conventional pipeline of human detection.

in shadow as background. Such defect cuts these approaches off from real applications. Thereby, some other image sensing (*e.g.*, infrared images) robust to lighting has to be considered, in particular, when human detection is applied in 24-hour applications.

1.1.2 Conventional Pipeline

As a canonical case of the generic object detection problem, in the past decades, human detection has attracted consistent attention from the computer vision community. To improve the detection performance, researchers have proposed a bunch of methods to handle the challenges mentioned above and have significant improvement in recent years [6, 7, 9, 8, 10]. Basically, the pipeline of these conventional methods consist several key components, as shown in Figure. 1.3, including low-level sensing, preprocessing, feature engineering, and recognition.

Low-level sensing. According to the number of cameras used in imaging, human detection can be categorized into monocular human detection [11] and binocular human detection [12]. With stereo estimation [13, 14], depth information can be leveraged into this problem. Beyond color cameras, other types of sensors have also been extensively used in human detection, such as depth camera [15] (*e.g.*, Kinect), thermal camera [16, 17], and Lidar [18, 19]. Compared to RGB cameras, these sensors are less sensitive to bad lightings while the imaging from these sensors is short of visual details.

Preprocessing. It is traditional to perform some image preprocessing before the feature extraction step, such as scaling, contrast enhancement, and color normalization. Scaling is commonly applied in preprocessing to detect people of low image resolutions. It is usually implemented by bilinear interpolation of pixel values. By doing these, visual features, e.g., Histogram of Gradient features [6], can be extracted from small image areas, since the extraction of most visual features relies on receptive fields of minimum image size (e.g., 64). Actually, even for some deep neural networks (DNNs) based detectors, e.g., R-CNN detection [20], image scaling

is also a necessary step. Contrast enhancement is a transformation of the sensory representation which makes the value of imaging signal cover a wider range, such that the regions of transition (e.g., edges) could be selectively emphasized [21]. We know that edges are very useful low-level visual features to represent human instances. Color normalization has been used to remove all intensity values from the image while preserving color values. With respect to human detection, especially in video surveillance, it is important to remove shadows or lighting changes on the same color pixels and then extract consistent visual features [22].

Feature engineering is typically composed by feature extraction and transformation. The purpose of feature engineering is to convert an image region into a visual representation (a vector or a matrix). Ideally, such image representation of human should invariant to scale, rotation, light change, partial occlusion. A bunch of visual features have been proposed and extensively study for human detection, such as Haar-like features [23], histogram of oriented gradient (HOG) feature [6], local binary pattern (LBP) feature [24], integrated channel feature (ICF) [25], and spatial-pooled features [26]. Although achieving good performance, these handcrafted features need delicate design and they cannot leverage the power of large training data. During the past five years, vision and machine learning communities have witnessed great success of deep learning for speech recognition [27], image classification [28], and general object detection [20]. Deep neural networks have also introduced into human detection field [29, 30, 31]. We will give more details of the related work in Chapter 2.

Recognition. A human instance could be represented by a holistic feature vector or a part-based model. With global representations, different classifiers have been introduced to improve the performance of human detection, *e.g.*, linear support vector machine (SVM) [6], random forest [32, 33], and AdaBoost classifier [23]. In general, better classifier leads to superior performance, in terms of detection accuracy and testing time. Some others proposed elastic models to represent human, based on a composition of several body parts. These part-based models, *e.g.*, pictorial structure (PS) model [34], deformable part model (DPM) [7], and regionlet detector [35], are robust to partial occlusion and pose variations in detection.

Post-processing. Non-maximum suppression (NMS) [36] is commonly applied to the original outputs of a human detector, which could remove bounding boxes that significantly overlap

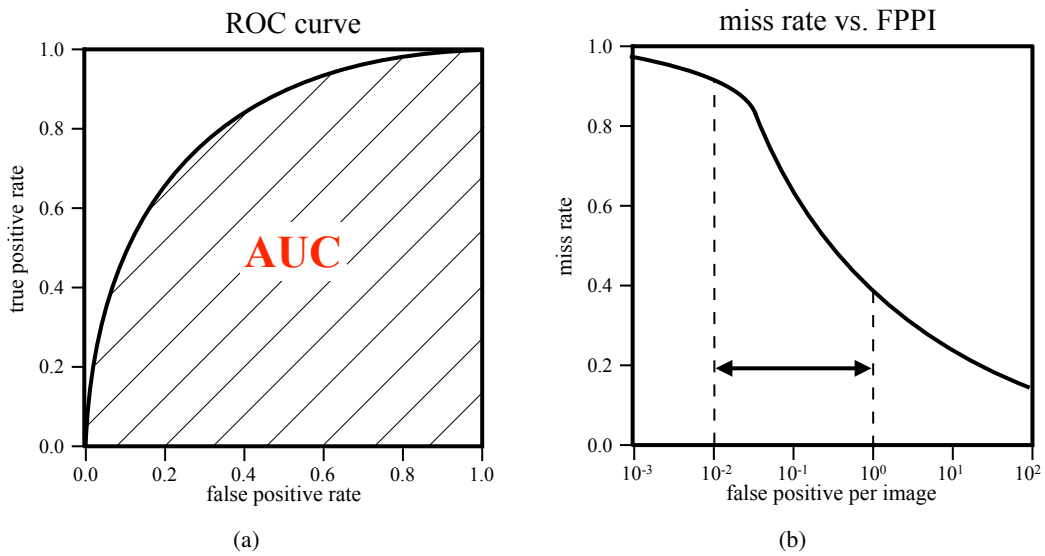


Figure 1.4: Plots of ROC curve and Miss rate vs. FPPI curve.

each other. It is a very important post-processing step, in order to avoid double detections which are regarded as false positives. NMS is a greedy algorithm, the parameter of which is set empirically and is usually sensitive to the density of people in images. In some scenarios, NMS would remove some true detections. Therefore, instead of using NMS, some others modeled post-processing as an optimization problem, *e.g.*, quadratic unconstrained binary optimization (QUBO) [37] or Latent SVM [38], which has demonstrated better performance in crowd scenes.

1.1.3 Evaluation Methodology

In terms of performance evaluation, Interaction-over-Union (IoU) is conventionally used in validating whether a detection is a true positive or a false positive. A human hypothesis is regarded as one true detection when the IoU against any ground truth annotation (represented by bounding box) is larger than a threshold (*e.g.*, 0.5). Otherwise, it is a false positive. The target of a human detector is to recall all true detections from images, in the meanwhile, ignoring any false positives. Apparently these two goals are exclusive and a good detector should balance the trade-off between the precision and the recall. Here, we introduce several evaluation criteria commonly used to assess a human detector.

F1-Score is a measurement which considers both the precision and the recall of detection outputs. The traditional F1-Score is the harmonic mean of precision and recall (as shown in Equation 1.1), while its variants with different focuses have also developed [39] .

$$\text{F1-Score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{1.1}$$

Receiver operating characteristic (ROC) curve is created by plotting the true positive rate against the false positive rate at various confidence threshold settings, as illustrated in Figure 1.4(a). It provides another access to investigate the comprehensive performance of a human detector. For instance, the area under the ROC curve (AUC) is frequently employed to compare different detectors. Additionally, ROC curve is also a useful reference in selecting the possibly optimal threshold. Such threshold is employed to distinguish true detections from false alarms, which could achieve the best trade-off between the precision and the recall.

Log-average miss rate (MR) was first proposed in [40], based on the miss rate versus false positive per image (FPPI) curve, as shown in Figure 1.4(b). It is the average value of the 10 evenly sampled miss rates in the FPPI range $[0.01, 1]$ which is considered as the most significant range in practice. MR is widely used to compare different human detectors on public pedestrian datasets, *e.g.*, INRIA pedestrian dataset [6] and Caltech-USA pedestrian dataset [41].

1.2 Contributions of the Dissertation

1.2.1 Human Detection by Deep Learning with Color-Thermal Imaging

Thermal imaging is helpful in detecting people under bad lighting, *e.g.*, standing in a shadow or at night time. It becomes more essential when human detection is applied in around-the-clock applications, such as self-driving systems. Most previous methods built human detectors by using color or thermal imaging independently, while ignoring the complementary potential between them. Thereby, it is necessary to study leveraging both color and thermal images (called multispectral image in this dissertation) simultaneously in constructing a human detector.

We exploit deep learning algorithm for multispectral human detection, which is the cutting-edge technology for a series of computer vision problems. We first show promising results on

human detection by using Faster R-CNN [42] with color images. Afterward, we thoroughly investigate the potential improvement for human detection by taking advantage of multispectral images rather than using only one image modality. We provide the more reliable ground truth for the test set of the KAIST multispectral pedestrian dataset (KAIST) [43] which is the largest multispectral pedestrian benchmark so far. The improved annotation brings in new discoveries on our proposed models and could further facilitate researcher to make more convincing conclusions. Along with the improved annotations of KAIST, we extensively study four distinct fusion architectures of deep convolutional neural network (ConvNet). Detections and proposals are evaluated in various scenarios, in terms of different lightings, scales, and occlusions. Our best fusion model significantly reduces the miss rate of baseline method Faster R-CNN by 16.2%, yielding 24.7% overall miss rate on KAIST. These results introduce useful insights into multispectral human detection. Besides, we discuss the bottleneck of the current framework, indicating some directions for future research.

1.2.2 Graph-Based Context Modeling to Optimize Human Detection

Exploiting contextual cues has been a key idea to improve human detection in crowded scenes. We develop a new framework to address the challenges for human detection by putting people in a global context and modeling their interactions, since in a crowded scene people usually form groups where they interact with each other, both geometrically and socially [44, 37]. By given the outputs of any underlying detector, our method could optimize the detections by enhancing true detection and suppressing false positives.

The main contribution of this approach is the application of graph-based label propagation to exploit contextual information for human detection. Compared to the structured learning framework which models contextual interactions only between local neighbors [45, 37, 46], our approach is clearly advantageous in that the graph-based propagation make interactions between any two nodes possible. Such a capability allows our approach to discover challenging weak human instances, even if they are not close to any strong detection. Based on the proposed context graph, we apply label propagation to discover weak detections contextually compatible with true (strong) detections while removing irrelevant false alarms. Four kinds of contextual information that incorporate both geometric and social contextual patterns are

used to construct the context graph, which are spatial context, scale context, social context, and layout context. We further propose a greedy-like technique, namely progressive potential propagation, to instance human hypotheses as true detections iteratively.

1.3 Outline of the Dissertation

The dissertation is organized as follows.

In Chapter 2, we review the relevant work on human detection, including conventional methods and DNN-based approaches. Human detectors with infrared input are discussed independently. Methods that facilitate deep learning with multiple modality fusion are also included.

In Chapter 3, we propose the multispectral deep neural networks which improves human detection with both color and thermal inputs. Extensive experiments are introduced, reported on the improved annotation of KAIST multispectral pedestrian dataset.

In Chapter 4, we propose the graph-based context modeling to enhance weak true detections and to suppress false positives, considering several contextual cues. The task is further modeled as an optimization problem which is solved by a greedy algorithm.

Finally, Chapter 5 contains the conclusions and several possible directions for the future research.

Chapter 2

Related Work

In this chapter, we briefly discuss some relevant work of this dissertation. We first introduce conventional human detectors, grouping them into three categories of different research focuses. Second, recent effort on deep neural network (DNN) based human detection is studied. Besides, we investigate some human detectors that take an infrared image as the input. Finally, we review how multimodal information is exploited in previous methods which leverage deep learning to tackle with different data sources.

2.1 Conventional Human Detection

Conventional people detectors localize human instances in color images, using vision techniques before deep learning introduced. We categorize these methods into three groups, each of which has different technical focuses.

Image Representation. To build up a robust human detector, the image representation of people has to be competent of differentiating human instance with other background. Such visual features should be invariant to different appearances, poses, scales, and dynamic light changing. Viola *et al.* [23] proposed Haar-like wavelet features and took advantage of the integral image for rapid feature computation. Based on such Haar-like feature, Sabzmeydani *et al.* [47] learned shapelet as mid-level features, which are more informative than low-level features in discriminating pedestrian from non-pedestrian objects. Dollár *et al.* [48] proposed an extension of [23], namely integral channel feature (ICF), where Haar-like features are computed over multiple channels of visual data, including LUV color channels, gray scale channel, gradient magnitude, and gradient magnitude quantized by orientation. Some variants of ICF were also developed to achieve better performance, such as aggregated channel feature (ACF) [25], SquaresChnFtrs filter [10], and LDCF filter [49]. Leibe *et al.* [50] employed SIFT feature [51]

to capture the local structure of pedestrian and then applied Hough voting to localize people. Ma *et al.* [24] learned discriminative local binary pattern (LBP) feature in personal albums for human detection. Contour information is also very useful in depicting human objects. Zhao *et al.* [52] used head-shoulder shape detector to generate human hypotheses. Leibe *et al.* [50] introduced chamfer matching [53] to add global shape constraint in verifying human detections. Most modern detectors used some form of gradient of histogram as human feature, derived from the pioneering work in [6] which proposed the histogram of oriented gradient (HOG) feature. For instance, Wang *et al.* [54] combined HOG with LBP feature in their human detector. Besides, nearly all part-based models used HOG feature to represent body parts [7, 55].

After feature extraction, different classifiers are trained to distinguish human from background objects, *e.g.*, linear SVM [6], random forest [32, 33], partial least squares analysis [56], and AdaBoost classifier [23, 57, 47]. Reviews of more related methods could be found in these nice surveys [9, 58].

Elastic Model. To deal with partial occlusions and poses, different elastic human model has been proposed, including pictorial structure (PS) model [34], hierarchical part-template matching model [59], DPM model [7], poselet detector [55, 60], flexible mixture model of parts [61], and regionlet detector [35]. In general, these models represent human with a mixture of body parts and implement detection by a bunch of part detectors. The kinematic dependencies between parts were commonly modeled by spatial constraints [34, 61, 61] or hierarchical part tree [59]. Depending on whether a whole body is first detected or not, these methods can further be separated into two categories: top-down [59, 62, 34, 7, 61, 35] and bottom-up [55, 60]. Top-down methods score each human hypothesis by considering both the detection score of each part and their deformation cost with respect to the location of the hypothesis. In opposite, bottom-up methods first detect body parts and then project the locations of these parts to a point which represent the center of a human body.

Context Modeling. Context here refers to the information beyond the visual features of a human instance. Ding *et al.* [63] have discovered that the enlarged image area of human which covers some background performs better in detecting people [63]. Image features extracted from such enlarged area could boost detection confidences and improve the localization capability of a human detector. A similar idea has also been exploited in [64] which used deep neural

network for human detection. Ouyang *et al.* [65] trained different human detectors, according to different layouts of occluded people, exploiting context from nearby human instances. However, these methods leveraged context information implicitly from low-level features, ignoring explicit context on instance-level which has achieved promising results on generic object detection [45, 66]. In fact, we could observe different kinds of contextual interactions between human instances, especially in crowd scenes, such as spatial proximity and scales similarity. Yan *et al.* [38] modeled the appearance and spatial interaction between human hypotheses as a maximum a posteriori (MAP) problem and approximated a solution by a greedy algorithm. Rujikietgumjorn [37] formulated multiple pedestrian detection as the quadratic unconstrained binary optimization (QUBO), which modeled spatial interactions into a quadratic loss function. Idrees *et al.* [67] incorporated scale information in Markov Random Field (MRF), in order to enforce a locally-consistent scale prior. Stewart *et al.* [68] modeled the instance-level interactions in the hidden units of long short-term memory (LSTM) [69] and predicted human hypotheses in one image in a sequential manner.

2.2 Infrared Image Based Human Detection

As mentioned in Chapter 1, infrared (IR) cameras are insensitive to bad lightings. There have been a couple of human detectors that exploited IR images as input. Despite different modality from RGB images, most of these detectors were inspired by methods developed for visible imaging. Actually, nearly all of these IR-based detectors used conventional image features to represent people, such as image density [70], LBP features [71, 72], HOG features [71, 73, 74, 75, 76, 77], shape template [16, 78], and ICF features [77], along with various classifiers. DPM [7] was also introduced in [75, 76], for the purpose of handling occlusions and poses. Overall, IR-based human detectors haven shown robust performance during night time, yet neglecting the complementary potential between infrared and visible channels that would boost detection results further.

Recently, researchers are becoming interested in devising human detectors by multispectral (color and infrared images) images rather than one image modality. Krotosky *et al.* [79] generated five-channel images by registering color, disparity and infrared inputs, and extracted

HOG feature from each channel to train a multispectral human detector. The similar idea was applied in [80], where gradient-based features were combined with a part-based model. Lee *et al.* [81] used thermal images to extract foreground image areas regarded as proposals of human instances. González *et al.* [82] studied the accuracy gain of different shallow models by using multispectral inputs. The most relevant work was proposed in [83]. They introduced the R-CNN framework [20] for multispectral human detection and employed the ACF+T+THOG detector [43] in both training and testing to generate human proposals. However, end-to-end training of ConvNet-based detector is worth to probe, and we need to investigate more fusion architectures that could employ multispectral inputs better.

2.3 Deep Neural Network Based Human Detection

As achieving great success in many computer vision tasks, deep neural network (DNN) has also been introduced to build human detectors. One pioneer work on pedestrian detection was proposed in [29], which trained convolutional neural networks (ConvNets) layer by layer using unsupervised learning. Then convolutional features from multiple stages and scales were extracted from warped sliding windows for classification. In [84], Ouyang *et al.* modeled visibility relationships among pedestrians using DNNs, which improved the visual confidences of human parts. Tian *et al.* [85] trained 45 independent part detectors with weakly annotated humans, in order to handle partial occlusions. Tian *et al.* [86] improved human detection by learning high-level features from DNNs along with multiple tasks, including people attribute prediction. Angelova *et al.* [87] built DNNs cascades that filtered candidates by tiny DNNs and speed up detection to 15 frame per second. Li *et al.* [88] proposed scale-aware DNNs with a scale gate function, which sent human hypotheses of different image sizes into different networks for classification.

Recent DNN-based methods (*e.g.*, [30, 5]) formulated human detection as a classification problem. Generally speaking, these methods followed R-CNN framework [20] which were consisted of two components (stages): a proposal generator and a ConvNet-based classifier. At the first stage, an external human detector is first applied to the image, generating human proposals represented by bounding boxes. Then these proposals are passed into

ConvNets at the second stage, to extract deep features for classification. Along with this research stream, many external human detectors and ConvNet architectures have been examined. Hosang *et al.* [30] tested different proposal generators, *e.g.*, ACF [25], SquaresChnFtrs [10], Katamari [10], and SpatialPooling+ [26], combined with two ConvNet models, *i.e.*, CifarNet [30] and AlexNet [28]. They discovered that better proposals and deeper neural networks would lead to better performance of human detection. Zhang *et al.* [5] studied more proposal generators, *i.e.*, LDCF [49] and Checkerboards [89], and compared the VGG16 model [90] with AlexNet sequentially, validating Hosang's observation.

Apparently, these methods rely on the qualities of proposals produced by an external human detector. These proposals not only decide the testing phases but also influence the qualities of training samples. In other words, a good proposal generator should produce hard negatives for training and enough positives for testing. Inevitably, the external human detectors largely affect the final performance of ConvNet-based detector. Differently, Faster R-CNN framework [42] trains ConvNets-based detectors in an end-to-end fashion. Compared to R-CNN, Faster R-CNN reduces the testing time by 250 times, only requiring 0.2 seconds in implementing detection on an image of 600 pixels. Faster R-CNN was first applied to pedestrian detection in [91] and further improved in [31] to tackle small human instances and negative hard examples.

2.4 Deep Learning with Multimodal Inputs

It is an essential challenge in many vision problems to integrate multimodal data sources. A bunch of DNN-based multimodal models have been developed, involving in data sources of different modalities, such as image *vs.* audio [92], image *vs.* text [93, 94], image *vs.* video [95, 96], etc. In [92], Naiam *et al.* learned a hidden layer from deep belief networks (DBNs) which represented the shared features of videos and audios. A similar network has also been applied in [93], to tolerate modality missing in image classification and retrieval. Wang *et al.* [94] imposed structure-preserving constraints into their similarity objective function for images and texts, achieving better phrase localizations in images. For video recognition, Simonyan *et al.* [95] proposed the two-stream ConvNets that simultaneously incorporated spatial and temporal information, while the optical flow ConvNet was combined with the key-frame based

ConvNets. The combination of color and depth images has also been exploited for 3D object classification [97] and 3D object detection [98] by using deep ConvNets. Most of the aforementioned methods used two-branch networks and then fused features from different channels at the very end. In other words, they implemented last feature layer fusion [97, 98, 94, 99] or confidence fusion [95]. Karpathy *et al.* [96] discussed more fusion schemes for video classification, integrating key frame features with video segment based features .

DNNs-based multispectral human detection has not been studied thoroughly. This problem is different from other vision problems in that, in some cases of unusual lighting, features extracted from color or thermal images could not work well independently for human detection. As a result, decision fusion at the very end stage seems not the straightforward solution for multispectral human detection. It is worthy of studying how color and thermal images could be fused properly in DNNs, in order to obtain 'optimal' synergy for human detection.

Chapter 3

Multispectral Deep Neural Networks

3.1 Introduction

Human detection is the principal technique for various applications, such as surveillance, tracking, autonomous driving, etc. However, many challenges, including occlusion, low image resolution, and cluttered background, prevent artificial vision systems from approaching human-level perception ability on identifying human objects [5]. To improve the performance of human detection, researchers have proposed abundant methods to handle these challenges, focusing on designing discriminative visual features [6, 100, 8, 101], devising robust human classifiers [7, 102, 33, 103], and leveraging context information [104, 105]. Although people have made numerous effort in this area and have achieved significant improvement in recent years [9, 10], there still exists an insurmountable gap between the current machine intelligence and the requirement of around-the-clock applications. Particularly, most of the current human detectors explored color images captured under good lighting and are thereby very likely to fail on images captured at night, due to the bad visibility of human instances. Such defect would cut these approaches off from the real applications, e.g., self-driven car and surveillance system.

In fact, the aforementioned difficulty not only exists in human detection but also presents in many other vision problems that have to handle cases with bad lighting. To deviate such inherent drawback of visible imaging, researchers are becoming to take advantages of invisible imaging captured by sensors of different spectrum, including depth cameras (time-of-flight, or near-infrared, *e.g.*, Kinect) and thermal cameras (long-wavelength infrared). Since ambient lighting has less effect on thermal imaging compared with near-infrared imaging, thermal cameras are widely used in various vision problems, such as saliency detection [106], human tracking [107], face recognition [108], and activity recognition [109]. When it comes to human detection, thermal images usually present clear silhouettes of people [110, 111] which

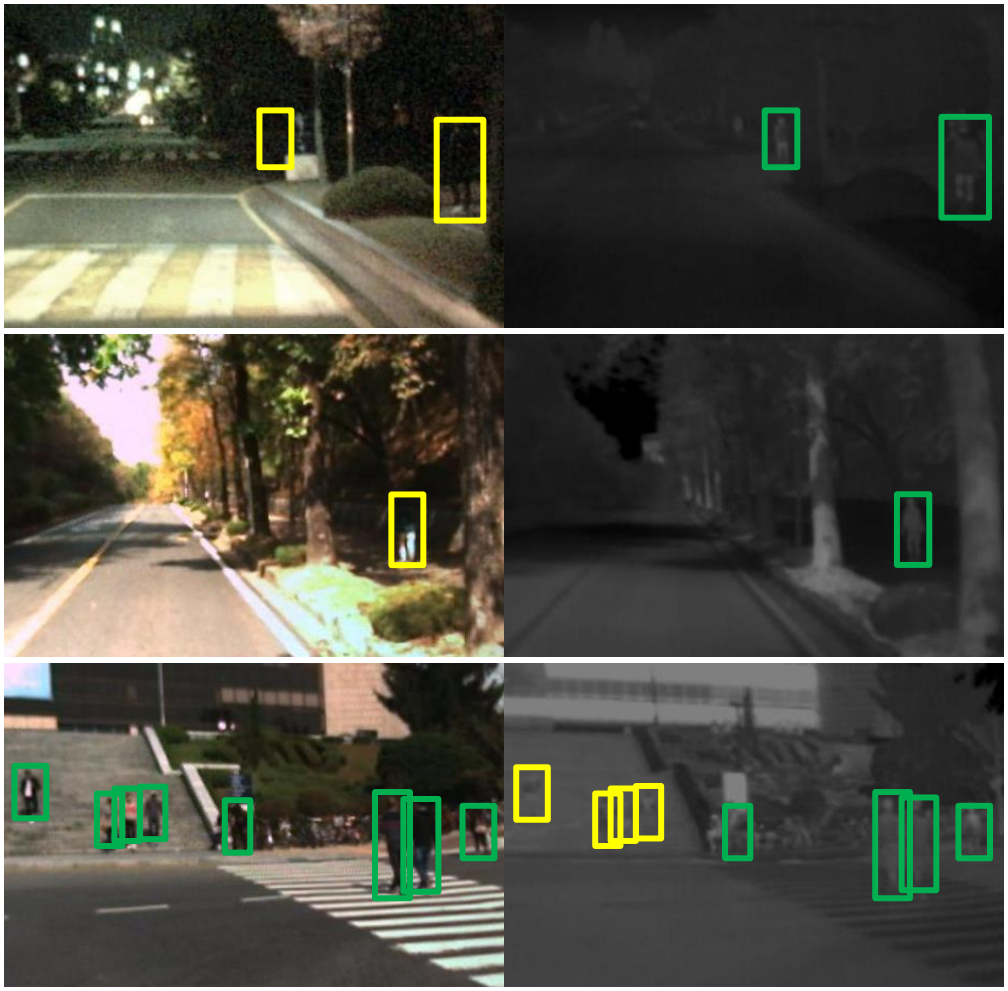


Figure 3.1: Yellow bounding boxes indicate detection failures with one image channel. Top and middle: the thermal images capture human shapes even in bad lighting, while the corresponding color images are messed up. Bottom: with bright background, color image provides more distinctive visual features for the pedestrians (standing on stairs) against background objects; in such scenario, human silhouettes in the thermal image are ambiguous.

is known as discriminative visual features in differentiating human objects from other background objects [52]. On the other hand, fine visual details of human objects (*e.g.*, clothing) are missing in thermal images which are useful in detecting human. Besides, a bright background would corrupt the thermal imaging of human in some cases during daytime [112]. As shown in Figure 3.1, if only one image modality is exploited for detection, human instances of yellow bounding boxes would be mistaken as background. Based on these observations, it is not difficult to conclude that one image modality (either color or thermal) cannot capture the image characteristic of human objects well.

In some sense, color and thermal image channels provide complementary information in depicting pedestrians. Therefore, it is reasonable to expect that when one image modality does not work or it is indecisive, the other one could help in the detection task; information from the two channels would lead to more reliable decisions on validating human hypotheses. Nevertheless, most of previous approaches (*e.g.*, [71, 74, 78, 72]) that considered infrared imaging for human detection only leveraged one image modality, without investigating both visible and non-visible modalities simultaneously. Some very recent research has shown that multispectral image [1] outperforms RGB image on pedestrian detection [80, 43, 82]. However, nearly all of these methods used hand-crafted image features to represent people, *e.g.*, histogram of oriented Gradient (HOG) [6] and aggregated channel feature (ACF) [8, 49], and trained shallow models to distinguish human objects with backgrounds, such as linear SVM and random forest, etc. In fact, compared to deep convolutional neural networks (ConvNets), handcrafted features and shallow models restrict these multispectral detectors from obtaining better performance.

During the past five years, vision and machine learning community have witnessed great success of deep neural networks (DNNs) [113] applied to image classification [28] and generic object detection [114]. Therefore, it becomes very natural and appealing to exploit the effectiveness of DNNs for multispectral human detection. Apparently, multispectral detection is a vision problem with multimodal inputs. Actually, there have been a couple of vision applications, involving in multimodal inputs, *e.g.*, action recognition [95], image classification [93], computer-aided diagnosis [115]. DNNs have shown significant achievement on these tasks by using multimode inputs than a single modality. However, it is still unknown how color and thermal images can be fused properly in DNNs to achieve the 'best' detection performance .

As shown in Figure 3.1, color and thermal images could help each other in different scenarios. They have consensus on detecting human to some extent, but not always. Thereby, it is not straightforward to make the most of the multispectral DNNs for human detection. Deep ConvNet was firstly introduced for multispectral pedestrian detection in [83]. In this work, Wagner *et al.* followed R-CNN [114] framework and studied several fusion architectures based

[1] In this dissertation, the multispectral image refers to the image of color and thermal channels.

on AlexNet [28]. However, such 2-step training pipeline requires external approaches to generate proposals, which is not competitive as end-to-end training of object detectors (we will discuss more in Section 3.2 on R-CNN). Consequently, it needs more effort to explore the end-to-end training framework for human detection, which could most effectively utilize both color and thermal images. In addition, other architectures of deep ConvNets should also be extensively investigated.

In our work, we adopt Faster R-CNN [42] as our vanilla ConvNet, which consists of a Region Proposal Network (RPN) and a Fast R-CNN detection network [116]. Faster R-CNN model uses RPN to generate proposals, rather than using an external proposal generator. Since RPN shares convolutional features with Fast R-CNN network, it does not need much extra computational overload. Such property also facilitates the end-to-end training of an object detector. Compared to R-CNN, Faster R-CNN has demonstrated superior performance on generic object detection [117], in terms of speed and accuracy. We validate our vanilla ConvNet on Caltech pedestrian benchmark [40] by training a human detector with color input and achieve state-of-art miss rate, outperforming R-CNN.

Now the question is whether multispectral images could ameliorate human detection with a substantial degree, even when a strong ConvNet-based detector is used. To quantitatively analyze the potential gain rather than taking it as granted, we train two separate ConvNet-based detectors using color and thermal images, respectively. Afterward, the study is implemented on KAIST multispectral pedestrian dataset (KAIST) [43]. Based on the detection results of the $2,252$ testing image, we reveal that there is a large margin in improving human detection by leveraging multispectral images, especially for around-the-clock applications. The problem of multispectral human detection then becomes to exploring the most effective deep ConvNet architecture that utilizes color and thermal images simultaneously. We treat this challenge as a ConvNet fusion problem. Motivated by the assumption that fusions on different DNNs stages would lead to distinct detections, we carefully design four ConvNet fusion architectures upon our vanilla ConvNet and then provide extensive experiments on these fusion models.

The primary contribution of this work is fourfold:

- We first report pedestrian detection results of Faster R-CNN on Caltech pedestrian benchmark, achieving a state-of-art performance (17% miss rate).

- We improve the ground truth of KAIST test set of $2,252$ images, with more annotations on valid human instances and less incorrect background labeling. KAIST dataset is the largest multispectral pedestrian benchmark so far. The improved annotation brings in new discoveries on our proposed models and could further facilitate researcher to make more convincing conclusions.

- We carefully design four distinct ConvNet fusion architectures that integrate two-branch ConvNets on different DNNs stages, corresponding to information fusion on low level, middle level, high level, and confidence level. All these models outperform the strong baseline detector (*i.e.*, Faster R-CNN) on KAIST dataset. Our best fusion model significantly reduces the miss rate of Faster R-CNN by 16.2%, yielding 24.7% overall miss rate on KAIST test set.

- Along with the improved annotations on KAIST, we extensively study the four ConvNet fusion models. Detections and proposals are evaluated in various scenarios, in terms of different lightings, scales, and occlusions. These results introduce useful insights into multispectral human detection. We discover that different fusion models are preferred in diverse scenarios. Discussion on the bottleneck of the current framework is also contained, inspiring some directions for future work.

3.2 Vanilla ConvNet

Most of deep ConvNet based human detectors followed R-CNN framework [30, 5]. The pipeline of R-CNN based human detector is illustrated in Figure 3.2. Basically, these detectors contains two stages. At stage one, an external proposal generator (*e.g.*, selective search [118] and Edge boxes [119]) is used to produce proposals, each of which corresponds to an image region. To achieve high recall, about $2,000$ proposals are typically generated. At stage two, R-CNN classify the $2,000$ proposals into different object categories by using deep ConvNet. Overall, R-CNN based detectors are very time-consuming. First of all, selective search at stage one commonly takes around 2 seconds to produce $2,000$ proposals. Moreover, convolutions on $2,000$ image regions requires a huge amount of computational time since computation on overlapped image regions is indeed redundant.

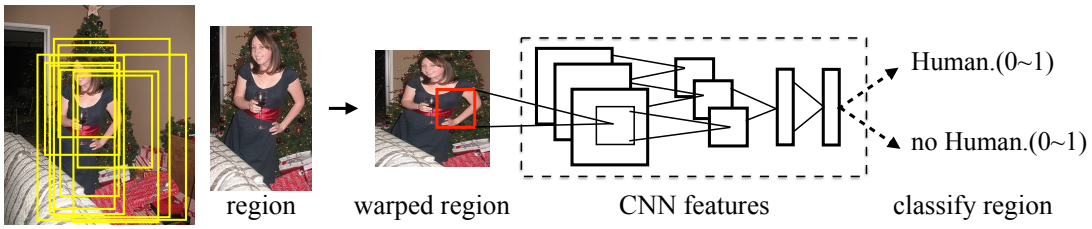
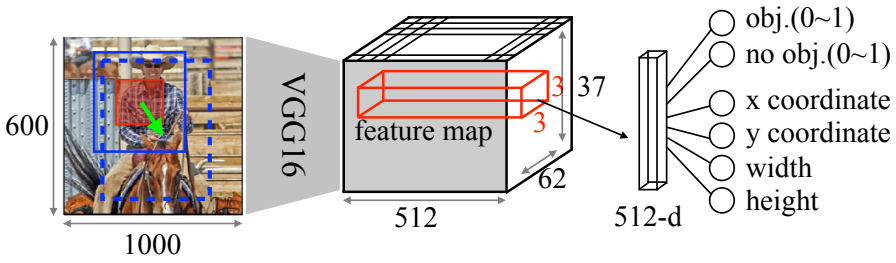


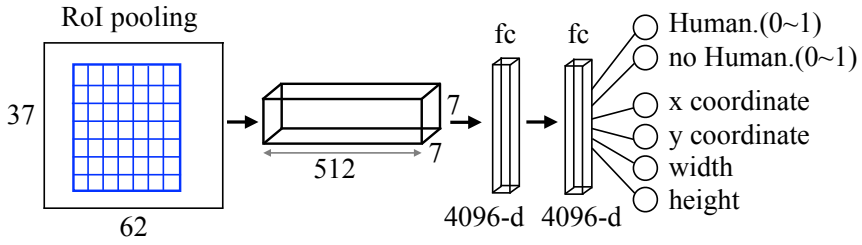
Figure 3.2: Pipeline of R-CNN based human detector. (a) Stage one: generate human proposals. (b) Stage two: classify image regions of proposal as human object or not.

Inspired by the recent success on generic object detection, we consider starting with Faster R-CNN [42] and verify its performance on Caltech pedestrian benchmark [40]. Faster R-CNN model consists of a Region Proposal Network (RPN) and a Fast R-CNN detection network [116]. Literally, RPN is used to generate proposals. As shown in Figure 3.3(a), it is a fully convolutional network that shares features with the detection network. On top of VGG16 model [90], an additional $512 \times 3 \times 3$ filter is applied, to generate one 512-d feature for every pixel location (x, y) on the last convolutional feature map. To be noticed, each 3×3 convolutional filter corresponds to a local area of the original image. Then these 512-d features are used to decide whether their corresponding local images belong to an object or not. Besides, a regression model is also introduced using the 512-d features as well, in order to obtain proposals of better localizations. In more detail, for an anchor (x, y, w, h) on the last feature map, the regression model outputs (dx, dy, dw, dw), which could shift and resize the original anchor. As illustrated in Figure 3.3(a), features from the red image areas could infer the proposal of dotted blue bounding box. To deal with multiple object sizes and layouts, anchors of 3 scales $(8 \times 8, 16 \times 16, 32 \times 32)$ and 3 ratios $(0.5, 1, 2)$ are considered at each pixel location (x, y). After ranking all of these local image areas, anchor locations of top objectness (typically 300) are regarded as proposals.

As we mentioned before, the receptive field of features in RPN commonly corresponds to the local part of an object. For example, it is possible that the feature of a left shoulder is used to deduce whether a true human exists and to infer the whole body location. Obviously, RPN in Faster R-CNN just thoroughly provides human candidates. In order to achieve better detections, refined feature have to be extracted which captures more object area for making decision. For



(a) Region Proposal Network (RPN). Red bounding box shows the receptive field of one 3×3 filter. Blue bounding box indicates the image region mapped by one square anchor, which is further shifted and resized after regression (the dotted blue bounding box).



(b) Fast R-CNN detection network. RoI pooling layer can be regarded as one variant of max pooling layer. It first divides a feature field into a grid of 7×7 cells and the maximum value in each cell is extracted.

Figure 3.3: Framework of Faster R-CNN.

this reason, proposals from RPN are passed into the Fast R-CNN detection network.

Fast R-CNN detection network was first proposed in [116], as shown in Figure 3.3(b), whose core part is the RoI (Region of Interest) pooling layer. Given the locations of proposals, RoI pooling layer projects the features of the last convolutional feature map into fixed size (*i.e.*, $512 \times 7 \times 7$) . Different from SPPNet [120], RoI pooling layer is differentiable, which enables back-propagation during training. Being connected to two more fully-connected layers, these features are used in differentiating a proposal as a human object or not. A regression model is also employed, targeting better localization results.

Compared to its methodological ancestries (*i.e.*, R-CNN [114]) that rely on independent proposal generators, Faster R-CNN generates proposals by RPN. Thereby, Faster R-CNN model trains ConvNet-based detector in an end-to-end fashion, which is getting appreciated by researchers due to its superior performance. More importantly, RPN shares features with detection network, which enables nearly cost-free region proposals. Consequently, Faster R-CNN takes 0.2 seconds to detect objects on an image of 600 pixels, while R-CNN based detectors require around 50 seconds. Moreover, compared to R-CNN pipeline (*e.g.*, [30, 5]) that has to

resize proposals, RoI Pooling layer extracts the features of a fixed size for any proposals, thus Faster R-CNN could handle human objects of arbitrary sizes.

Implementation details: We adapt the original Faster R-CNN model with a few twists and use it as our vanilla ConvNet. First of all, we remove the fourth max pooling layer of the very deep VGG16 model. Since each max pooling layer down samples feature map by half, such modification results in a four times lager feature map. This is encouraged by the observations in [88] that larger feature maps are beneficial to detecting human of small image sizes. Besides, original Faster R-CNN uses reference anchors of multi-scale ($\times 3$) and multi-ratios ($\times 3$) to predict locations of region proposals. Given the typical aspect ratio of people (they are upright most of the time), we discard the anchor ratio 0.5 in our vanilla ConvNet, to accelerate the training and testing of RPN.

Caltech$\times$10 training set is used for fine-tuning. We exclude occluded, truncated, and small (< 50 pixels) pedestrian instances, resulting in around $7,000$ training images. Since Faster R-CNN regards one image as a training mini-batch, these $7,000$ images are guaranteed to contain at least one valid pedestrian instance, such that each mini-batch includes positive samples for training. Proposals of intersection-over-union (IoU) with any ground truth annotation larger than 0.5 are regarded as positive training samples, otherwise, as negative samples. The vanilla network is initialized by the VGG16 model pre-trained on ImageNet database [121]. We fix the parameters of the first two convolutional layers of the VGG16 model and fine-tune other layers with Stochastic Gradient Descent (SGD). Since the lower layer extracts low-level features, such as, edges, corners, and lines, we speculate that the first two layers pre-trained with millions of images have been well trained for this task. Hence, it is unnecessary to update them further. Following the alternative training routine of Faster R-CNN, we fine-tune the networks for about 6 epochs. Learning rate (LR) of parameters is set to 0.001 at first and then reduced to 0.0001 after 4 epochs. Single image scale (600 pixels) is employed, without using image pyramids nor feature pyramids.

Comparison of Detections: We compare our vanilla ConvNet (denoted as FasterRCNN in Figure 3.4) with some other methods reported on Caltech test set, including HOG [6], DBN-Mut [84], SpatioPooling [101], LDCF [49], Katamari [10], Checkerboards+ [89], TA-CNN [86], FastRCNN [116], and DeepParts [85]. For FastRCNN, we use ACF pedestrian detector [8] to

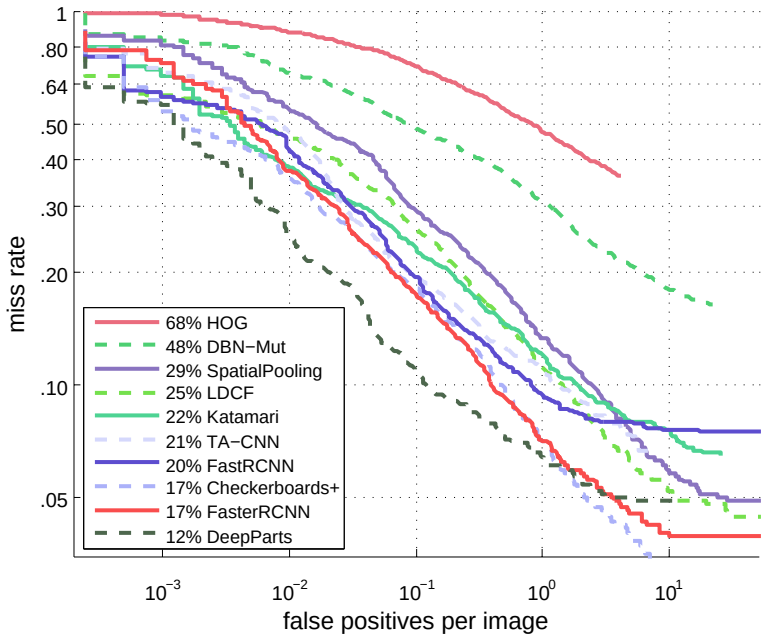


Figure 3.4: Comparison of detection results reported on the test set of Caltech pedestrian benchmark. Our vanilla ConvNet achieved 17% MR.

produce human proposals which are used to train and test Fast R-CNN detection network. A low threshold (*i.e.*, -50) is set for ACF detector, in order to gain high recall of proposal. We use IoU on ground truth to validate detections, *i.e.*, a hypothesis of IoU larger than 0.5 is regarded as true positive (TP). Detection performance is then measured by log-average miss rate (MR) over the range of $[10^{-2}, 10^{0}]$ (as discussed in Chapter 1.1.3) under reasonable configuration [9]. Lower MR indicates better results. It is easy to conclude from Figure 3.4 that our vanilla ConvNet beats most of the state-of-art approaches that depend on sophisticated features crafting or network designs. With completely data-driven and end-to-end training, we achieve 17% MR, which is lower than the ConvNet-based methods, such as DBN-Mut (48%), TA-CNN (21%), and FastRCNN (20%). One leading performance (12% MR) on this benchmark is reported by DeepParts detector which is an assembly of 45 ConvNets-based part detectors. Apparently, this method requires much more computational overhead than our vanilla network. Besides, annotations on body parts are needed in training.

Given its state-of-art performance and other important advantages, *e.g.*, end-to-end training and capability of handling human instances of arbitrary sizes, in our paper, the vanilla ConvNet is used as the fundamental ConvNet architecture in designing multispectral detectors.

(a) (b) (c) (d)

Figure 3.5: Improved annotation of selected frames from KAIST×30 test set, compared to the original ones. (a) Removed *pedestrians* labeling that are background objects; (b-d) Additional annotations (either *pedestrian*, *people* or *person?*) of valid human objects, with tighter bounding boxes.

3.3 Complementary Potential

Intuitively, color and thermal channels offer complementary visual information. However, a quantitative study on such complementary potential is necessary, in order to approximate the improvement by leveraging both color and thermal images for human detection. In this section, we answer two following questions: 1) when strong ConvNet-based detectors are involved, whether color and thermal images still provide complementary information. 2) To what extent the improvement should be expected by using such multispectral inputs. These analyses are implemented on KAIST multispectral pedestrian dataset (KAIST) [43], while ground truth of test images are corrected. Without doubts, better annotations are critical in making credible comparisons and convincing conclusions.

3.3.1 Improved Annotations of KAIST Test Set

As a standard routine, $2,252$ test images are sampled from test videos of KAIST benchmark with 30-frame skips [43]. We denote them as KAIST×30 test set. Each pedestrian in KAIST×30 test set was annotated by using Piotr's Computer Vision Toolbox [122]. However, only human instances in selected video frames were manually labeled, while the ground truth bounding boxes of others were yielded by interpolations between annotations in consequential frames. In short, not all of images in KAIST×30 test set were manually annotated by human.

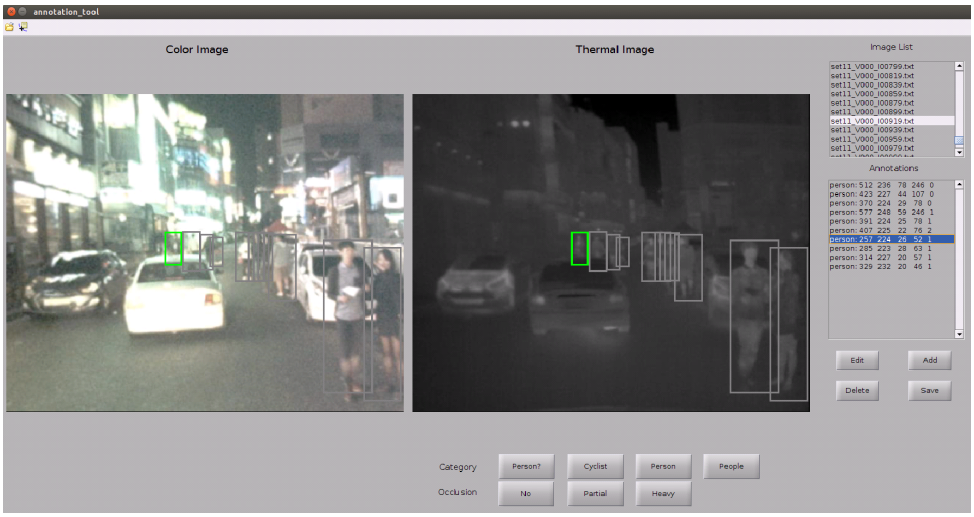


Figure 3.6: Interface of annotation tool.

Therefore, some problematic annotations can be found in KAIST×30 test set, *e.g.*, missing annotation of valid human instances and incorrect labeling on background objects, as illustrated in the top row of Figure 3.5. Obviously, such mistakes in ground truth would lead to error-prone evaluations on human detectors.

To obtain reliable ground truth, we develop an annotation tool in MATLAB which enables us to manually label every image in KAIST×30 test set, rather than skipping some of them. The interface of our annotation tool is demonstrated in Figure 3.6. Both color and thermal images are simultaneously displayed to annotators since some human instances of bad lighting are hardly observable in one image modality. Annotations of each image are listed on the left and illustrated on the images. The chosen human instance is shown in green bounding box for checking, while others in gray. Following the annotation protocol proposed in [43], each human instance is labeled as one of the four categories: *pedestrian*, *people*, *cyclist*, and *person?*. We add missing labeling, remove incorrect annotation, and refine ground truth bounding boxes of human objects in KAIST×30 test set. Their occlusion levels are also double-checked. To be noticed, the qualities (tightness) of bounding boxes influence the assessment of pedestrian detectors, with regard to localization capabilities.

The statistical differences between the original annotations of KAIST×30 test set and our improved ones are presented in Figure 3.7. In terms of categories, the improved annotation has $1,145$ more *pedestrian*, 183 more *people*, 28 more *cyclist*, and 151 *person?* instances. In

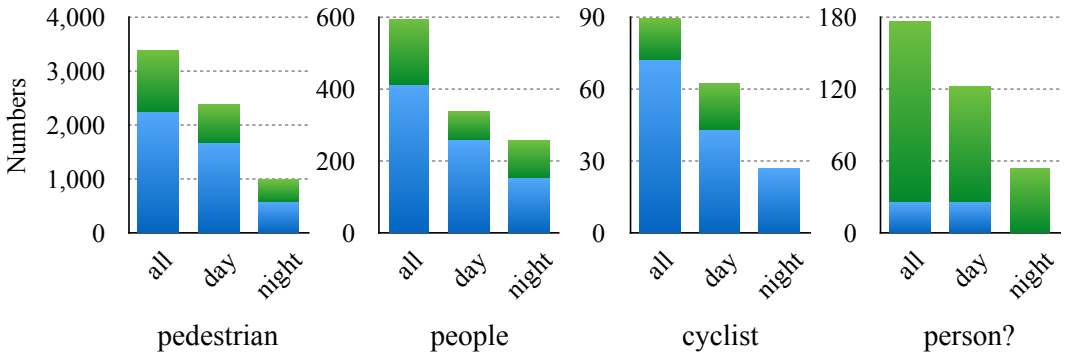


Figure 3.7: Statistical differences between the original and our improved annotations. The numbers of annotations, in terms of categories. Blue bars represent the original annotation, and green bars as additional human instance in our annotation.

fact, *people*, *cyclist*, and *person?* instances are also critical to fair comparison since human objects of these categories have to be ignored in the evaluation. If their annotation is missing, detections on these instances would be mistakenly regarded as false positives. With regard to scale, as shown in Figure 3.8(b), we have 583 new *medium* pedestrians, 603 additional *far* pedestrians. Besides, 42 *near* pedestrians are removed due to tighter bounding boxes[2]. Such scale definition helps in evaluating human detectors on detecting instances of different image resolutions. Figure 3.5 (bottom row) shows the enhanced ground truth of human instance in some sample frames, compared to the original annotations (top row) released by [43].

Eventually, we have $3,390$ *pedestrian*, 597 *people*, 90 *cyclist*, and 177 *person?* instances on KAIST$\times$30 test set. The $3,390$ pedestrians include $2,383$ instances from daytime images, and the other $1,007$ captured at night time. As illustrated in Figure 3.8, we have 261 *near*, $2,295$ *medium*, and 834 *far* pedestrian instances. Moreover, $1,400$ frames out of the $2,252$ ones have pedestrian annotations, among which about $1,000$ frames contain $1-2$ pedestrians and 38 images have more then 8 pedestrian instances[3]. The improved annotation of KAIST$\times$30 test set are used in all the following evaluations.

[2]The scale criteria of KAIST is different from Caltech benchmark: *far* (less than 45 pixel), *medium* ($45-115$ pixel), *near* (115 pixel or more).

[3]We released the improved annotations of KAIST$\times$30 to vision community for future use: http://paul.rutgers.edu/ jl1322/multispectral.htm

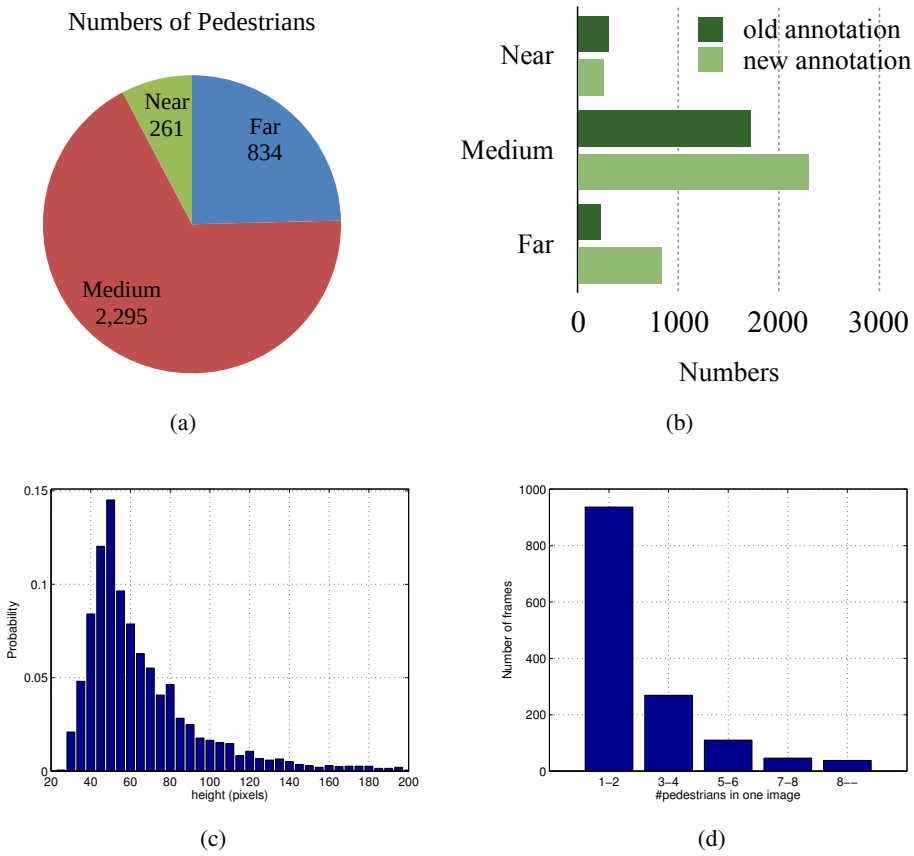


Figure 3.8: Statistics of improved annotation on KAIST×30 test set. (a) Numbers of *pedestrian* instances of different scales. (b) Number changes of *pedestrian* instances, in terms of *Near*, *Medium*, and *Far* scales. (c) Height (unit: pixel) distribution of *pedestrian* instances. (d) Number of *pedestrians* per frame.

3.3.2 Complementary Potential

To study the complementary potential between color and thermal channels, we first train two separate human detectors with either color or thermal images, depending on our vanilla ConvNet, namely FasterRCNN-C and FasterRCNN-T. The training set of KAIST (more details on this dataset will be given in Section 3.5) is used in fine-tuning these two ConvNets. Scheme of training these two detectors is similar to that in Section 3.2.

We only consider detections of more than 0.5 confidence scores in this potential analysis. Detections of IoUs with any ground truth (GT) larger than 0.5 are regarded as true positives (TPs), otherwise as false positives (FPs). Multiple detections on the same human are treated as FPs. In Table 3.1, we enumerate the numbers of GTs, TPs, and FPs of FasterRCNN-C and

	GT	$TP_{(C,T)}$	$TP_{(C)}$	$TP_{(T)}$	$FP_{(C,T)}$	$FP_{(C)}$	$FP_{(T)}$
All	$3,390$	$1,149$	440	634	85	973	780
Day	$2,383$	917	403	307	73	562	604
Night	$1,007$	232	37	327	12	411	176

Table 3.1: Detections performed by FasterRCNN-C and FasterRCNN-T pedestrian detectors. Numbers of ground truths (GTs), true positives (TPs), and false positives (FPs) are reported on KAIST$\times$30 test set, in terms of all-day, daytime, and night time images.

FasterRCNN-T, in terms of all-day, daytime, and night time images. $TP_{(C,T)}$ denotes pedestrians detected by both of the two detectors. $TP_{(C)}$ and $TP_{(T)}$ represent pedestrians exclusively detected by FasterRCNN-C or FasterRCNN-T. Analogously, we have $FP_{(C,T)}$, $FP_{(C)}$, and $FP_{(T)}$ for false alarms.

In general, color and thermal images provide complementary information on human detection, even when strong ConvNet-based detectors are employed. The two detectors have $1,1149$ mutual true detections, while 440 and 634 human hypotheses are regarded as TPs by color and thermal input, respectively. Without a doubt, color and thermal images have consensuses to a substantial extent on validating true pedestrian instances. In contrast, the number of shared FPs is relatively fewer, while each imaging modality individually brings in many FPs. It seems color and thermal images have controversy in identifying false alarms.

We can also observe different performances of the detectors on daytime and night time images. During daytime, color image works better than the thermal. As in Table 3.1, color-based detector captures more TPs ($1,302$ *vs.*$1,224$), while mistaking fewer FPs (635 *vs.*677). The result is anticipated since during daytime most pedestrians are in good lighting, except some extreme cases (standing in shadow). Thermal imaging is apt to be corrupted by hot background during daytime, resulting in slightly worse performance. Differently, color channels struggle with night time images on pedestrian detection. It captures significantly fewer TPs, compared to the thermal-based detector (269 *vs.*559), yet bringing in more FPs (423 *vs.*188). Clearly, the thermal channel is more competent in detecting people under bad lighting or during night time.

Based on the detection results shown in Table 3.1, we can make an extreme assumption: if all true detections from FasterRCNN-C and FasterRCNN-T were kept and only shared false alarms were retained, then the detection rate can be increased from 46.9% to 65.6%, while the FPPI (false positives per image) will be reduced from 0.470 to 0.04. We would say there is

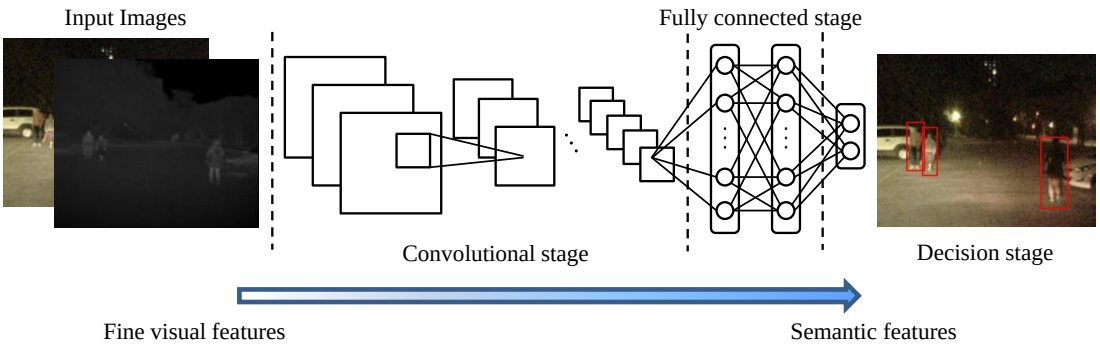


Figure 3.9: Different stages in a ConNet. Features at different stages correspond to various levels of semantic meaning and fine visual details.

a large improvement margin in recalling more true detections and excluding false alarms by leveraging both color and thermal images. Therefore, we should pay serious attention to the improvement potential that would be raised for human detection by using multispectral images.

3.4 Multispectral ConvNet

The question now is how a good multispectral human detector could be achieved that explores the most complementary visual information from color-thermal image pairs. It can be concluded from Table 3.1 that, in order to obtain more true detections, we need the 'union' of detections from color- and thermal-based detectors. In the meanwhile, their 'intersection' could significantly suppress false positives. Unfortunately, we cannot achieve these two goals at the same time. Thereby, a desired multispectral human detector should balance such trade-off when making decisions on human hypotheses.

Basically, a ConvNet-based detector is composed of several stages: the input stage, the convolutional stage (a series of convolutional layers), the fully-connected stage, and the decision stage, as shown in Figure 3.9. Recent research on deep convolutional networks revealed that features at different stages corresponding to various levels of semantic meanings. Features from higher level layers carry with more semantic meaning, while losing fine visual details. Visualizations of convolutional filters in [123] clearly show their semantic differences. We conjecture that fusion at different stages would lead to distinct detection results. Therefore, the multispectral human detection task turns out to be a ConvNet fusion problem, which studies various architectures of fusion model to approach best detection synergy.

C1	$3 \times 3 \times 3 \times 64$ + ReLU $64 \times 3 \times 3 \times 64$ + ReLU
P1	2×2 max pooling
C2	$64 \times 3 \times 3 \times 128$ + ReLU $128 \times 3 \times 3 \times 128$ + ReLU
P2	2×2 max pooling
C3	$128 \times 3 \times 3 \times 256$ + ReLU $256 \times 3 \times 3 \times 256$ + ReLU $256 \times 3 \times 3 \times 256$ + ReLU
P3	2×2 max pooling
C4	$256 \times 3 \times 3 \times 512$ + ReLU $512 \times 3 \times 3 \times 512$ + ReLU $512 \times 3 \times 3 \times 512$ + ReLU
C5	$512 \times 3 \times 3 \times 512$ + ReLU $512 \times 3 \times 3 \times 512$ + ReLU $512 \times 3 \times 3 \times 512$ + ReLU
F6	$(512 \times 7 \times 7) \times 4096$
F7	4096×4096

Table 3.2: Details of our vanilla ConvNet. Convolutional filters are denoted in form of Dimension×H×W×Number. For layer F6, $512 \times 7 \times 7$ represented the dimension of feature after RoI pooling layer.

In this work, we make thorough inquiries on four distinct fusion models designed upon our vanilla ConvNet, while each fusion model represents a multispectral human detector. We denote them as Early Fusion, Halfway Fusion, Late Fusion, and Score Fusion. Basically, these fusion models are two-branch ConvNets which merge at some point of the network, as illustrated in Figure 3.10, such that features of different levels of semantic meaning are fused. To facilitate following explanations on these fusion models, we symbolize our vanilla ConvNet as C1 - P1 - C2 - P2 - C3 - P3 - C4 - C5 - F6 - F7. To make this dissertation self-contained, we list details of these layers in Table 3.2. To be noticed, each convolutional component is followed by a non-linear activation function, *i.e.*, rectified linear unit (ReLU) [124].

Early Fusion concatenates feature maps of color and thermal branches immediately after the second convolutional layers (C2). We reuse the convolutional filters in C1 and C2 of the VGG16 model without fine-tuning, similarly to the training protocol of Faster R-CNN. The reason is we think fine-tuning of C1 and C2 layers with thousands of training images would hit over-fitting. Vanishing gradient problem during back-propagation would also affect the convergence of C1 and C2 layers. Besides, filters in C1 and C2 were pre-trained by millions of neural images,

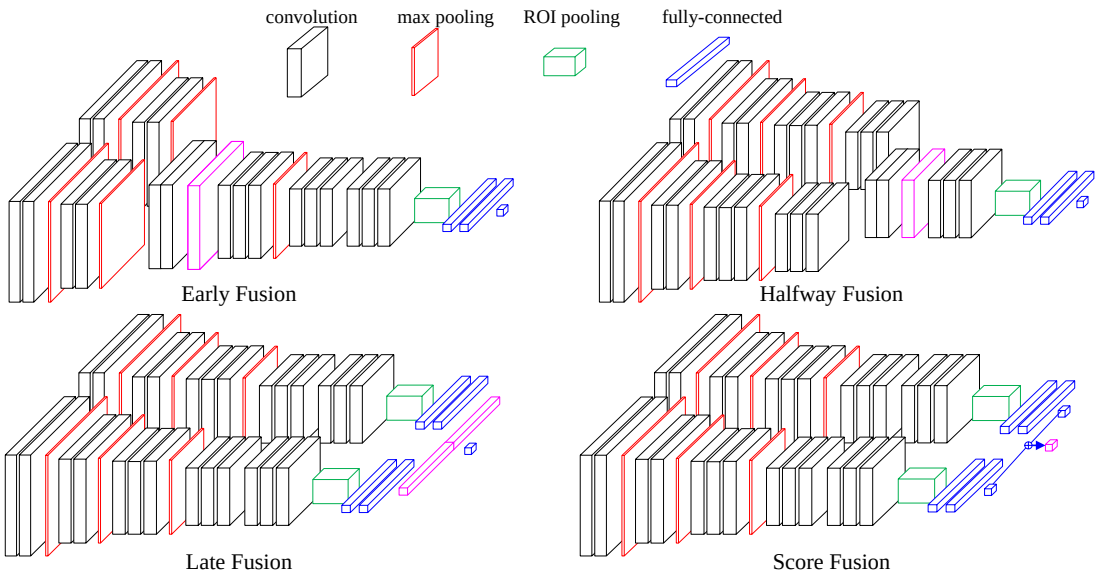


Figure 3.10: Explored ConvNet fusion models to leverage color and thermal images for multispectral pedestrian detection, *i.e.*, *Early Fusion*, *Halfway Fusion*, *Late Fusion*, and *Score Fusion*. These fusion models correspond to low-level feature fusion, middle-level feature fusion, high-level feature fusion, and confidence-level fusion, respectively. Magenta box in each model represent layers (or scores) immediate after fusion. For the sake of conciseness, ReLU layers and dropout layers are hidden from view in this figure. (Best viewed in color.)

which are powerful in extracting low-level features, even when thermal inputs are given.

After feature concatenation, we introduce the Network-in-Network (NIN) [125, 126], which is indeed a $256 \times 1 \times 1 \times 128$ convolutional layer followed by ReLU. Here, 256 is the dimension of the concatenated feature map; 1×1 is the filter size; 128 is the number of filters. Clearly, NIN reduces the dimension of the concatenated layer to 128. Consequently, filters of other layers in the pre-trained VGG16 model can be reused in initialization. Besides, NIN applies ReLU on the linear combinations of features from color and thermal branches, which could enhance the discriminability of local patches. Since C2 captures low-level visual features, such as corners, lines, line segments, etc., we expect *Early Fusion* model fuses features at low level.

Halfway Fusion also executes fusion at the convolutional stage. Different from *Early Fusion*, it places the fusion module after the fourth convolutional layers (C4). NIN ($512 \times 1 \times 1 \times 256$ convolutional layer + ReLU) is also used after feature concatenation, for the same reasons as discussed above. Compared to features from the C2 layer, C4 extracts features of more semantic meaning, while retaining some visual details.

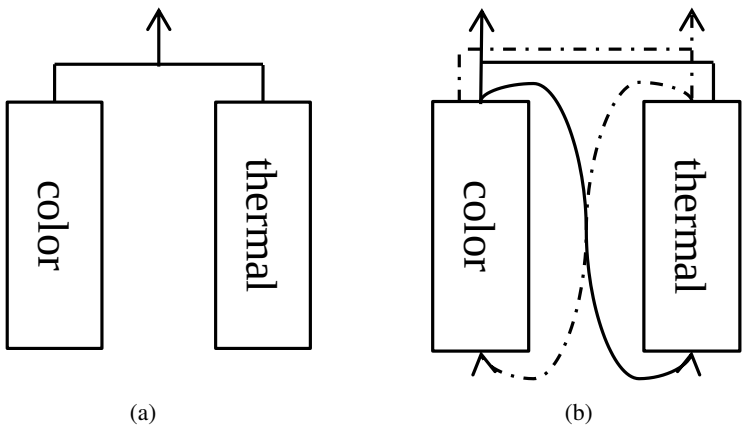


Figure 3.11: Pipelines of different schemes applicable to *Score Fusion* model. (a) Parallel scheme. (b) Cascade scheme.

Late Fusion concatenates features of last fully-connected layers (F7). In image classification [28] or face recognition [127], F7 feature conventionally stands for the new visual representation. Therefore, we say *Late Fusion* model executes high-level feature fusion. Concatenated features ($8,192$-d) are then used in the softmax classifier to identify human objects. To be noticed, *Early Fusion* and *Halfway Fusion* execute branch merging before RPN, such that they can directly use the C5 features to generate proposals. In contrast, fusion in *Late Fusion* model happens after RPN. In order to make a fair comparison, RPN in *Late Fusion* model also concatenates C5 features of the color and thermal branches to predict human proposals.

Score Fusion merge detection confidences from color and thermal branches with equal weights (*i.e.*, 0.5). There exist two schemes applicable for Score Fusion, as shown in Figure 3.11. (I) Parallel scheme. We first generate proposals by using concatenated C5 features, similarly to that in *Late Fusion* model. Then, each proposal collects detection confidences from the parallel color and thermal branches, while regression results are also merged. (II) Cascade scheme. We get detections from color images at the first step and then use them as proposals for the other ConvNet detector, such that we could obtain their corresponding scores by thermal inputs. The similar process is executed again to start with detections from thermal images. At the end, non-maximum suppression (NMS) is applied to all detections from color and thermal channels, in order to avoid double detections. Since we employ RoI Pooling layer, when C5 feature maps from the two ConvNets are computed, detection at the second step can be accomplished

without extra computational overload. In practice, we find cascade scheme performs better than the parallel one. Hence, unless otherwise specified, we refer *Score Fusion* to the one with cascade scheme.

3.5 Experiments

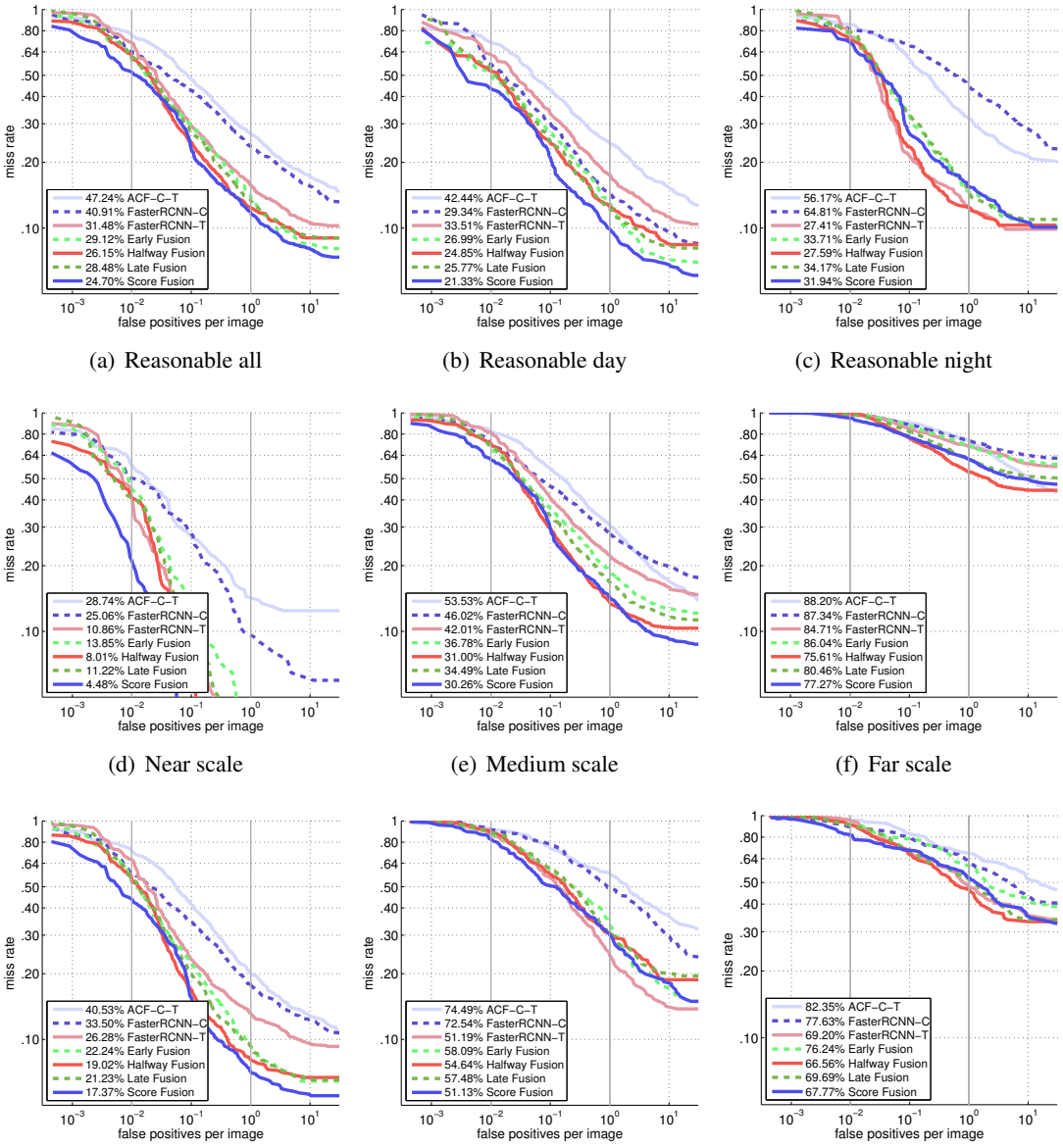
Dataset: KAIST multispectral pedestrian dataset (KAIST) [43] contains $95,328$ aligned color-thermal frame pairs, with $103,128$ dense annotations on $1,182$ unique pedestrians. We sample images from training videos with 2-frame skips, and finally, obtain $7,095$ training images with qualified pedestrians (the same criteria as discussed in Section 3.2). KAIST$\times$30 test set contains $2,252$ images, among which $1,455$ images were captured during daytime and the other 797 from night time. Some statistics on KAIST$\times$30 test set can be found in Section 3.3.1.

Implementation Notes: Parameters in all the four ConvNet fusion models are initialized by the pre-trained VGG16 model, except newly introduced layers. For instance, in *Early Fusion* and *Halfway Fusion* models, weights of NINs are initialized by Gaussian distributions. Parallel branches in the fusion models do not share weights. The two branches in *Early Fusion*, *Halfway Fusion*, and *Late Fusion* are trained simultaneously. In *Score Fusion*, the two branches are trained by color and thermal images, respectively. In fact, *Score Fusion* is a cascade model consisted of FasterRCNN-C and FasterRCNN-T detectors. All the models are fine-tuned with SGD for 4 epochs with LR 0.001 and 2 more epochs with LR 0.0001.

3.5.1 Evaluation of Detection

We evaluate the proposed four ConvNet fusion models on KAIST$\times$30 test set, compared to FasterRCNN-C and FasterRCNN-T reported in Section 3.3.2, as well as ACF-C-T detector [8]. The ACF-C-T detector used 10-channel aggregated features constructed by both color and thermal images [43]. Comparisons of these detectors are presented in Figure 3.12, in terms of various subsets suggested by [43].

Day and night. The *reasonable* subset includes non-/partially occluded pedestrians of heights more than 55 pixels, which is divided into *reasonable day* and *reasonable night*. In general, detectors with single image modality obtain worse results than fusion models. Overall, we



(a) Reasonable all (b) Reasonable day (c) Reasonable night

(d) Near scale (e) Medium scale (f) Far scale

(g) No occlusion (h) Partial occlusion (i) Heavy occlusion

Figure 3.12: Miss rates versus false positive per-image curves shown for various subsets of the data. Lower curves indicate better performance; the log-average miss rate (MR) for each detector is shown in plot legends. (a-c) Performance w.r.t. time (computed for no or partial occluded pedestrians of 55 pixels or more). (d-f): Performance w.r.t. scale (computed for non-occluded pedestrians). (g-i): Performance under varying levels of occlusion (computed for pedestrians of 55 pixels or more).

achieve 24.7% MR on KAIST$\times$30 test set with *Score Fusion*. Such conclusion is different from our published work in [91] due to the improved annotations.

It is easy to conclude from Figure 3.12(b) that FasterRCNN-C outperforms FasterRCNN-T on the daytime images. This means color imaging is more competent than the thermal in differentiating human objects against backgrounds when environmental illumination is good. Compared to FasterRCNN-C and FasterRCNN-T, all the fusion models produce significantly better detection results. Among the four ConvNet fusion models, *Score Fusion* achieves the lowest overall MR (21.3%) that is 8% lower than FasterRCNN-C, showing the most effective synergy for multispectral pedestrian detection. We speculate that, during the daytime, both of color and thermal modalities work well. Decision-level fusion could perform the best in boosting true detections (both votes *pedestrian*) and suppressing false positives (both votes background). Unfortunately, color channels suffer on night time images due to bad lighting. As shown in Figure 3.12(c), FasterRCNN-C even works worse than the ACF-C-T detector (64.8% *vs.* 56.2%). It is clear that FasterRCNN-T outperforms other detectors on night time images and the thermal channel seems to be dominant in the fusion models. We speculate that for night time images, semantic noises and decision mistakes from the color channels are very difficult to thoroughly eliminate through fusion.

Scales. In this experiment, we examine the detectors on the pedestrians of different image sizes, *i.e.*, *near* (115 pixels$\sim$), *medium* ($45 \sim 115$ pixels), and *far* (~ 45 pixels). These subsets only contain non-occluded pedestrians. As shown in Figure 8(d)$\sim$(f), the thermal detector outperforms the color detector, especially for *near* pedestrians (10.9% *vs.* 25.1%). Fusion models can further reduce the miss rate, while *Score Fusion* works the best for *near* and *medium* pedestrians. For the *far* subset, *Halfway Fusion* obtains lowest miss rate. We believe that color and thermal detectors make consistent decisions to some degree on *near* and *medium* pedestrian instances. As a result, *Score Fusion* works the best. When color and thermal channels are controversial in identifying *far* human objects, *Halfway Fusion* could implicitly temper mistakes of the two modalities, thereby achieving the lowest MR (75.8%).

Occlusions. These subsets contain pedestrians of varying levels of occlusions, *i.e.*no occlusion, partial occlusion ($\sim 50\%$ occluded), and heavy occlusion ($50\% \sim$ occluded). *Score Fusion* beats other detectors on non-occlusion and partial-occlusion pedestrians, while *Halfway Fusion*

Figure 3.13: Detection samples. Red bounding boxes denote detections. Yellow arrows indicate false positives and green ellipses represent miss detections. First row: detections by FasterRCNN-C detector. Bottom two rows: first two (daytime images) by *Score Fusion* detector, the others (night time images) *Halfway* Fusion detector, illustrated in both color and thermal images.

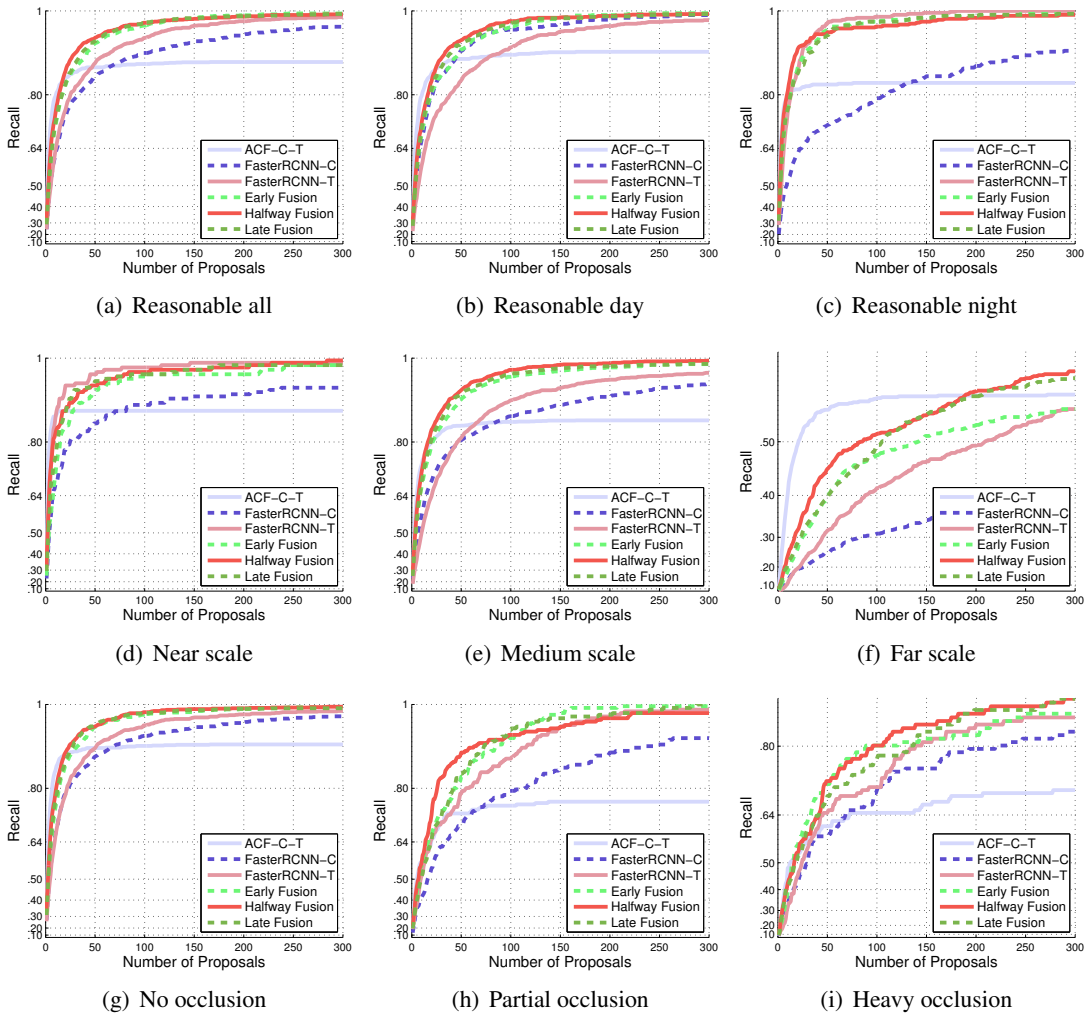
model works the best for heavy-occluded pedestrians. Our deduction is similar to the one above involved in scale that the compatibility of color and thermal imaging decides the choice of the proper fusion model.

Some detection samples are presented in Figure 3.13. Detections with confidences large than 0.5 are illustrated. Obviously, compared to the color image based detector, our multispectral human detectors achieve more true detections, especially when some pedestrians are in bad external illumination. Meanwhile, false alarms are also removed.

3.5.2 Evaluation of Proposals

We also assess the proposals generated by RPNs in different detectors, with respect to recalls. We considered the RPNs of FasterRCNN-C and FasterRCNN-T in comparison, as well as the ACF-C-T pedestrian detector. The comparisons performed on the KAIST$\times$30 test is shown in Figure 3.14. To have a deeper understanding of the proposals, recalls are evaluated under different criteria, i.e., under various image scales and occlusion levels.

Recall *vs.* number of proposals: On the *reasonable* subset, given IoU 0.5 and fixed number of



(a) Reasonable all (b) Reasonable day (c) Reasonable night

(d) Near scale (e) Medium scale (f) Far scale

(g) No occlusion (h) Partial occlusion (i) Heavy occlusion

Figure 3.14: Miss rates versus false positive per-image curves shown for various subsets of the data. Lower curves indicate better performance; the log-average miss rate for each detector is shown in plot legends. (a-c) Performance w.r.t. time (computed for no or partial occluded pedestrians of 55 pixels or more). (d-f): Performance w.r.t. scale (computed for non-occluded pedestrians). (g-i): Performance under varying levels of occlusion (computed for pedestrians of 55 pixels or more).

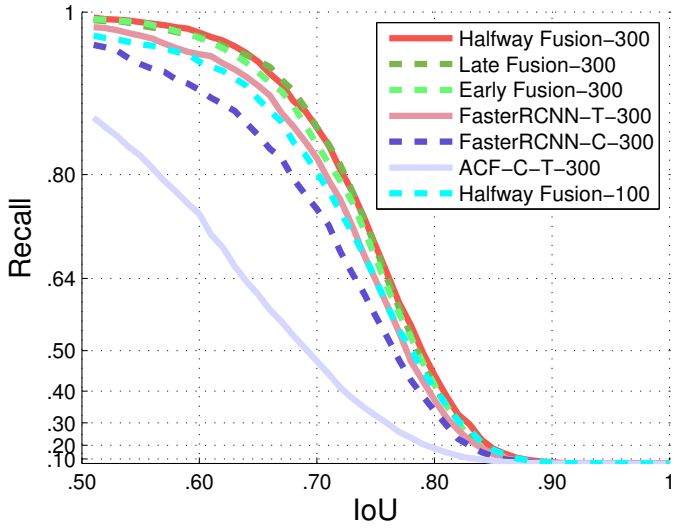
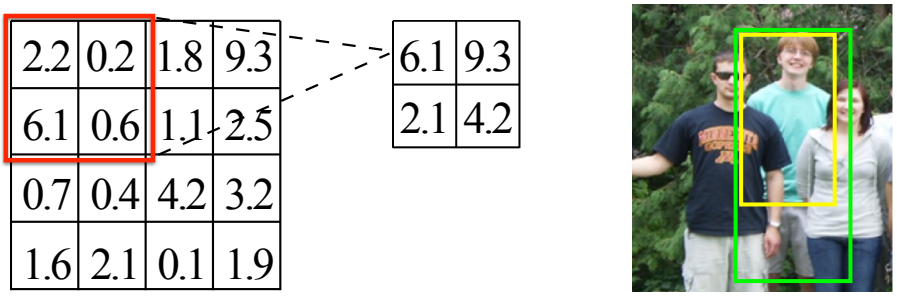


Figure 3.15: Comparison of 300 pedestrian proposals reported on *reasonable* subset of KAIST×30 test set dataset: Recall *vs.*IoU.

proposals, RPNs of fusion models obtain higher recall than the RPNs using one image modality. For instance, *Halfway Fusion* model achieves 94.1% recall with 50 proposals. By contrast, FasterRCNN-T obtains 88.1% of and 84.7% for FasterRCNN-C. In other words, fusion models could reach the same recall with fewer proposals. This is very useful in practice since fewer proposals could save time in classification. Overall, *Halfway Fusion* gets 90% recall with 30 proposals, while FasterRCNN-C and FasterRCNN-T require around 80 proposals to achieve a competitive recall. This result indicates that fusion models retrieve more true detection and we believe such improvement benefits the classification step. It is also notable that the performances of the RPNs in the three fusion models do not have much difference.

Without surprise, color imaging works better than the thermal in generating human proposals from daytime images (as shown in Figure 3.14(b)), while thermal imaging turns out to outperform the color on night time images. From Figure 3.14(d), we can observe that thermal channel works best on *near* pedestrians. This is because local body configuration extracted by thermal imaging is more useful in inferring global locations of human objects, rather than visual details. In Figure 3.14(e) and 3.14(f), fusion performs better than single image modality, while both color and thermal channels play important roles. To be mentioned, the minimum scale of anchor we use is 8×8, corresponding to image area of 64×64 pixel. Apparently, such scale cannot capture small human instances well. Consequently, the recall on *far* instances is



(a) Max pooling layer (b) Rgression model

Figure 3.16: Merits of deep convolutional networks for human detection. (a) Max pooling layer. The maximum value in each cell (*e.g.*, 2×2) is picked up and used in forward- and back-propagation. (b) Regression model. Yellow bounding box represents receptive filed of deep feature, which can be deformed to the green one by a regression model. Obviously, the green bounding box has better localization on the person in the middle of the image.

relative lower. A possible solution is using more scales for anchors (*e.g.*, 5×5), which would be interesting to investigate. According to Figure 3.14(g)- 3.14(i), fusion models could obtain nearly 100% recall, even with partial occlusion. While the occlusion ratio is larger than 0.5, the recall still keeps at an acceptable level.

Recall *vs.* IoU: The comparison of experimental results is illustrated in Figure 3.15. Given 300 proposals from RPN, fusion models obtain around 97% recall at IoU 0.6, which is better than FasterRCNN-C (90.0%) and FasterRCNN-T (95.0%). With 100 proposals, *Halfway Fusion* model accomplishes comparative recalls against FasterRCNN-T with 300 proposals. Clearly, fusion models produce proposals with better overlaps on true detections. These proposal of better localizations would lead to more human-relevant convolutional features since these feature are pooled according to the locations of proposals.

3.6 Discussions

3.6.1 Merits of Deep ConvNets

Convolutional filters capture local visual structures and feature correlations. However, not all of the visual information is useful in human detection. By using ReLU (activation function) and max pooling, task-independent features can be discarded or suppressed. Equation 3.1 shows the operation of ReLU. Apparently, neurons with negative values in the network will not be

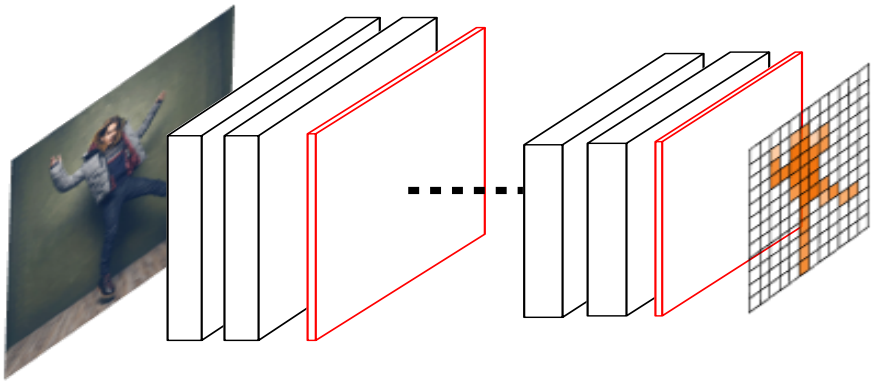


Figure 3.17: Deep feature extracted through a series of convolutional layers and max pooling layers. Only human object related features are activated.

fired and their connections with the classification stage are cut off. These neurons are regarded of carrying irrelevant information, and would not have the influence on the final decision. Besides, compared to the logistic sigmoid function used previously, ReLU is more practical in back-propagation, since fired neurons get constant gradients in activation layers,. Such advantage would accelerate the training convergence of a deep neural network. Some other activation functions, such as LeakyReLU [128] and ELU [129], do not have non-zero response on negative value domain. They were proposed for training very deep networks (*e.g.*, ResNet [130]). Although negative signal cannot be cut off completely, compared to positive neural response, they are still suppressed to a substantial extent.

$$y = \max(0, x) \tag{3.1}$$

Figure 3.16(a) shows the operation of a max pooling layer. The maximum value in each cell (*e.g.*, 2×2) is picked up and used in forward- and back-propagation. These values represent the most informative features in identifying human objects. By leveraging a large scale of training data, the most common and robust visual features of human can be extracted from deep ConvNet, as illustrated in Figure 3.17. Such features could capture the visual complexity of human samples in training images, which are robust to various appearances and poses to a certain extent.

Another advantage of deep ConvNets we use in this work is the application of regression models. As shown in Figure 3.16(b), even the original location of a human hypothesis is not accurate, possibly due to partial occlusion, a well trained regression model can deform the

Figure 3.18: Some examples from KAIST×30 test set, showing mismatches in the aligned color-thermal image pairs. The mismatches could be observed according to the image margins of green and red lines.

bounding box to an image area that better matches the ground truth. Since at the training stage of Faster R-CNN image areas of IoU larger than 0.5 with a ground truth annotation are regarded as positive samples, the deformation (from (x, y, w, h) to $(x + dx, y + dy, w + dw, h + dh)$) can be learned. We speculate that deep features trained with regression loss would carry some layout information (*e.g.*, orientations of body and limbs), which is useful in inferring the whole body area by given locations of parts.

3.6.2 Some Bottlenecks

We have validated our proposed model on multispectral human detection through extensive experiments, demonstrating promising results on KAIST pedestrian benchmark. However, there exist some issues in this work, which prevent current framework from obtaining better results. We believe the future research solving these issues could further improve the performance of multispectral pedestrian detection.

Alignment of multispectral image. As mentioned in [112], one important issue in multispectral detection is the registration of images from different sensors. Color-thermal image pairs in KAIST have been aligned using hardware-based calibration approach [43]. However, we still observe mismatches in the registered pairs, as shown in Figure 3.18. Such mismatching would definitely affect the training of fusion models, although it is hard to make a quantitative analysis. For instance, a local patch belonging to a human body in the color image is actually the background in the thermal image. On the other hand, error in alignment could result in wrong decisions during testing, especially for pedestrians of small image sizes. Since each detection

is validated by IoU (usually 0.5), misaligned annotations would mistake a indeed true detection as a false alarm. Consequently, alignment of multispectral imaging is a crucial preprocessing for pedestrian detection, which requires more attention.

Adaptive weights for color and thermal channels. Recall that *Score Fusion* model merges two detection confidences with equal weights (0.5 *vs.* 0.5). Actually, we can regard FasterRCNN-T as a spatial case of *Score Fusion* model which gives 0 weight to the detections based on color inputs. Figure 3.14(b) and 3.14(c) demonstrate that *Score Fusion* works best for daytime images, while FasterRCNN-T for night time images. Imagine that we have a component in *Score Fusion* model that could adaptively adjust weights for detection confidences of color and thermal branches, the overall performance of *Score Fusion* can be further improved. Inspired by this idea, we have tried an auxiliary ConvNet that takes convolutional features from color and thermal images as input. We expect this network to output adaptive weights for detections coming from color and thermal images. However, the learned auxiliary ConvNet even deteriorated the detections of *Score Fusion*. More efforts could be put in this direction. Besides, we are thinking, a lighting sensor may also help in this scenario rather than an autonomous algorithm.

Scales. Our proposed models achieve worse performance on *pedestrians* of small image sizes, compared to *near* and *medium* instances, as shown in Figure 3.12(f) and Figure 3.14(f). Recently, there are several methods working on this problem which used scale dependent pooling strategies [131, 64] or Generative Adversarial Network [132]. It is worthy of indicating that these approaches can be seamlessly incorporated into our current framework, which can make our human detectors perform better on pedestrians of small sizes.

3.7 Summary

In this chapter, we have focused on leveraging deep convolutional neural networks for multispectral (color and thermal images) human detection. Our multispectral detectors were designed based on Faster R-CNN framework, which achieved the state-of-art performance on Caltech pedestrian benchmark. We have devised four ConvNet fusion architectures that fused two-branch ConvNets at different stages, corresponding to low-level, middle-level, high-level

feature fusion, and confidence fusion. All of them have yielded better detection results, compared to the baseline detector, *i.e.*, Faster R-CNN detector. Through extensive experiments, we have validated three hypotheses: 1) End-to-end learning of ConvNet-based human detector is superior to the R-CNN framework that replies other independent human detectors. 2) Based on deep ConvNets, multispectral imaging still demonstrates better capability of detecting human objects in various conditions, rather than using color or thermal images only. 3) Although achieving good performance, different fusion models are preferred in different scenarios, which should be further investigated with more training and testing data. Given the improved annotation, our trained detector obtained overall 24.7% MR on KAIST, showing great improvement in this area.

Chapter 4

Graph-based Context Modeling

4.1 Introduction

Vision-based human detection has witnessed great advances recent years, a bunch of approaches such as deformable part model (DPM) [7], poselets [55] and deep convolutional neural network (ConvNet) [20]. Although these approaches and their variants have achieved tremendous promising results, the problem still remains quite challenging when it comes to the scenario where cluttered background, various occlusions and large pose variations exist. Despite these difficulties, some research effort has exploited contextual cues in a scene to improve human detection in adverse conditions. For example, two-person or multiple-person classifiers were built directly in several approaches [54, 65] to handle partial occlusion. Other approaches [45, 66, 46] explored pairwise spatial relationships between local neighbors to boost detection performance, under the framework of structured prediction.

Inspired by the above approaches, we develop a new framework to address the challenges of human detection by putting people in a global context and modeling their interactions. In a crowded scene, people usually form group, where they interact with each other both geometrically and socially [44, 37, 46, 133]. A group of people, either queuing, sitting or walking together, indicate spatial closeness and similar scales (Figure 4.1). There also exist strong social patterns in the group such as facing (i.e., two people face each other) and following (i.e., people stand or sit side by side).

To effectively leverage the geometric and social contexts in crowds, we propose a unified framework for human detection that integrates visual recognition with graph-based context modeling. We formulate the detection task as an optimization problem where the goal is to find a maximum set of human hypotheses that agrees on both visual detections and their contextual interactions in an image. While such optimization problem is theoretically intractable, we show

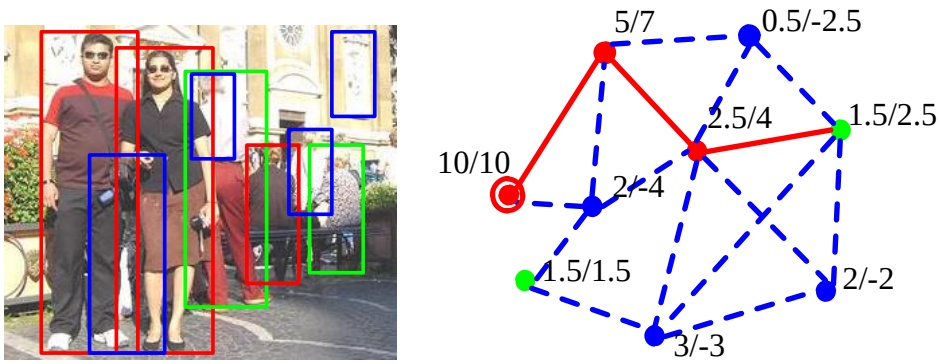


Figure 4.1: Context-drive label propagation for people detection. (a) Grouped people tend to present spatial closeness and similar scales (image from VOC 2012 for illustration). There also exist social interactions in a group such as *facing* and *following* (e.g., the two people on the left stand side by side). (b) A context graph captures the interaction strength between human hypotheses (or detections). Each node here is a human hypothesis from an underlying detector. True detections are colored as either red (high confidence) or green (low confidence) while false alarms as blue. A bold edge indicates mutual *attraction* between two nodes, i.e. they are contextually compatible. Oppositely a dotted edge suggests an opposite relationship, i.e. *repulsion*. Our approach applies label propagation to boost up weak detections (greed nodes) while suppressing irrelevant false alarms.

that it can be approximately addressed by *label propagation* [134] in a progressive way. Label propagation was originally proposed for predicting unlabeled instances based on annotations of labeled data [134]. It propagates labels to data iteratively by instance similarity. In our case, true detections are supposed to be contextually compatible with each other, but inconsistent with false alarms. This suggests that strong detections with high confidence can boost up weak ones by spreading rewards through contextual proximity and meanwhile penalize false positives according to contextual incompatibility, in a similar spirit to label propagation.

More specifically, our approach starts by taking input from an underlying detector, possibly with a large number of false alarms. We build a context graph that exploit contextual information available in a scene (spatial, scale, social, and overlap cues) for label propagation. For the purpose of rewarding true detections as well as suppressing irrelevant false alarms, our approach enables the graph to spread both positive and negative contextual potentials along edges in the context graph, which depends on contextual attraction and repulsion. As a result, contextually compatible human hypotheses get reinforced by receiving positive potentials during the propagation while false alarms are contained due to being negated by their incompatibility

with true detections. The idea of our proposed context graph is illustrated in Figure 4.1. If we start with a strong detection of scored by 10 (in a red circle), the context graph then enhances the detection scores of the true human instances, *e.g.*, from 5 to 7, and decrease the scores of fake human illusions to negative values.

Compared to the structured output learning framework that models contextual interactions only between local neighbors [45, 37, 46], our approach is clearly advantageous in that the graph-based propagation make interactions between any two nodes possible. Such a capability allows our approach to discover challenging weak human instances, even if they are not close to any strong detection, as illustrated later in Figure 4.10. We would also like to point out that Conditional Random Fields (CRFs) [135] methods while being another option for contextual modeling, are less suitable for our problem. This is largely because capturing long-range interactions requires fully connected CRFs, which need expensive inference mechanisms or strong assumptions such as Gaussian edge potentials [136] that does not fit our problem. Besides, modeling of fully connected CRFs requires large-scale training data and fine annotations covering contextual interactions.

Since we do not have truly 'labeled' data to start with, we further design a greedy-like inference similar to the approximation for structured output learning [45, 46], which iteratively adds the best hypothesis with the most potential gain at each time to initialize a new round of propagation. The process repeats until convergence when no new hypothesis can be instanced. Hence, this work presents a unified framework of graph-based label propagation to exploit contextual information for human detection. The main contribution of our approach is threefold:

- We devise the context graph which simultaneously considers four contextual information, *e.g.*, spatial proximity, scale consistency, social interaction, and layout. The context graph could spread award (positive potential) and punishment (negative potential) between vertices in the graph, such that weak detection can be reinforced, with false alarmed suppressed.

- We propose the progressive inference algorithm to approximately solve the optimization problem modeled for human detection, including the potential propagation algorithm which computes the contextual potential of each human hypothesis obtained through a

pre-defined context graph.

- We validate our approach on two challenging crowd datasets, one for detecting people with variations in pose and size, and the other for pedestrian detection in low-resolution images. Our experimental results confirm that the proposed method can significantly improve human detection in crowded scenarios, achieving performance comparable to the state of the art approaches reported in the literature.

This chapter is organized as follows. In Section 4.2, we describe an optimization problem formulated to leverage both unary detection confidences and contextual information for human detection in crowded scenes. Our approach models people interactions by constructing a *context graph*, which is introduced in Section 4.3. We propose a greedy-like method to approximately solve the optimization problem, in Section 4.4. Experimental results and ablation analysis will be presented in Section 4.5.

4.2 Problem Formulation

Given an image, let $\mathbb{X} = \{x_i, i = 1 : m\}$ be a set of m human hypotheses generated by an underlying people detector. We purposely set a low detection threshold to allow for more true detections in $\mathbb{X}$, which unfortunately also gets many more undesirable false alarms. Therefore, our task is to find a subset of $\mathbb{X}$ that covers as many as possible true detections, meanwhile with the fewest false alarms brought in. Mathematically, we aim at seeking an indicator vector $\mathbf{Y} = \{y_1, y_2, ..., y_m\}^T \in \{0, 1\}$ (1 means true people detection, otherwise background or other objects) that maximizes a potential function $\Psi(\mathbb{X}, \mathbf{Y})$, such that the visual detections agree on the contextual setting of the image. We define the potential function as follows:

$$\Psi(\mathbb{X}, \mathbf{Y}) = \sum_{i=1}^{m} y_i \psi^u(x_i) + \alpha \sum_{i=1}^{m} y_i \psi^c(x_i, \mathbb{X}, \mathbf{Y}) \tag{4.1}$$

where $\psi^u(\cdot)$ is the unary potential that utilizes the original detection score of a hypothesis in our case. $\psi^c(\cdot)$ represents the total contextual potentials (support) of a human hypothesis received from others. α is a constant number balancing these two terms. In previous approaches such as proposed in [45, 46, 37], $\psi^c(x_i, \mathbb{X}, \mathbf{Y})$ represents the support of a hypothesis received from its neighbors. For example, Desai *et al.* [45] modeled 6 contextual patterns based on relative

spatial locations of two hypotheses. While this proves effective in some cases, it lacks a way to model interactions beyond the 6 spatial patterns. Instead of pre-specifying local spatial relationships, in our approach, we implicitly model the contextual interaction between any two human hypotheses via a *context graph* $\mathcal{G}$. Thus we have,

$$\psi^c(x_i, \mathbb{X}, \mathbf{Y}) \triangleq \psi^{\mathcal{G}}(x_i, \mathbf{Y}) \tag{4.2}$$

where $\psi^{\mathcal{G}}(x_i, \mathbf{Y})$ measures how much contextual potential hypothesis x_i can obtain from validated human hypotheses (reflected by $\mathbf{Y}$) based on $\mathcal{G}$. We drop $\mathbb{X}$ in $\psi^{\mathcal{G}}$ here for clarity. While contextual confidence in previous methods [45, 46, 37] can be regarded as a linear combination of interactions between a node and its local neighbors, our graph-based context modeling enables a node to interact with any node in the graph in a nonlinear way through label propagation, effectively and efficiently.

4.3 Context Graph

A *context graph* in our approach is an undirected graph $\mathcal{G} = (V, E)$ used for label propagation, where V corresponds to a set of human hypotheses and E indicates the strength of contextual interaction between any pair of hypotheses. While our focus is to reward human hypotheses contextually consistent to true detections, suppressing false alarms is equally important during label propagation as our input contains a substantial number of errors. For such a purpose, we consider two types of strengths when constructing $\mathcal{G}$: *attraction* e^+ and *repulsion* e^-. Here *attraction* measures contextual compatibility between two hypotheses while *repulsion* relates to contextual inconsistency.

In our approach, we deliberate 4 types of contextual cues, namely scale, spatial, layout and social context, and denote their attraction strengths with e^+_{sc}, e^+_{sp}, e^+_{la}, and e^+_{so}, respectively. Similarly, their repulsion strengths are denoted by e^-_{sc}, e^-_{sp}, e^-_{la}, and e^-_{so}. We further define the overall attraction strength $e^+(i, j)$ between hypothesis x_i and x_j by

$$e^+_{i,j} = \min\left(e^+_{\text{sc}}, e^+_{\text{sp}}, e^+_{\text{la}}, e^+_{\text{so}}\right) \tag{4.3}$$

and their overall repulsion strength by

$$e^-_{i,j} = \max\left(e^-_{\text{sc}}, e^-_{\text{sp}}, e^-_{\text{la}}, e^-_{\text{so}}\right) \tag{4.4}$$

Finally, the context graph is represented by a symmetric matrix $\mathcal{G} \in \mathbb{R}^{m \times m}$, where $\mathcal{G}_{i,j} = e^{+}_{i,j} - e^{-}_{i,j}$. To be noticed, $G_{i,j}$ may have negative values, *i.e.*, edges in the context graph would have negative weights, through which repulsion can be spread from true detections to false alarms. The 'min' operation in Equation (4.3) implies the attraction of two hypotheses is low whenever one of the contextual attractions is weak since we expect giving reliable awards to human hypotheses of contextual compatibility which are more likely to be true detections. Differently, the 'max' operation in Equation (4.4) indicates any incompatible pattern should lead to high repulsion. Recall that the underlying detector would bring more false alarms than true positives. In order to remove these false positives from final detections, the 'max' operation would spread the strictest punishment to hypotheses of contextual inconsistency.

4.3.1 Feature Representation of $\mathcal{G}$

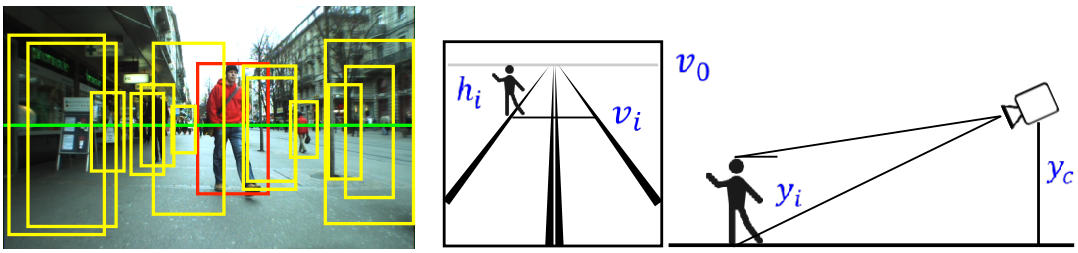
We describe below the feature representation for each type of context considered in our paper. These features will be further mapped to a value between 0 and 1 to indicate the strength of the contextual interaction.

Spatial context. Two observations motivate us to exploit spatial context. 1) People coexist nearby in crowd scenes. 2) People usually occupy comfortable zones. The first observation helps in boosting human hypotheses close to a true detection, while the second one could infer unreasonable detections (*e.g.*, double detections). Here, the image distance $d(x_i, x_j)$ between two hypotheses x_i and x_j is used to represent spatial context. To eliminate the effects of image resolution and camera perspective, $d(x_i, x_j)$ is further normalized by two additional items:

$$f_{\text{sp}}(x_i, x_j) = \frac{2d(x_i, x_j)}{h_i + h_j} \cdot \max\left(\frac{h_i}{h_j}, \frac{h_j}{h_i}\right) \tag{4.5}$$

where h_i and h_j are the image heights of hypothesis x_i and x_j. h_i/h_j compensates camera perspective as this ratio can sort of reflect the depth change of the two hypotheses. Here, we use the '$\max$' operation to obtain symmetric feature. The sum of h_i and h_j further normalizes the distance into the unit of human height, which counterbalances different image resolutions.

Scale context. We assume people roughly have the same height and most of them rest on the ground plane, when using scale context. The idea of scale context is demonstrated in Figure 4.2(a). Given the image location of the horizontal line (green line) and the scale of a



(a) Motivation (b) Variable

Figure 4.2: Scale context. (a) Motivation. The scale of the blue hypothesis is clearly controversy to that of others (in yellow). (b) Illustration of variables used in computing scale context. v_0: image location of the horizon line; h_i: image height of a person; y_i: physical height of a person; y_c: physical height of the camera; v_i: image location of a person.

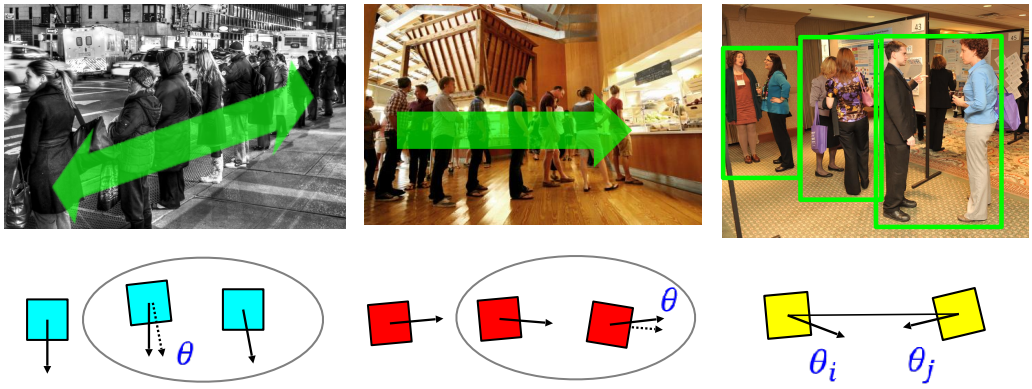
true detection (red bounding box), we can impose a scale prior for each hypothesis in the image. Apparently, in this case, the blue hypothesis does not coincide with such prior. Therefore, it should be regarded as a false alarm.

The scale context is approximated by the height (under world coordinate system) ratio of two hypotheses. The first step is to estimate the image location of the horizon line v_0. We apply the method in [137], while the involved variables are illustrated in Figure 4.2(b). By assuming all hypotheses are all grounded and upright, the physical height of hypothesis x_i can be represented as $y_i = h_i y_c/(v_i - v_0)$, where h_i and v_i encode the physical height and image location of x_i, respectively. y_c is the camera height. Consequently, given any two detections, we can obtain a value pair $(v_0(h_i - h_j), h_i v_j - h_j v_i)$. With multiple ($\geq 3$) strong detections of high confidences, v_0 can be easily estimated by least square fitting. Then the physical height ratio of two hypotheses x_i and x_j can be defined as:

$$f_{\text{sc}}(x_i, x_j) = \min\left(\frac{h_i(v_j - v_0)}{h_j(v_i - v_0)}, \frac{h_j(v_i - v_0)}{h_i(v_j - v_0)}\right) \tag{4.6}$$

where the 'min' operation guarantees $f_{\text{sc}} \in (0, 1]$. In other words, if f_{sc} is close to 1, the two hypotheses have similar scale (physical height).

Social context. We can observe different kinds of social interaction between people and here we model three of them, *i.e.*, following in line, following in queue, and facing each other, as shown in Figure 4.3. These social interactions among human hypotheses are measured in virtue of pose and body (or head) orientations. Obviously, in these interactions, the body orientation



(a) Following in line (b) Following in queue (c) Facing each other

Figure 4.3: Social context. Each circle represents one human hypothesis; arrows illustrate body (head) orientations. We model three kinds of social interaction patterns between any two hypotheses, i.e., (a) *following in line*, (b) *following in queue* and *facing each other*, which can be indicated by their body (head) orientations.

of a person indicates the position of another. As shown in the bottom row of Figure 4.3, considering a pair of hypotheses, $\angle\theta$ represents the included angle of body orientations; $\angle\theta_i$ and $\angle\theta_j$ are the included angles of head orientations and the connecting line of two hypotheses. A small $\angle\theta$ indicates the following pattern, while small $\angle\theta_i$ and $\angle\theta_j$ exhibit a facing pattern. Estimation of body (or head) orientations takes advantage of poselet activation vector (PAV) [138], which is capable of handling both profile and back views. Besides, we also train a pose classifier (standing vs. sitting) as a RBF-kernel SVM with 1200-dim PAV. The 2-dim probability output p of this classifier is then used to evaluate pose similarity. Finally, the strongest orientation pattern is used to represent social context, thus we have:

$$f_{\text{so}}(x_i, x_j) = \min(\angle\theta, \max(\angle\theta_i, \angle\theta_j)) \times ||p_i - p_j||_2 \tag{4.7}$$

Layout context. Typically, detections partially overlapping (IoU < 0.5) true detections are treated as false positives. In order to remove them, conventional approaches used non-maximum suppression (NMS). NMS is a greedy search algorithm, which is controlled by a threshold to tolerate crowd of high density in images. However, NMS cannot deal with cases as illustrated in Figure 4.4, where the overlapped areas between these false positives and true detections only occupy a small ratio (typically $0 <$ IoU ≤ 0.3). To further exclude these false positives, we use the overlap ratio of two bounding boxes to express their layout compatibility following the

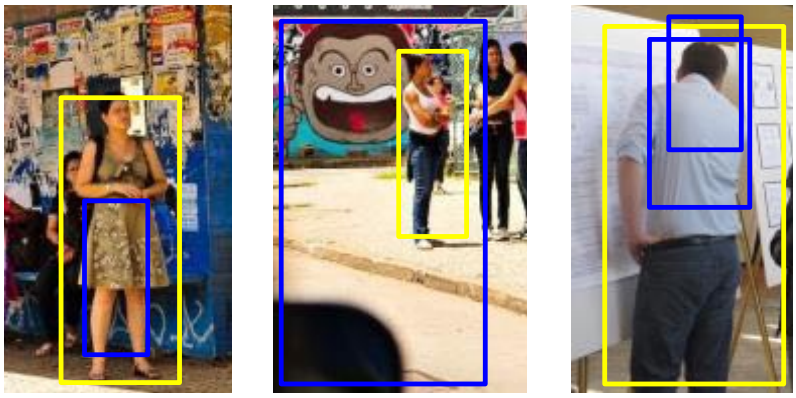


Figure 4.4: Result samples of orientation estimation on ground truth annotations. Magenta lines represent head orientations, while the cyan is for body orientation.

conventional approaches [37, 139]:

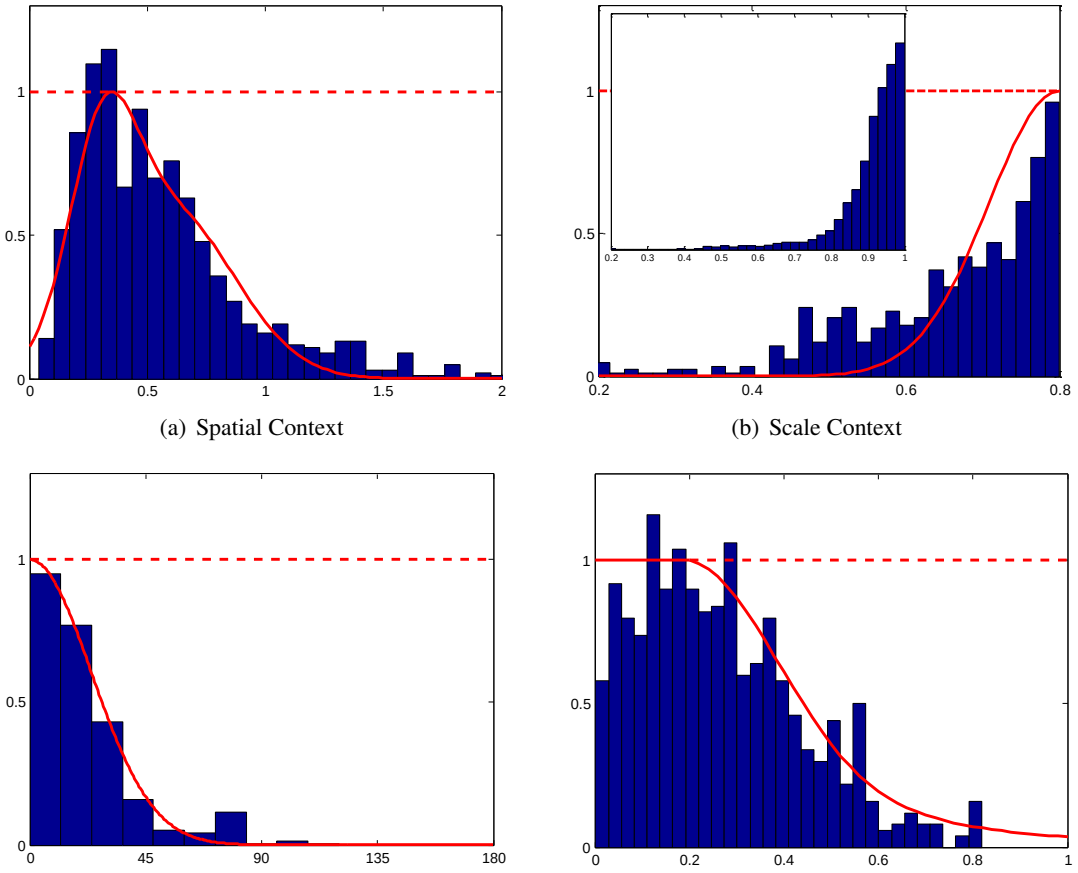
$$f_{\text{la}} = (B_i \cap B_j)/(B_i \cup B_j) \tag{4.8}$$

where B_i and B_j are bounding boxes of x_i and x_j, respectively.

4.3.2 Model Parameter Fitting

We adopt a data-driven approach to learn a mapping function $\mathcal{F} : f \rightarrow e$ for each contextual pattern, using 120 images from Structured Group Dataset *SGD* [133] that are independent of the evaluation subset. Since these images were captured at bus stops, classrooms, cafeterias, conferences, libraries, and parks, the contextual information between human beings in these images can help in estimating the parameters of our mapping functions. The PAV features are applied to the ground truth annotations at first, to obtain poses and orientations for modeling social context. Basically, we use the Gaussian kernels to model mapping functions $\mathcal{F}$, which has been widely used in constructing affinity graph [134, 140]. Parameters of these mapping functions are estimated by fitting data to the distributions of the 4 context patterns, which are illustrated in Figure 4.5. Single Gaussian is approximated by computing mean and variance, while Gaussian mixture model (GMM) is estimated by expectation maximization (EM) [141].

We model the spatial attraction e^{+}_{sp} as a 2-component Gaussian distribution, as shown in Equation 4.9, which corresponds to a maximum influence around 0.4 human height and vanishes after 1.5. When $f_{\text{sp}} < 0.1$, *i.e.*, we set $e^{+}_{\text{sp}} = 0$ since people are commonly not quite close to each other due to comfortable zone. The two ranges $[0.1, 0.3)$ and $[0.3, 1.5)$ represent



(a) Spatial Context (b) Scale Context

(c) Social Context (d) Layout Context

Figure 4.5: Feature distributions of the four contexts.

different probabilities of human coexistence. In other words, a hypothesis is more likely to be a true detection if it is in the neighborhood of a valid human instance. When $f_{\text{sp}} < 0.1$, we set $e^{-}_{\text{sp}} = 1$ for the reason discussed above. However, we cannot conclude that a hypothesis is not a true positive when it is remote from any valid instances. Consequently, for $f_{\text{sp}} \geq 0.1$, we have $e^{-}_{\text{sp}} = 0$.

$$e^{+}_{\text{sp}} = \begin{cases} 0, & \text{if } f_{\text{sp}} < 0.1 \\ \exp\left(-\frac{(0.4-f_{\text{sp}})^2}{0.2^2}\right), & \text{if } 0.1 \leq f_{\text{sp}} < 0.3 \\ \exp\left(-\frac{(0.4-f_{\text{sp}})^2}{0.7^2}\right), & \text{if } 0.3 \leq f_{\text{sp}} < 1.5 \\ 0, & \text{otherwise} \end{cases} \tag{4.9}$$

$$e^{-}_{\text{sp}} = \begin{cases} 1 - e^{+}_{\text{sp}}, & \text{if } f_{\text{sp}} < 0.1 \\ 0, & \text{otherwise} \end{cases}$$

For scale attraction e_{sc}^{+}, we fit it a Gaussian function in the range of $[0, 0.8]$, shown in Equation 4.10. Clearly, we have $e_{\text{sc}}^{+} \to 1$ if $f_{\text{sc}} \to 1$, *i.e.*, two hypotheses are compatible to each other if they have similar scales. To counteract height differences of individuals and mild errors of detection bounding boxes, we set $e_{\text{sc}}^{+} = 1$ when $f_{\text{sc}} \in [0.8, 1]$. Since compatibility and inconsistence of scale context are mutually exclusive, we let $e_{\text{sc}}^{-} = 1 - e_{\text{sc}}^{+}$.

$$
\begin{aligned}
e_{\text{sc}}^{+} &= \begin{cases} \exp\left(-\frac{(0.8-f_{\text{sc}})^2}{0.2^2}\right), & \text{if } 0 < f_{\text{sc}} < 0.8 \\ 1, & \text{otherwise} \end{cases} \\
e_{\text{sc}}^{-} &= 1 - e_{\text{sc}}^{+}
\end{aligned}
\tag{4.10}
$$

In Equation 4.11, the attraction strength of social context is set to 1 if $f_{\text{so}} \geq \pi/6$, since we assume strong social interaction happening between two human hypotheses in such scenario. In range $[\pi/6, \pi]$, e_{so}^{+} is modeled as a Gaussian function of f_{so}. Similarly to e_{sp}^{-}, weak social context does not indicate existence of false positives. Hence, we have $e_{\text{so}}^{-} = 0$.

$$
\begin{aligned}
e_{\text{so}}^{+} &= \begin{cases} 1, & \text{if } f_{\text{so}} < \pi/6 \\ \exp\left(-\frac{(f_{\text{so}}-\pi/6)^2}{(\pi/6)^2}\right), & \text{if } f_{\text{so}} \geq \pi/6 \end{cases} \\
e_{\text{so}}^{-} &= 0
\end{aligned}
\tag{4.11}
$$

The layout context e_{la}^{+} is also defined as piecewise-defined function, as shown in Equation 4.12. When f_{la} is small, two human hypotheses are more likely to coexist with partial overlap; thereby, we set e_{la}^{+} close to 1. Otherwise, large repulsion should be configured between two hypotheses, *i.e.*, $e_{\text{la}}^{-} \to 1$, in order to keep a stronger detection, in the meanwhile, ignoring the other.

$$
\begin{aligned}
e_{\text{la}}^{+} &= \begin{cases} 1, & \text{if } f_{\text{la}} < 0.4 \\ \exp\left(-\frac{(0.4-f_{\text{la}})^2}{0.2^2}\right), & \text{if } 0.4 \leq f_{\text{la}} < 0.8 \\ 0, & \text{otherwise} \end{cases} \\
e_{\text{la}}^{-} &= 1 - e_{\text{la}}^{+}
\end{aligned}
\tag{4.12}
$$

Note that scale and overlap contexts are *deductive* patterns, i.e., they are discriminative to tell whether a hypothesis is true or not. For instance, if a hypothesis goes against true detections with regard to scale (small e_{sc}^{+}), then a strong repulsion should be given (large e_{sc}^{-}). On the opposite, spatial and social cues are not deductive, suggesting that we cannot infer whether or

not a hypothesis is invalid, even if it is remote from true detections or no social interactions are observed. To sum up, we have four mapping functions for the contextual features exploited in this method, as shown in Equation 4.13.

$$\begin{aligned}
&\mathcal{F}_{\text{sp}} : f_{\text{sp}} \in (0, +\infty) \rightarrow e_{\text{sp}} \in [0, 1] \\
&\mathcal{F}_{\text{sc}} : f_{\text{sc}} \in (0, 1] \rightarrow e_{\text{sc}} \in [0, 1] \\
&\mathcal{F}_{\text{so}} : f_{\text{so}} \in [0, \pi] \rightarrow e_{\text{so}} \in [0, 1] \\
&\mathcal{F}_{\text{la}} : f_{\text{la}} \in [0, 1] \rightarrow e_{\text{la}} \in [0, 1]
\end{aligned} \tag{4.13}$$

4.4 Progressive Potential Propagation

Label propagation has been widely used in graph-based semi-supervised learning (GSSL) [134, 142] to perform transductive inference. We employ label propagation in our problem to estimate confidence obtained from contextual information while with several specific modification. Besides, since we do not have any 'labeled' data in our case, we propose a greedy-like technique, namely progressive potential propagation, to iteratively verify one hypothesis as a true detection in one run and use it for propagation in the next.

4.4.1 Potential Propagation

Given a context graph $\mathcal{G}$, the first question arising is how the contextual potential $\psi^{\mathcal{G}}(x_i, \mathbf{Y})$ ($i \in [1 : m]$) in Equation (4.2) can be obtained. Suppose that we have a label vector $\mathbf{Y} \in \mathbb{R}^m$ for m hypotheses, we first initialize a potential vector $\mathbf{Z} \in \mathbb{R}^m$ as $\mathbf{Y}$. Since the contextual potential of hypothesis x_i is targeted, we set $z_i = 0$ to avoid *self-reinforcement* (defined in [134]). Under such a setting, a hypothesis $x_j (j \in [1 : m])$ can be seemed as labeled when $y_j \neq 0$ and unlabeled otherwise. Intuitively, strong true detections is more robust in propagating potential, therefore we need to re-weight the context graph $\mathcal{G}$. We apply logistic regression to normalize the unary score $\psi^u(x_j)(j \in [1 : m])$ into $w_j \in (0, 1)$, which can be regarded as the true detection probability of hypothesis x_j; we set $w_j = 1$ if $y_j = 1$, since $y_j = 1$ means validated true detection. Then each column of $\mathcal{G}$ is re-weighted by w_j, i.e., $\mathcal{G}'_{\cdot j} = w_j \mathcal{G}_{\cdot j}$. Matrix $\mathcal{G}'$ is further row-normalized such that $\overline{\mathcal{G}}_{ij} = \mathcal{G}_{ij} / \sum_k |\mathcal{G}'_{ik}|$, which is critical for the convergence of the propagation algorithm. We summarize the potential propagation algorithm in Algorithm 1.

Algorithm 1 Potential propagation for $\psi^{\mathcal{G}}(x_i, \mathbf{Y})$

1: **Input:** given $\mathcal{G}$,$\mathbf{Y}$.
2: **Output:** $\psi^{\mathcal{G}}(x_i, \mathbf{Y})$
3: Initialize $\mathbf{Z} = \mathbf{Y}$, $z_i = 0$
4: Obtain $\overline{\mathcal{G}}$ by re-weighting and row-normalization
5: **while** $\mathbf{Z}$ does not converge **do**
6: $\quad \mathbf{Z} \leftarrow \overline{\mathcal{G}}\mathbf{Z}$
7: $\quad \mathbf{Z} = \max(\mathbf{Z}, 0)$
8: $\quad \forall j \in [1:m], j \neq i$, if $y_j = 1$, $z_j = 1$,
9: **end while**
10: $\psi^{\mathcal{G}}(x_i, \mathbf{Y}) \leftarrow \overline{\mathcal{G}}\mathbf{Z}$

In line 6, the potential is propagated based on $\overline{\mathcal{G}}$ and the potential vector $\mathbf{Z}$ in the previous iteration. In line 7, we reset all negative elements in potential vector $\mathbf{Z}$ back to zero, to cut off the potential defused by false alarms. The reason behind is: even there is a strong exclusion between a false positive and a hypothesis, it would be still hard to determine whether the hypothesis is a true detection or a false alarm. True detections definitely expose repulsions to false positives. However, since false alarms are heterogeneous, they probably also appear intense exclusive patterns. Similar to [134], in line 8, we replenish elements with initial value 1. The propagation algorithm repeats from line 6 to line 8 until $\mathbf{Z}$ converges. One more propagation is executed in line 10, in order to output the $\psi^{\mathcal{G}}(x_i, \mathbf{Y})$ with both positive and negative contextual potential.

4.4.2 Progressive Inference

Given the contextual potential $\psi^{\mathcal{G}}(x_i, \mathbf{Y})$, it is still NP-hard to optimize the objective function in Equation (4.1). We thus combine the potential propagation algorithm with a greedy forward search, aiming at a sub-optimal $\hat{\mathbf{Y}} = \max_{\mathbf{Y}} \Psi(\mathbb{X}, \mathbf{Y})$. Different from conventional graph-based propagation with fixed 'labeled' instances, we progressively validate unconfirmed hypotheses and propagate their potential in next run. Such approximation is similar to learning structured output, which has been applied in [45, 46]. Let $S^t = \{i | y_i = 1, i \in [1:m]\}$ denote the confirmed set of hypotheses at t iteration. We define the potential change by instancing

Algorithm 2 Progressive inference for $\hat{\mathbf{Y}}$,$\hat{S}$

1: **Input:** given $\mathcal{G}$, $\mathbb{X}$
2: **Output:** $\hat{\mathbf{Y}}$, $\hat{S}$
3: Initialize $\mathbf{Y} = 0$ and $S^0 = \emptyset$
4: First instance $i^* = \arg\max_i \psi^u(x_i)$, set $S^1 = \{i^*\}$.
5: **while** $\triangle(x_i) > 0$ **do**
6: $\quad \forall i \in [1:m], y_i \neq 1, i^* = \arg\max_i \triangle(x_i)$ (Algorithm1)
7: $\quad$ Update $\mathbf{Y}$: $y_{i^*} = 1$
8: $\quad S^{t+1} \leftarrow S^t \cup i^*$
9: $\quad t+1 \leftarrow t$
10: **end while**
11: $\hat{\mathbf{Y}} \leftarrow \mathbf{Y}$, $\hat{S} \leftarrow S^t$

hypothesis x_i as follows:

$$\begin{aligned} \triangle(x_i) &= \Psi(\mathbb{X}, \mathbf{Y}(S^t \cup i)) - \Psi(\mathbb{X}, \mathbf{Y}(S^t)) \\ &= \psi^u(x_i) + \alpha \left(\psi^{\mathcal{G}}(x_i, \mathbf{Y}(S^t)) + \sum_{j \in S^t} \psi^{\mathcal{G}}(x_j, \mathbf{1}(i)) \right) \end{aligned} \tag{4.14}$$

where $\psi^u(x_i)$ is the unary potential; $\psi^{\mathcal{G}}(x_i, \mathbf{Y}(S^t))$ measures the contextual potentials hypothesis x_i obtains from instanced hypotheses in S^t; $\mathbf{Y}(S^t)$ is the label vector that $y_k = 1$, if $k \in S^t$ and 0 otherwise; $\sum_{j \in S^t} \psi^{\mathcal{G}}(x_j, \mathbf{1}(i))$ represents potentials that hypothesis x_i imposes onto instance(s) in S^t, where $\mathbf{1}(i)$ is an indicator vector with only ith element equals 1 and others are 0s. These two terms can be achieved by using potential propagation algorithm proposed in Section 4.4.1. Then we devise our progressive inference algorithm as detailed in Algorithm 2.

The algorithm starts with an empty set S and a zero vector $\mathbf{Y}$. We select the first hypothesis according to unary potential only. During each iteration, we instance one unconfirmed hypothesis with the largest potential change $\triangle(x_i)$ defined in Equation (4.14), and update S^t and $\mathbf{Y}$ accordingly. The algorithm runs line 6 to 9 repetitively and instance one hypothesis in each iteration, until adding any other detections could not enhance the total potential $\Psi(\mathbb{X}, \mathbf{Y})$. Obviously, by growing S^t alternatively, contextual potentials from true detections are progressively propagated. An illustration of the progressive inference with potential propagation is shown in Figure 4.6. When the algorithm ceases, the hypotheses in $\widehat{S}$ are regarded as true detections while others as false alarms. We further rescore all detections by summing up their

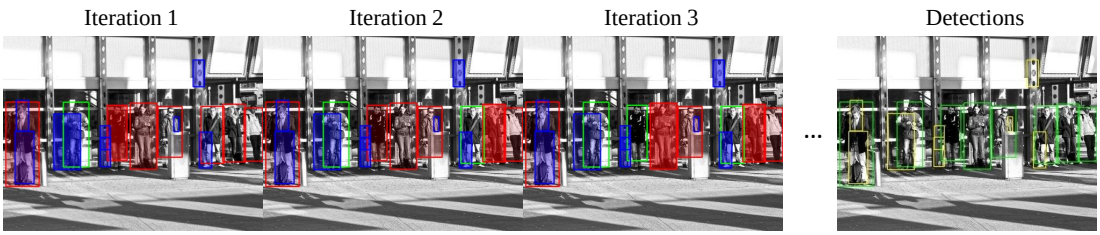


Figure 4.6: Illustration of progressive inference with potential propagation (image from VOC 2012). Iteration 1 selects the first hypothesis (green bounding box) and propagates contextual potentials (positive or negative) to others in the image. A red bounding box indicates a hypothesis getting positive potentials while a blue one receives negative potentials. The darkness of colors shows the amounts of their contextual potentials. After iteration 1, the algorithm picks the hypothesis with the highest potential change composed of both unary and contextual potentials and then starts the second iteration with 2 instanced human hypotheses. The process repeats until no hypothesis has a positive potential gain. In this example, our algorithm ends in 8 iterations, resulting in 8 true detections.

unary and contextual potentials, i.e., we have,

$$\psi'(x_i) = \psi^u(x_i) + \alpha \left(\psi^{\mathcal{G}}(x_i, \hat{\mathbf{Y}}) + \sum_{j \in \hat{S}} \psi^{\mathcal{G}}(x_j, \mathbf{1}(i)) \right) \tag{4.15}$$

where $\psi'(x_i)$ could be positive or negative. After rescoring, value 0 is further used as the cutoff threshold to differentiate true detections and false alarms (as shown in Figure 4.8). Although the value of this operational point is fixed, the detection confidences of human hypotheses are adaptive to images of different detection outputs. As a result, our method is more flexible to detecting human in images of different scenarios. Differently, a traditional human detector (*e.g.*, in [55, 7]) needs a threshold of fixed value to decide whether a detection is true or not. Without a doubt, an empirical setting of this threshold value would not be applicable to every testing image, which hurts the performance of human detection in practice.

4.5 Experimental Results

4.5.1 Datasets

We evaluate our proposed approach on two public datasets: Structured Group Dataset (*SGD*)[1] [133] and *ETH* pedestrian dataset (*ETH*) [143]. *SGD* presents people with variant poses and layouts while *ETH* contains only pedestrians in low image resolution.

[1] http://cvgl.stanford.edu/projects/groupdiscovery/

There is a total of 599 images in *SGD*, taken from 6 scenarios (bus stops, classrooms, cafeterias, conferences, libraries, and parks). Crowds in different scenes show various layouts, e.g., queuing, standing in line, sitting in a circle, etc. More than $5,000$ people were annotated with tight bounding boxes, torso orientations, and poses (sitting or standing). We randomly choose 20 images from each scenario to form a training subset of 120 images for fitting model parameters (Section 4.3.1) and learning the pose classifier. Images without sufficient information for horizon line estimation are excluded, leading to 308 images totally for evaluation.

ETH is a standard benchmark for pedestrian detection, containing videos captured in urban settings by a pair of cameras mounted on a chariot. In our experiments, we explore all left video sequences from "Setup 1", which include $1,804$ images and a total of $14,167$ annotated human instances down to a size of about 48 pixels. All these $1,804$ images were used for evaluation.

4.5.2 Experimental Setup

Our approach can take input from any detector. On *SGD* dataset, we choose poselet [55] as the underlying detector for its good ability of handling pose variations. On *ETH*, DPM (LatSvm-V2) [7] is used to verify our proposed method. Parameter α in Equation (4.1) weighs the contextual potential over unary scores. Since detection scores of different detectors are normally not within the same range, α needs to be empirically determined for each underlying detector used in our approach. In our experiments, we set α to 3.0 for poselet, and 10.0 for DPM, respectively. For evaluation purpose, we use the popular PASCAL interaction-over-union (IoU) as the measurement to verify the correctness of a human hypothesis. Unless otherwise specified, we set IoU as 0.5 for reporting performance.

We first briefly evaluate the performance of pose classification and body orientation estimation on the *SGD* dataset. Over 379 test images, the method [138] achieves an overall accuracy of 84.27% (standing: 81.64%; sitting: 86.41%) for pose classification and a mean error of 20.7° for body orientation estimation. Some examples of orientation estimation are demonstrated in Figure 4.7. Magenta lines represent the head orientations of the human instances and the cyan ones indicate the body orientations. It can be observed that the estimated orientations are accurate enough in capturing their social interactions. The results suggest that reliable social context can be integrated into the contextual graph by using these orientations .

Figure 4.7: Some samples of orientation estimation on ground truth annotations. Magenta lines represent head orientations, while the cyan ones are for body orientation.

Method	Recall	Precision	F1-score
Proposed	0.6686	0.7829	**0.7212**
Poselet	0.6686	0.7037	0.6857
	0.6280	0.7829	0.6969

Table 4.1: Evaluation on *SGD*: recalls and precisions of our proposed method at the operational point in comparison with poselet detector.

4.5.3 Results

Performance on *SGD*. We validate our approach with $6,161$ human hypotheses from poselet that are above a unary detection threshold of 0.5, and compared the results with those of poselet [55], and DPM (LatSvm-V2 [7]).As shown in Figure 4.8, our approach outperforms the baseline detector, demonstrating the effectiveness of context modeling. Poselet is improved from 0.669 to 0.684 with regard to average precision (AP).

To further understand the improvement, we report in Table 4.1 the precisions and recalls corresponding to the default operational point of our approach (red dot in Figure 4.8). At the same recall, our approach yields a much higher precision (78.3% vs. 70.4%) than the baseline detector, i.e., poselet, while at the same precision, it improves the recall by 4.1% (66.9% vs. 62.8%).

Performance on *ETH*. We use DPM as the underlying detector for our approach due to its

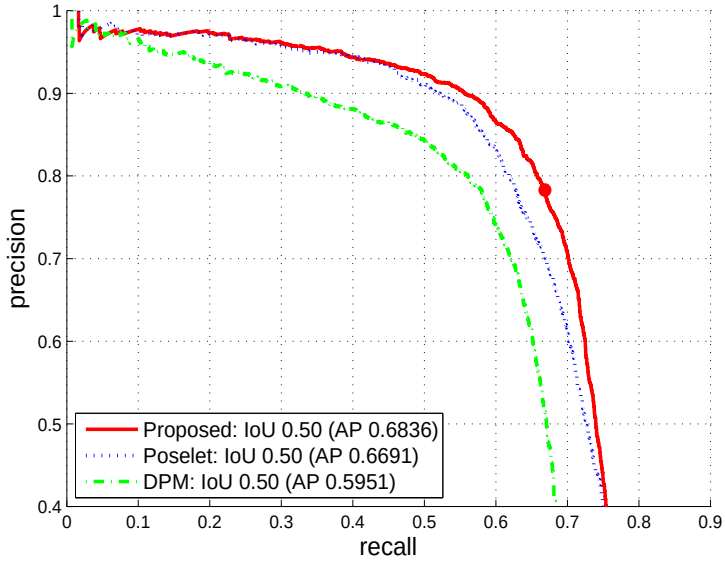


Figure 4.8: Recall-precision curves of different approaches on *SGD* dateset. The operational point of our approach (at the cutoff threshold of 0.0) is marked as a red dot on the corresponding curve.

better performance on *ETH*. The threshold of DPM is set to -0.9, generating a total of $41,744$ pedestrian hypotheses as inputs to our approach. PAV features for predicting pose and body orientations are extracted based on their bounding boxes. In addition to poselet and DPM (LaSvm-V2), other pedestrian detection methods, such as HOG [6], ConvNet [144], MultiSDP [145], JointDeep [146], Roerei [147], SDN [148], Franken [149], and SpatioPooling [150] are included for comparison, following the evaluation routine in [41][2]. These detectors, except HOG and DPM, are top methods proposed recently for pedestrian detection. As can be seen in Figure 4.9, using DPM as the underlying detector (Proposed–DPM), our approach is as comparative as SDN, achieving a log-average miss rate of 43% and performing better than other deep learning methods including ConvNet, MultiSDP, and JointDeep. The results are encouraging as our method is built upon some well-developed detectors without requiring massive training data and computational training time. We also apply our approach to a strong detector, i.e., Proposed–SDN. The results show that our approach can further reduce the miss rate of SDN from 41% to 39%.

[2] `http://www.vision.caltech.edu/Image_Datasets/CaltechPedestrians/`

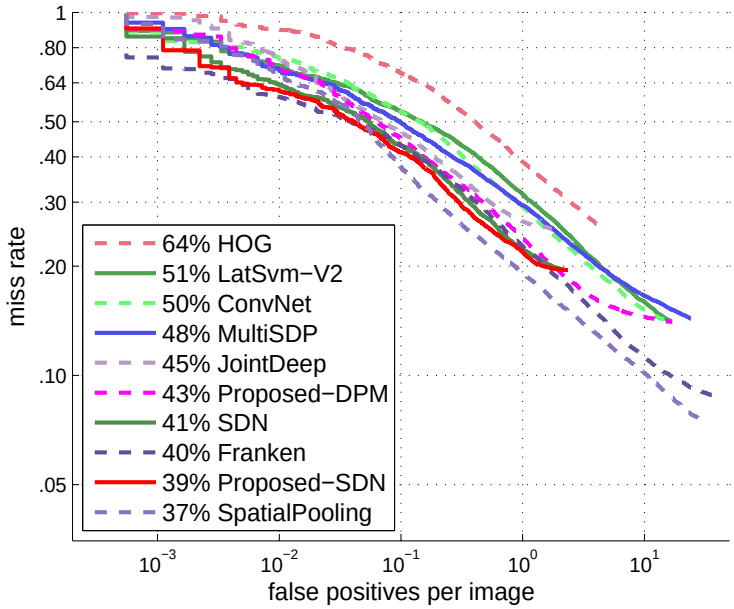


Figure 4.9: Overall performance on *ETH* dateset in terms of miss rates and false positives per image.

Sampled Detections We illustrate a few sampled detections by our proposed approach in Figure 4.10, in which correct detections from the underlying detectors are colored as red while additional detections discovered by our approaches marked as green. These results clearly demonstrate the efficacy of our approach in context modeling. In the second image, we observe a child (marked with a yellow bounding box) incorrectly being suppressed by our approach due to scale inconsistency, since we assume people have roughly equal heights. The last image indicates two false alarms in our approach, which actually highly resemble a human.

4.6 Discussions

4.6.1 Ablation Study

We perform an ablation study on the *SGD* dataset to understand the contribution of each contextual cue considered in our approach, in regards of average precision. We single out one cue each time and reported the performance in Table 4.2. Apparently, removing any single cue from the context graph leads to deteriorative results, suggesting that all the cues are helpful for people detection. Among them, overlap context contributes most to the final performance as it

Figure 4.10: Sampled detections by our proposed approach. Correct detections from the underlying detector are colored as red while detections discovered by our approaches marked as green. Red arrows point out some failed cases in our approach.

Dataset	Baseline	-Scale	-Spatial	-Social	-Overlap	All
SGD	0.6691	0.6632	0.6702	0.6755	0.6620	0.6836

Table 4.2: Effects of different contextual patterns, in regards of average precision. Contextual information was discarded respectively.

acts as a primary force in suppressing false alarms.

As described in Section 4.4.2, our algorithm iteratively picks the best 'true detection' with the largest potential gain at each iteration, and then in the next run uses it as a 'labeled' instance to propagate contextual potential to other unconfirmed hypotheses. To validate the effectiveness of such a progressive fashion, we compare it with a 'threshold-based' method that only does potential propagation once using high-confidence hypotheses, since strong hypotheses are usually associated with true detections.

As can be seen in Table 4.3, a large threshold (e.g., 20) leads to fewer 'labeled' data samples, thus is incompetent to suppress false alarms (lower precision). Oppositely, a small threshold (e.g., 1) would take many false positives as 'labeled' data that would propagate potentials improperly to true detections (lower recall). As a comparison, our progressive inference is adaptive and can grasp a good trade-off between precision and recall.

Method	Proposed	Threshold				
		20	10	5	1	$-\infty$
F1-score	**0.7212**	0.6482	0.6971	0.7115	0.7029	0.6894
Recall	0.6686	0.7293	0.7047	0.6814	0.6036	0.5728
Precision	0.7829	0.5834	0.6896	0.7444	0.8414	0.8656

Table 4.3: Performance comparisons of our progressive inference and the threshold-based approach on *SGD*, in terms of F1-score at the default operational point.

4.6.2 Differences against Conventional GSSL

As discussed above, the motivation of our method is similar to that of conventional graph-based semi-supervised learning (GSSL). However, there are several critical differences between our method and the conventional GSSL.

1. Typically, GSSL uses similarities (*e.g.*, Euclidean distance) between data samples to construct an affinity graph. Here, we model the instance-level contextual interaction as the 'similarity' to propagate potential, resulting in the context graph.

2. GSSL employs annotated data samples as seeds to propagate labeling. In our detection problem, we do not have any labeled hypothesis. Since we assume strong detections in crowd scenes are indeed valid human instances, we start potential propagation using these strong detections as pseudo annotated seeds.

3. The conventional assumption of GSSL is that both positive and negative data samples have been labeled, such that unlabeled instances would receive either positive and negative potential. However, false positives (could be regarded as negative) in human detection can be used to infer the existence of neither true detections nor false alarms. In order to propagate negative labeling, we design edges of negative weights in the context graph, by introducing contextual repulsion.

4. In GSSL, the final labeling of unannotated samples only rely on their neighborhood. However, the inputs of our method are detections from an underlying detector. Thereby, the unary detection confidences cannot be ignored. Our method is different from GSSL that we formulate human detection as an optimization problem, which simultaneously leverages both unary and contextual confidences.

4.7 Summary

In this chapter, we have proposed a novel approach to improve human detection in crowded scenes by exploring contextual cues. Our approach have modeled people interactions through a context graph, via attraction and repulsion built up on both geometric and social cues available in crowded scenarios. Four kinds of contextual information have been employed, including spatial context, scale context, social context, and layout context. Then, potential can be progressively spread by label propagation, such that contextually compatible human hypotheses would be reinforced by receiving positive potentials while false alarms would be contained due to being negated by contextual incompatibility. We have shown results comparable to state of the art on two public datasets for people and pedestrian detection. With a human detector of a shallow model, our method has achieved comparative results to deep ConvNet based detector. Besides, our method has reduced the miss rate of SDN detector [148] by $\%2$, indicating its flexibility to any underlying human detector.

Chapter 5

Conclusions and Future Work

5.1 Conclusions

The dissertation has exploited multispectral and contextual information for human detection.

First, we have employed multispectral images (color and thermal channels) for human detection. Rather than using handcrafted image features, we have used features extracted from deep convolutional neural networks (ConvNets) to represent human objects. Faster R-CNN framework has been first investigated for human detection, which has achieved state-of-art results on Caltech pedestrian benchmark. We have leveraged Faster R-CNN as our vanilla network and our study on such vanilla network has shown that there exists promising complementary potential between RGB and thermal images. We have modeled the multispectral human detection a fusion problem of deep ConvNets. Motivated by the hypothesis that fusion at different stages of neural networks would lead to distinct detections, we have proposed four fusion models, *i.e.*, *Early Fusion*, *Halfway Fusion*, *Late Fusion*, and *Score Fusion*, corresponding to low-level, middle-level, high-level feature fusions and confidence fusions. Based on the corrected annotations of KAIST multispectral pedestrian benchmark, our fusion models have all yielded better detection results, compared to human detectors of using one single image modality. Overall, the *Score Fusion* model has achieved the best performance on multispectral human detection, due to cascade structure. In the context of feature fusion, *Halfway Fusion* has obtained the lowest MR, indicating that middle-level feature fusion performs better than low-level and high-level feature fusion. We have also demonstrated experimental results in terms of various scales and occlusion levels, giving more insights on human detection of exploiting color and thermal images simultaneously. Besides, we have also discussed why deep ConvNets work better than handcrafted features for human detection and have mentioned several bottlenecks of current multispectral detection pipeline.

In the second part of this dissertation, we have proposed a unified framework that improves human detection by leveraging the instance-level contextual information extracted in crowd scenes. In order to model the contextual interactions between human hypotheses generated from any underlying detector, we have exploited four kinds of contexts, *i.e.*, spatial context, scale context, social context, and layout context. The *context graph* has been developed, considering both contextual compatibility and inconsistency in the four contexts. Since both positive and negative potentials can be propagated through the graph, our method has shown the capability of boosting weak true detection, in the meanwhile, suppressing false positives. Human detection has been formulated an optimization problem, the optimum of which has been approximated by the proposed progressive potential propagation. Our method has achieved promising results on two public datasets. The experimental results have demonstrated that the performance of a shallow human detector improved by our approach is comparable to some deep ConvNet based detectors.

5.2 Directions of Future Work

5.2.1 Better Multispectral Image

Multispectral images of a higher resolution would definitely improve the human detection performance. However, if high-resolution multispectral images are exploited, the trade-off between accuracy and computational complexity of the human detector should be investigated. Regression based deep ConvNet detectors (*e.g.*, YOLO [151]) or single shot detector (*e.g.*, SSD [152]) could be considered which have shown preliminary results on generic object detection. Besides, image alignment is another crux for multispectral human detection, especially in detecting people of small image sizes. We have shown mismatching between color and thermal images in KAIST benchmark which are aligned based on hardware design. Such misalignment could be tempered through camera calibration or image registration (*e.g.*, key point matching or disparity estimation). Additionally, the raw image of a thermal camera has relatively low contrast. Although thermal image enhancement has been studied in [153], it is more interesting to develop task-specific enhancement techniques similar to the idea of saliency detection, *i.e.*, only potential image areas are enhanced.

5.2.2 Smarter Fusion Scheme

In our multispectral deep ConvNets, we assume that fusions of features at different network stages would lead to distinct detection performance. In our fusion models, the concatenations of convolutional features are predefined. For instance, *Halfway Fusion* model connects the fourth layer of the color branch to the fourth one of the thermal. However, an interesting question arises: how is the performance if we combine low-level features from color images with middle-level features extracted from thermal inputs? Since thermal imaging is good at capturing global shape while color channels carry more visual detail, it is possible that such fusion works better. The idea of recurrent rolling convolution [154] could be applied here to let a recurrent scheme adaptively learn the fusion.

5.2.3 More Information Modalities

Beyond the thermal channel, in fact, there is some other information we can leverage for human detection. If images are captured by fixed cameras, we can exploit additional motion patterns, *e.g.*, moving foreground extraction (such as in [155]) or optical flow. With a binocular camera, stereo estimation can be regarded as a 'depth' prior in human detection. Besides, dense Lidar signal can also be used to verify human instance. Recently, we have witnessed great improvement of semantic segmentation. It would be worthy of trying to fuse color images and semantic masks for human detection.

5.2.4 Deeper Context Model

Deep learning has been employed for context modeling for human detection, such as LSTM [68]. Besides, deep reinforcement learning has also applied to object detection, showing promising results in localization [156, 157]. However, these methods designed the policies and actions of their agents individually based on the deep features of each single object. If we have semantic masks (such as sky, road, grown, tree, and building), detections of objects of several categories (such as car and traffic sign), better deep reinforcement learning approach can be developed for human detection, by using the high-level context information.

References

[1] R. Cutler and L. Davis, "Look who's talking: Speaker detection using video and audio correlation," in *IEEE International Conference on Multimedia and Expo*, vol. 3. IEEE, 2000, pp. 1589–1592.

[2] C. Wu, Z. Yang, Z. Zhou, X. Liu, Y. Liu, and J. Cao, "Non-invasive detection of moving and stationary human with wifi," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 11, pp. 2329–2342, 2015.

[3] P. Viola, M. J. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," *International Journal of Computer Vision*, vol. 63, no. 2, pp. 153–161, 2005.

[4] [Online]. Available: https://en.wikipedia.org/wiki/Amazon_Go

[5] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele, "How far are we from solving pedestrian detection?" *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[6] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, 2005, pp. 886–893.

[7] 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*, vol. 32, no. 9, pp. 1627–1645, 2010.

[8] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 8, pp. 1532–1545, 2014.

[9] 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*, vol. 34, no. 4, pp. 743–761, 2012.

[10] R. Benenson, M. Omran, J. Hosang, and B. Schiele, "Ten years of pedestrian detection, what have we learned?" in *Proceedings of European Conference on Computer Vision*. Springer, 2014, pp. 613–627.

[11] M. Enzweiler and D. M. Gavrila, "Monocular pedestrian detection: Survey and experiments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 12, pp. 2179–2195, 2009.

[12] R. Muñoz-Salinas, E. Aguirre, and M. García-Silvente, "People detection and tracking using stereo vision and color," *Image and Vision Computing*, vol. 25, no. 6, pp. 995–1007, 2007.

[13] K. Yamaguchi, T. Hazan, D. McAllester, and R. Urtasun, "Continuous markov random fields for robust stereo estimation," *Proceedings of European Conference on Computer Vision*, pp. 45–58, 2012.

[14] W. Luo, A. G. Schwing, and R. Urtasun, "Efficient deep learning for stereo matching," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5695–5703.

[15] L. Xia, C.-C. Chen, and J. K. Aggarwal, "Human detection using depth information by kinect," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 2011, pp. 15–22.

[16] J. W. Davis and M. A. Keck, "A two-stage template approach to person detection in thermal imagery," in *Proceedings of IEEE Winter Conference on Applications of Computer Vision*, vol. 1. IEEE, 2005, pp. 364–369.

[17] C. Dai, Y. Zheng, and X. Li, "Pedestrian detection and tracking in infrared imagery using shape and appearance," *Computer Vision and Image Understanding*, vol. 106, no. 2, pp. 288–299, 2007.

[18] M. Szarvas, U. Sakai, and J. Ogata, "Real-time pedestrian detection using lidar and convolutional neural networks," in *Proceedings of IEEE Symposium on Intelligent Vehicles*. IEEE, 2006, pp. 213–218.

[19] C. Premebida, O. Ludwig, and U. Nunes, "Lidar and vision-based pedestrian detection system," *Journal of Field Robotics*, vol. 26, no. 9, pp. 696–711, 2009.

[20] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 580–587.

[21] T. A. Cleland, *Contrast Enhancement*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 876–880. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-29678-2_1250

[22] M. Vanrell, F. Lumbreras, A. Pujol, R. Baldrich, J. Llados, and J. J. Villanueva, "Colour normalisation based on background information," in *Proceedings of International Conference on Image Processing*, vol. 1. IEEE, 2001, pp. 874–877.

[23] P. Viola, M. J. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2003, pp. 734–741.

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

[25] P. Dollár, S. Belongie, and P. Perona, "The fastest pedestrian detector in the west." in *Proceedings of British Machine Vision Conference*, vol. 2, no. 3, 2010, p. 7.

[26] S. Paisitkriangkrai, C. Shen, and A. van den Hengel, "Pedestrian detection with spatially pooled features and structured ensemble learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014.

[27] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[28] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

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

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

[31] L. Zhang, L. Lin, X. Liang, and K. He, "Is faster R-CNN doing well for pedestrian detection?" in *Proceedings of European Conference on Computer Vision*. Springer, 2016, pp. 443–457.

[32] D. Tang, Y. Liu, and T.-K. Kim, "Fast pedestrian detection by cascaded random forest with dominant orientation templates." in *Proceedings of British Machine Vision Conference*, 2012, pp. 1–11.

[33] J. Marin, D. Vázquez, A. M. López, J. Amores, and B. Leibe, "Random forests of local experts for pedestrian detection," in *Proceedings of IEEE International Conference on Computer Vision*, 2013, pp. 2592–2599.

[34] M. Andriluka, S. Roth, and B. Schiele, "Pictorial structures revisited: People detection and articulated pose estimation," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1014–1021.

[35] X. Wang, M. Yang, S. Zhu, and Y. Lin, "Regionlets for generic object detection," in *Proceedings of IEEE International Conference on Computer Vision*, 2013, pp. 17–24.

[36] R. B. Girshick, P. F. Felzenszwalb, and D. McAllester, "Discriminatively trained deformable part models, release 5," http://people.cs.uchicago.edu/ rbg/latent-release5/.

[37] S. Rujikietgumjorn and R. T. Collins, "Optimized pedestrian detection for multiple and occluded people," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3690–3697.

[38] J. Yan, Z. Lei, D. Yi, and S. Z. Li, "Multi-pedestrian detection in crowded scenes: A global view," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 3124–3129.

[39] D. M. Powers, "Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation," 2011.

[40] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: A benchmark," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 304–311.

[41] 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*, vol. 34, no. 4, pp. 743–761, 2012.

[42] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proceedings of Advances in Neural Information Processing Systems*, 2015, pp. 91–99.

[43] S. Hwang, J. Park, N. Kim, Y. Choi, and I. So Kweon, "Multispectral pedestrian detection: Benchmark dataset and baseline," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1037–1045.

[44] S. Pellegrini, A. Ess, and L. Van Gool, "Improving data association by joint modeling of pedestrian trajectories and groupings," in *Proceedings of European Conference on Computer Vision*, 2010, pp. 452–465.

[45] C. Desai, D. Ramanan, and C. C. Fowlkes, "Discriminative models for multi-class object layout," *International Journal of Computer Vision*, vol. 95, no. 1, pp. 1–12, 2011.

[46] J. Yan, X. Zhang, Z. Lei, S. Liao, and S. Z. Li, "Robust multi-resolution pedestrian detection in traffic scenes," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[47] P. Sabzmeydani and G. Mori, "Detecting pedestrians by learning shapelet features," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2007, pp. 1–8.

[48] P. Dollár, Z. Tu, P. Perona, and S. Belongie, "Integral channel features," in *Proceedings of British Machine Vision Conference*, 2009.

[49] W. Nam, P. Dollár, and J. H. Han, "Local decorrelation for improved pedestrian detection," in *Proceedings of Advances in Neural Information Processing Systems*, 2014, pp. 424–432.

[50] B. Leibe, E. Seemann, and B. Schiele, "Pedestrian detection in crowded scenes," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1. IEEE, 2005, pp. 878–885.

[51] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[52] T. Zhao and R. Nevatia, "Bayesian human segmentation in crowded situations," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2. IEEE, 2003, pp. II–459.

[53] M.-Y. Liu, O. Tuzel, A. Veeraraghavan, and R. Chellappa, "Fast directional chamfer matching," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 1696–1703.

[54] X. Wang, T. X. Han, and S. Yan, "An HOG-LBP human detector with partial occlusion handling," in *Proceedings of IEEE International Conference on Computer Vision*, 2009, pp. 32–39.

[55] L. Bourdev, S. Maji, T. Brox, and J. Malik, "Detecting people using mutually consistent poselet activations," in *Proceedings of European Conference on Computer Vision*, 2010, pp. 168–181.

[56] W. R. Schwartz, A. Kembhavi, D. Harwood, and L. S. Davis, "Human detection using partial least squares analysis," in *Proceedings of IEEE International Conference on Computer Vision*, 2009, pp. 24–31.

[57] Q. Zhu, M.-C. Yeh, K.-T. Cheng, and S. Avidan, "Fast human detection using a cascade of histograms of oriented gradients," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2. IEEE, 2006, pp. 1491–1498.

[58] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele, "Towards reaching human performance in pedestrian detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[59] Z. Lin and L. S. Davis, "A pose-invariant descriptor for human detection and segmentation," in *Proceedings of European Conference on Computer Vision*. Springer, 2008, pp. 423–436.

[60] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele, "Poselet conditioned pictorial structures," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 588–595.

[61] Y. Yang and D. Ramanan, "Articulated human detection with flexible mixtures of parts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2878–2890, 2013.

[62] D. Tran and D. A. Forsyth, "Configuration estimates improve pedestrian finding," in *Advances in neural information processing systems*, 2008, pp. 1529–1536.

[63] Y. Ding and J. Xiao, "Contextual boost for pedestrian detection," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2895–2902.

[64] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," in *Proceedings of European Conference on Computer Vision*. Springer, 2016, pp. 354–370.

[65] W. Ouyang and X. Wang, "Single-pedestrian detection aided by multi-pedestrian detection," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3198–3205.

[66] M. A. Sadeghi and A. Farhadi, "Recognition using visual phrases," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 1745–1752.

[67] H. Idrees, K. Soomro, and M. Shah, "Detecting humans in dense crowds using locally-consistent scale prior and global occlusion reasoning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015.

[68] R. Stewart, M. Andriluka, and A. Y. Ng, "End-to-end people detection in crowded scenes," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2325–2333.

[69] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[70] F. Xu, X. Liu, and K. Fujimura, "Pedestrian detection and tracking with night vision," *IEEE Transactions on Intelligent Transportation Systems*, vol. 6, no. 1, pp. 63–71, 2005.

[71] Y. Fang, K. Yamada, Y. Ninomiya, B. K. Horn, and I. Masaki, "A shape-independent method for pedestrian detection with far-infrared images," *IEEE Transactions on Vehicular Technology*, vol. 53, no. 6, pp. 1679–1697, 2004.

[72] H. Sun, C. Wang, B. Wang, and N. El-Sheimy, "Pyramid binary pattern features for real-time pedestrian detection from infrared videos," *Neurocomputing*, vol. 74, no. 5, pp. 797–804, 2011.

[73] F. Suard, A. Rakotomamonjy, A. Bensrhair, and A. Broggi, "Pedestrian detection using infrared images and histograms of oriented gradients," in *Proc. IEEE Intell. Veh. Symp.* IEEE, 2006, pp. 206–212.

[74] L. Zhang, B. Wu, and R. Nevatia, "Pedestrian detection in infrared images based on local shape features," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2007, pp. 1–8.

[75] D. Olmeda, C. Premebida, U. Nunes, J. M. Armingol, and A. de la Escalera, "Pedestrian detection in far infrared images," *Integr. Comput.-Aided Eng.*, vol. 20, no. 4, pp. 347–360, 2013.

[76] J. Portmann, S. Lynen, M. Chli, and R. Siegwart, "People detection and tracking from aerial thermal views," in *Proceedings of International Conference on Robotics and Automation*. IEEE, 2014, pp. 1794–1800.

[77] M. Teutsch, T. Muller, M. Huber, and J. Beyerer, "Low resolution person detection with a moving thermal infrared camera by hot spot classification," in *Proc. IEEE CVPR Workshops*, 2014, pp. 209–216.

[78] M. Bertozzi, A. Broggi, C. Caraffi, M. Del Rose, M. Felisa, and G. Vezzoni, "Pedestrian detection by means of far-infrared stereo vision," *Computer Vision and Image Understanding*, vol. 106, no. 2, pp. 194–204, 2007.

[79] S. J. Krotosky and M. M. Trivedi, "On color-, infrared-, and multimodal-stereo approaches to pedestrian detection," *IEEE Transactions on Intelligent Transportation Systems*, vol. 8, no. 4, pp. 619–629, 2007.

[80] Y. Yuan, X. Lu, and X. Chen, "Multi-spectral pedestrian detection," *Signal Process.*, vol. 110, pp. 94–100, 2015.

[81] J. H. Lee, J.-S. Choi, E. S. Jeon, Y. G. Kim, T. T. Le, K. Y. Shin, H. C. Lee, and K. R. Park, "Robust pedestrian detection by combining visible and thermal infrared cameras," *Sensors*, vol. 15, no. 5, pp. 10 580–10 615, 2015.

[82] A. González, Z. Fang, Y. Socarras, J. Serrat, D. Vázquez, J. Xu, and A. M. López, "Pedestrian detection at day/night time with visible and fir cameras: A comparison," *Sensors*, vol. 16, no. 6, p. 820, 2016.

[83] J. Wagner, V. Fischer, M. Herman, and S. Behnke, "Multispectral pedestrian detection using deep fusion convolutional neural networks," in *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2016, pp. 509–514.

[84] W. Ouyang, X. Zeng, and X. Wang, "Modeling mutual visibility relationship in pedestrian detection," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3222–3229.

[85] Y. Tian, P. Luo, X. Wang, and X. Tang, "Deep learning strong parts for pedestrian detection," in *Proceedings of IEEE International Conference on Computer Vision*, 2015, pp. 1904–1912.

[86] ——, "Pedestrian detection aided by deep learning semantic tasks," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5079–5087.

[87] A. Angelova, A. Krizhevsky, V. Vanhoucke, A. Ogale, and D. Ferguson, "Real-time pedestrian detection with deep network cascades," in *Proceedings of British Machine Vision Conference*, 2015.

[88] J. Li, X. Liang, S. Shen, T. Xu, and S. Yan, "Scale-aware fast r-cnn for pedestrian detection," *arXiv preprint arXiv:1510.08160*, 2015.

[89] S. Zhang, R. Benenson, and B. Schiele, "Filtered channel features for pedestrian detection," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1751–1760.

[90] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[91] J. Liu, S. Zhang, S. Wang, and D. N. Metaxas, "Multispectral deep neural networks for pedestrian detection," *arXiv preprint arXiv:1611.02644*, 2016.

[92] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proceedings of International Conference on Machine Learning*, 2011, pp. 689–696.

[93] N. Srivastava and R. R. Salakhutdinov, "Multimodal learning with deep boltzmann machines," in *Proceedings of Advances in Neural Information Processing Systems*, 2012, pp. 2222–2230.

[94] L. Wang, Y. Li, and S. Lazebnik, "Learning deep structure-preserving image-text embeddings," *arXiv preprint arXiv:1511.06078*, 2015.

[95] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proceedings of Advances in Neural Information Processing Systems*, 2014, pp. 568–576.

[96] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.

[97] R. Socher, B. Huval, B. Bath, C. D. Manning, and A. Y. Ng, "Convolutional-recursive deep learning for 3d object classification," in *Proceedings of Advances in Neural Information Processing Systems*, 2012, pp. 665–673.

[98] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik, "Learning rich features from rgb-d images for object detection and segmentation," in *Proceedings of European Conference on Computer Vision*. Springer, 2014, pp. 345–360.

[99] F. Yan and K. Mikolajczyk, "Deep correlation for matching images and text," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3441–3450.

[100] X. Wang, T. X. Han, and S. Yan, "An HOG-LBP human detector with partial occlusion handling," in *Proceedings of IEEE International Conference on Computer Vision*, 2009, pp. 32–39.

[101] S. Paisitkriangkrai, C. Shen, and A. van den Hengel, "Strengthening the effectiveness of pedestrian detection with spatially pooled features," in *Proceedings of European Conference on Computer Vision*, 2014, pp. 546–561.

[102] M. Mathias, R. Benenson, R. Timofte, and L. Gool, "Handling occlusions with franken-classifiers," in *Proceedings of IEEE International Conference on Computer Vision*, 2013, pp. 1505–1512.

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

[104] X. Chen, K. Kundu, Y. Zhu, A. G. Berneshawi, H. Ma, S. Fidler, and R. Urtasun, "3d object proposals for accurate object class detection," in *Proceedings of Advances in Neural Information Processing Systems*, 2015, pp. 424–432.

[105] J. Liu, Q. Fan, S. Pankanti, and D. N. Metaxas, "People detection in crowded scenes by context-driven label propagation," in *Proceedings of IEEE Winter Conference on Applications of Computer Vision*, 2016, pp. 1–9.

[106] Q. Wang, P. Yan, Y. Yuan, and X. Li, "Multi-spectral saliency detection," *Pattern Recognition Lett.*, vol. 34, no. 1, pp. 34–41, 2013.

[107] A. Torabi, G. Massé, and G.-A. Bilodeau, "An iterative integrated framework for thermal–visible image registration, sensor fusion, and people tracking for video surveillance applications," *Computer Vision and Image Understanding*, vol. 116, no. 2, pp. 210–221, 2012.

[108] M. S. Sarfraz and R. Stiefelhagen, "Deep perceptual mapping for thermal to visible face recognition," *Proceedings of British Machine Vision Conference*, 2015.

[109] J. Han and B. Bhanu, "Human activity recognition in thermal infrared imagery," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2005, pp. 17–17.

[110] ——, "Fusion of color and infrared video for moving human detection," *Pattern Recognition*, vol. 40, no. 6, pp. 1771–1784, 2007.

[111] Y. Socarrás, S. Ramos, D. Vázquez, A. M. López, and T. Gevers, "Adapting pedestrian detection from synthetic to far infrared images," in *Proceedings of International Conference on Computer Vision Workshop: Visual Domain Adaptation and Dataset Bias*, vol. 7, 2011.

[112] T. Gandhi and M. M. Trivedi, "Pedestrian protection systems: Issues, survey, and challenges," *IEEE Transactions on Intelligent Transportation Systems*, vol. 8, no. 3, pp. 413–430, 2007.

[113] I. Goodfellow, Y. Bengio, and C. Aaron, "Deep learning," 2016, book in preparation for MIT Press. [Online]. Available: http://www.deeplearningbook.org

[114] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 580–587.

[115] T. Xu, H. Zhang, X. Huang, S. Zhang, and D. N. Metaxas, "Multimodal deep learning for cervical dysplasia diagnosis," in *Proceedings of International Conference on Medical Image Computing and Computer Assisted Intervention*. Springer, 2016, pp. 115–123.

[116] R. Girshick, "Fast R-CNN," in *Proceedings of IEEE International Conference on Computer Vision*, 2015, pp. 1440–1448.

[117] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results," http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

[118] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *International Journal of Computer Vision*, vol. 104, no. 2, pp. 154–171, 2013.

[119] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *Proceedings of European Conference on Computer Vision*. Springer, 2014, pp. 391–405.

[120] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *Proceedings of European Conference on Computer Vision*. Springer, 2014, pp. 346–361.

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

[122] P. Dollár, "Piotr's Computer Vision Matlab Toolbox (PMT)," https://github.com/pdollar/toolbox.

[123] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proceedings of European Conference on Computer Vision*. Springer, 2014, pp. 818–833.

[124] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of International Conference on Machine Learning*, 2010, pp. 807–814.

[125] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.

[126] 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 IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[127] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1701–1708.

[128] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of IEEE International Conference on Computer Vision*, 2015, pp. 1026–1034.

[129] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," *arXiv preprint arXiv:1511.07289*, 2015.

[130] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[131] F. Yang, W. Choi, and Y. Lin, "Exploit all the layers: Fast and accurate cnn object detector with scale dependent pooling and cascaded rejection classifiers," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2129–2137.

[132] J. Li, X. Liang, Y. Wei, T. Xu, J. Feng, and S. Yan, "Perceptual generative adversarial networks for small object detection," *arXiv preprint arXiv:1706.05274*, 2017.

[133] W. Choi, Y.-W. Chao, C. Pantofaru, and S. Savarese, "Discovering groups of people in images," in *Proceedings of European Conference on Computer Vision*, 2014, pp. 417–433.

[134] X. Zhu and Z. Ghahramani, "Learning from labeled and unlabeled data with label propagation," Technical Report CMU-CALD-02-107, Carnegie Mellon University, Tech. Rep., 2002.

[135] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of International Conference on Machine Learning*, 2001, pp. 282–289.

[136] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," in *Proceedings of Advances in Neural Information Processing Systems*, 2011, pp. 109–117.

[137] D. Hoiem, A. A. Efros, and M. Hebert, "Putting objects in perspective," *International Journal of Computer Vision*, vol. 80, no. 1, pp. 3–15, 2008.

[138] S. Maji, L. Bourdev, and J. Malik, "Action recognition from a distributed representation of pose and appearance," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 3177–3184.

[139] R. Rothe, M. Guillaumin, and L. van Gool, "Non-maximum suppression for object detection by passing messages between windows," in *Proceedings of Asian Conference on Computer Vision*, 2014.

[140] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," *Proceedings of Advances in Neural Information Processing Systems*, vol. 16, no. 16, pp. 321–328, 2004.

[141] L. Xu and M. I. Jordan, "On convergence properties of the em algorithm for gaussian mixtures," *Neural Computation*, vol. 8, no. 1, pp. 129–151, 1996.

[142] B. Wang, Z. Tu, and J. K. Tsotsos, "Dynamic label propagation for semi-supervised multi-class multi-label classification," in *Proceedings of IEEE International Conference on Computer Vision*, 2013, pp. 425–432.

[143] A. Ess, B. , and L. Van Gool, "Depth and appearance for mobile scene analysis," in *Proceedings of IEEE International Conference on Computer Vision*, 2007, pp. 1–8.

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

[145] X. Zeng, W. Ouyang, and X. Wang, "Multi-stage contextual deep learning for pedestrian detection," in *Proceedings of IEEE International Conference on Computer Vision*, 2013, pp. 121–128.

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

[147] R. Benenson, M. Mathias, T. Tuytelaars, and L. Van Gool, "Seeking the strongest rigid detector," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3666–3673.

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

[149] M. Mathias, R. Benenson, R. Timofte, and L. Van Gool, "Handling occlusions with franken-classifiers," in *Proceedings of IEEE International Conference on Computer Vision*, 2013, pp. 1505–1512.

[150] S. Paisitkriangkrai, C. Shen, and A. van den Hengel, "Strengthening the effectiveness of pedestrian detection with spatially pooled features," in *Proceedings of European Conference on Computer Vision*, 2014, pp. 546–561.

[151] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," *arXiv preprint arXiv:1612.08242*, 2016.

[152] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Proceedings of European Conference on Computer Vision*. Springer, 2016, pp. 21–37.

[153] Y. Choi, N. Kim, S. Hwang, and I. S. Kweon, "Thermal image enhancement using convolutional neural network," in *Proceedings of IEEE International Conference on Intelligent Robots and Systems*. IEEE, 2016, pp. 223–230.

[154] J. Ren, X. Chen, J. Liu, W. Sun, J. Pang, Q. Yan, Y.-W. Tai, and L. Xu, "Accurate single stage detector using recurrent rolling convolution," *arXiv preprint arXiv:1704.05776*, 2017.

[155] X. Du, M. El-Khamy, J. Lee, and L. Davis, "Fused dnn: A deep neural network fusion approach to fast and robust pedestrian detection," in *Proceedings of IEEE Winter Conference on Applications of Computer Vision*. IEEE, 2017, pp. 953–961.

[156] J. C. Caicedo and S. Lazebnik, "Active object localization with deep reinforcement learning," in *Proceedings of IEEE International Conference on Computer Vision*, 2015, pp. 2488–2496.

[157] M. Bellver, X. Giró-i Nieto, F. Marqués, and J. Torres, "Hierarchical object detection with deep reinforcement learning," *arXiv preprint arXiv:1611.03718*, 2016.