Czech Technical University in Prague
Faculty of Electrical Engineering

Department of Cybernetics

DIPLOMA THESIS ASSIGNMENT

Student: Bc. Hana Š a r b o r t o v á

Study programme: Open Informatics

Specialisation: Computer Vision and Image Processing

Title of Diploma Thesis: Individual Person Counting in Semi-Crowded Environment

Guidelines:

State of the art video surveillance trackers fail to correctly operate when presented with crowded or semi-crowded scene at the input. The main problem is inability to determine whether the tracked object consists of a single or multiple individuals. The number of persons in each tracked object of a single camera video sequence is needed in order to solve multi-camera correspondence problem.

1. Review methods suitable for estimating the number of persons in images/video sequences of a semi-crowded scene with frequently occurring occlusions.
2. Select a method suitable for efficient person counting in parts of images/video sequences delimited by the output of a legacy video surveillance tracker.
3. Implement the selected method in Matlab or C/C++.
4. Evaluate the performance of the implemented method using available data (CAVIAR [1], PETS 2009 [2]).

Bibliography/Sources:

[1] CAVIAR: Context Aware Vision using Image-based Active Recognition URL: http://homepages.inf.ed.ac.uk/rbf/CAVIAR/
[2] PETS 2009 Benchmark Data, URL: http://cs.binghamton.edu/~mrldata/pets2009.html
[3] Corvee, E., Bremond, F.: Body Parts Detection for People Tracking Using Trees of Histogram of Oriented Gradient Descriptors. Proceedings of the 2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance (2010)
[4] Jia, HX., Zhang, YJ.: Fast Human Detection by Boosting Histograms of Oriented Gradients. Proceedings of the Fourth International Conference on Image and Graphics (2007)

Diploma Thesis Supervisor: Ing. Vít Líbal, Ph.D.

Valid until: the end of the winter semester of academic year 2015/2016

L.S.

doc. Dr. Ing. Jan Kybic
Head of Department

prof. Ing. Pavel Ripka, CSc.
Dean

Prague, January 27, 2015

České vysoké učení technické v Praze
Fakulta elektrotechnická

Katedra kybernetiky

ZADÁNÍ DIPLOMOVÉ PRÁCE

Student: Bc. Hana Š a r b o r t o v á

Studijní program: Otevřená informatika (magisterský)

Obor: Počítačové vidění a digitální obraz

Název tématu: Počítání individuálních osob v částečně zaplněné scéně

Pokyny pro vypracování:

Algoritmy sledování pohybujících se objektů v současných kamerových systémech selhávají pokud ve sledovaném prostoru dochází k častým zákrytům sledovaných objektů, jako například při sledování osob v davu, nebo v hustěji obsazené scéně. Jednou z hlavních příčin je problém rozlišit jednotlivý objekt od skupiny pohybujících se objektů a odhadnout počet objektů v rámci jedné detekce. Spolehlivý odhad počtu objektů v pohledu z jedné kamery je nezbytný k určení správných korespondencí mezi objekty různých kamer.

1. Proveďte rešerši metod vhodných pro odhad počtu osob v obrazech/video sekvencích hustě a středně hustě zaplněných prostor s častými překryvy.
2. Zvolte vhodnou metodu pro efektivní počítání osob v částech obrazu vymezených sledovacím algoritmem existujícího kamerového systému.
3. Implementujte tuto zvolenou metodu v Matlabu nebo C/C++.
4. Vyhodnoťte chování implementované metody na vhodných datech (CAVIAR [1], PETS 2009 [2]).

Seznam odborné literatury:

[1] CAVIAR: Context Aware Vision using Image-based Active Recognition URL: http://homepages.inf.ed.ac.uk/rbf/CAVIAR/
[2] PETS 2009 Benchmark Data, URL: http://cs.binghamton.edu/~mrldata/pets2009.html
[3] Corvee, E., Bremond, F.: Body Parts Detection for People Tracking Using Trees of Histogram of Oriented Gradient Descriptors. Proceedings of the 2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance (2010)
[4] Jia, HX., Zhang, YJ.: Fast Human Detection by Boosting Histograms of Oriented Gradients. Proceedings of the Fourth International Conference on Image and Graphics (2007)

Vedoucí diplomové práce: Ing. Vít Líbal, Ph.D.

Platnost zadání: do konce zimního semestru 2015/2016

L.S.

doc. Dr. Ing. Jan Kybic
vedoucí katedry

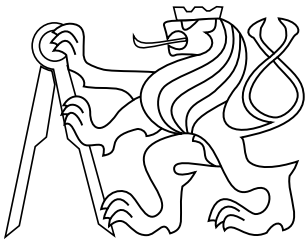
prof. Ing. Pavel Ripka, CSc.
děkan

V Praze dne 27. 1. 2015

master's thesis

Individual Person Counting in Semi-Crowded Environment

Hana Šarbortová



July 2015

Ing. Vít Líbal PhD.

Czech Technical University in Prague
Faculty of Electrical Engineering, Department of Cybernetics

Acknowledgement

I would like to express the greatest appreciation to my supervisor Vít Líbal for his patience, a magnificent guidance throughout this thesis and the amount of time he spent on consultations.

Declaration

Prohlašuji, že jsem předloženou práci vypracovala samostatně, a že jsem uvedla veškeré použité informační zdroje v souladu s Metodickým pokynem o dodržování etických principů při přípravě vysokoškolských závěrečných prací.

....
Date Signature

Abstract

Algoritmy sledování pohybujících se objektů v současných kamerových systémech selhávají, pokud ve sledovaném prostoru dochází k častým zákrytům sledovaných objektů, jako například při sledování osob v davu nebo v hustěji obsazené scéně. Hlavním problémem je neschopnost určit, zdali je ve sledovaném objektu jeden nebo několik jedinců. Cílem této práce je odhadnout počet lidí v každém z těchto sledovaných objektů. Takto odhadnutý počet lidí je nutný k řešení problému korespondence více kamer. Navržená metoda odhaduje počet lidí počítáním detekcí ve sledovaných objektech. Byly testovány dva typy detektorů, detektor celého těla a detektor hlavy a ramen. Všechny zkoumané detektory jsou složené kombinováním HOG nebo "channel features"s SVM nebo DF klasifikátory. Bylo natrénováno několik detektorů pro různé velikosti osob a pro každou část obrazu se odhadla očekávaná velikost jedince. Takto mohou být klasifikační znaky spočítány pouze jednou na celém obraze a na každou část obrazu se použije detektor, který velikostí nejvíce odpovídá očekávané velikosti detekce. Nejslibnější výsledky dává detektor kombinující "channel features"a DF klasifikátor.

Klíčová slova

Detekce osob, počítání osob, HOG, DPM, ICF, ACF, SVM, DF

Abstract

State of the art video surveillance trackers fail to correctly operate when presented with crowded or semi-crowded scene at the input. The main problem is inability to determine whether the tracked object consists of a single or multiple individuals. The objective of this thesis is to provide an estimation of people count for each of the tracked objects. The number of persons in each tracked object of a single camera video sequence is needed in order to solve multi-camera correspondence problem. The proposed method estimates the number of people by counting detections inside of the tracked objects. Two types of detectors, full body and head-and-shoulders, were tested. All investigated detectors are based on combining HOG or channel features with SVM or decision forest (DF) classifiers. Multiple detectors for different scales are trained and an expected person size is estimated for all image parts. Therefore, the features are computed only once at one scale and the detection is done by different detectors based on the estimated size. The most promising results are given by the head detector combining channel features with decision forest.

Keywords

Person detection, person counting, HOG, DPM, ICF, ACF, SVM, DF

Contents

Abbreviations

ACF	Aggregate Channel Features
DPM	Deformable Part Models
FP	False Positive
FPPI	False Positives per Image
GT	Ground Truth
HOG	Histogram of Oriented Gradients
ICF	Integral Channel Features
PR	Precision-Recall
LBP	Local Binary Patterns
ROC	Receiver operating characteristic
ROI	Region of interest
SVM	Support Vector Machine

1 Introduction

Analysis of people behavior in video surveillance is a very important task and information about the people count throughout the analyzed video streams makes the steps easier. The processing of surveillance videos involves several problems including person detection and tracking, crowd counting, behavioral analysis of individuals and groups, detection of suspicions events est. The approaches to solve these tasks highly depends on expected density of people in a scene. In general, the difficulty of any analysis increases with people density and frequent partial and full occlusions.

State of the art video surveillance trackers fail to correctly operate when presented with a crowded or semi-crowded scene at the input. Tracked objects often contain multiple individuals when the people density is high. The main problem is inability to determine whether they consist of multiple individuals or a single person even though the number of persons in each tracked object of a single camera video sequence is needed in order to solve multi-camera correspondence problem. The goal of this thesis is to determine how many individuals are in each tracked object.

1.1 Motivation

Analysis of surveillance videos is a very important task that have impact on several fields. One of the most important concern is security in public spaces such as airports, bus and train stations, shopping centers or streets but also in private sector including office houses or industrial areas. These places are often monitored by a huge number of surveillance cameras and their efficient analysis would be a very hard task without any help provided by computer vision algorithms. The analysis, as performed today, is a labour intensive task that is infeasible or near infeasible for a larger amount of cameras. The problem with such monitoring is that a camera system operator is very likely to stop paying attention while nothing is happening on screens for a long time. It is a known fact that a common attention span of a security operator is ten to twenty minutes while a single operator is capable of monitoring only a several camera views at a time depending on the scene and the scene traffic complexity. Therefore, a real-time monitoring of a large number of camera views to prevent or react to a security/safety incidents requires a rotating teams of trained security operators. Automated or semi-automated analysis using computers is a way to cope with the large number if cameras so that the security and safety demands are met. A computer aided analysis can have several purposes depending on the analyzed scene and security concerns. The methods differ depending on expected people density and the fact whether someone seen in the scene is an alarm or just an input to another analysis.

There are scenes where people are not supposed to be and detection of anyone should cause an alarm. This includes a theft detection or any protection of restricted areas. In this case, the number of people is not very important as the fact that there is at least one person should cause an alarm. The basic functionality of a system should be highlighting of unusual things in a scene so that the operator can react properly. In often

needed forensic analysis, a surveillance system operator needs to find a specific piece of information in possibly huge amount of video data which consequently requires a long search time or a large number of people involved. The computer aided automated or semi-automated video data analysis holds a great potential for improvement.

The other type of scenes, were people are expected to be, require a different philosophy when designing the analysis algorithms. The main concern is to provide a security measures and ensure public safety. However, properly functioning algorithms can provide useful inputs to many other seemingly unrelated fields. These include data for economic purposes, resource management, scheduling of public transportation, indexing multimedia archives, or advertising. The analysis have to deal with individual persons as well as crowds. The applications can vary from counting people at certain places to tracking groups and individuals throughout the entire monitored space that can be very complex. These data can be further used in more fine purposes such as behavioral analysis and detection of uncommon events.

The baseline to any analysis is a properly functioning detecting and tracking algorithm. Once the algorithms work for one camera, the analysis can expand to multi-camera problem which can cover entire complex monitored spaces. The complexity of such problems increases with the density of people. The increasing density introduces problems such as frequent partial or full occlusion. The processing of a crowded scenes becomes very difficult in surveillance applications as the data are often in low resolution and it has to be processed in real time.

1.2 Problem statement

The goal of this thesis is to provide an estimation of people count in a scene captured by a static overhead placed camera. The scene is being analyzed by a legacy tracker which is supposed to track persons within the captured scene. However, it fails to correctly operate when presented with a crowded or semi-crowded scene. In this work, the Active Alert video analytics suite of the Honeywell's Digital Video Manager is used in place of the legacy tracker. The active Alert is a state of the art commercial video analytics software providing a robust tracking of persons and objects in the camera views. While the Active Alert tracker works reliably across all kinds of environments, similar to other state of the art commercial solutions it does not combine the tracking information of overlapping cameras and works only when occlusions among multiple tracked objects do not occur very frequently. In practice, the detector considers any overlapping moving objects as one tracked object. The tracking works correctly when the scene contains only few unoccluded people. Nevertheless, if two tracked objects become even slightly occluded, they are merged into one. The task is to provide a number of people for all tracked objects.

The tracker objects were provided for a publicly available dataset, namely Pets2009 [1]. The given log files contain coordinates of a rectangular bounding box, ID, confidence, validity and other information. During the tracking process, a tracked object can be created and removed anytime and anywhere in the video. An ID is assigned to a newly detected object; no number is repeated in the given logs of a particular video sequence even if the detection with a particular ID does not exist any more. A new object can be created by a new detection or by splitting an existing object. In such a

case one newly created object keeps the previous ID and the other is assigned a new one. If two objects are merged into one, the lower ID is kept.

The estimation of people count would significantly improve informativeness of the tracker. It would provide more relevant information about what is happening in the scene; and more importantly, the people count estimation would help to solve multiple-camera correspondence problem. Such an information can allow more profound analysis of public spaces. It can lead to behavioral analysis of groups and individuals and significantly improve the efficiency of a surveillance operator. Also, searching in the past data would be more straightforward if the data were previously well analyzed.

The constraints on the proposed solution are not only accuracy but also the computational speed. The method should provide reasonable results preferably in real time. The problem is expected to be solvable by detecting individuals as the scene is not supposed to be highly crowded.

1.3 Solution overview

Current approaches to people count estimation are generally classified into three categories: model-based methods, trajectory-clustering-based methods, and map-based methods [2]. The model-based approaches attempt to segment and detect every single person in the scene. The trajectory-clustering-based approaches try to detect every independent motion by clustering interest-points on people tracked over time. In contrast, the map-based approaches count the number of people without having to segment or detect each individual. These approaches generally map the number of people to foreground pixels or some other features by training. The proposed method should work for a semi-crowded scene, which is understood as a scene that contains several people with frequent occlusions but where individual persons are still detectable. Therefore, this work follows the concept of model-based approaches, i.e. the count is given by counting detected individuals.

The objective is to keep the detector simple while obtaining sufficiently accurate results. Detection of persons in a scene with frequent occlusions is rather difficult as detectors are very sensitive to particular features computed on the whole body. The most important features seem to come from the parts like head and shoulders and the transition between body and legs. It is almost impossible to find a person with most of the parts hidden behind other person without any additional information. The most basic simplification is to make a size prediction. As the camera is placed overhead and is static, it is possible to make a rough estimation about expected body size for each pixel. This information can be taken from camera calibration, homography or by estimation from several body size examples coming from different parts of the image.

The detector can overcome the fact that most of the body parts are hidden by detecting only some of them. As the camera is placed overhead, heads are almost always visible. For this reason, the proposed approach focuses on head detection as it has the potential to give more accurate estimation for a situation with most of the other body parts hidden. However, a head-only detector would very likely give a lot of false positives. This fact is eliminated by detecting the head-and-shoulders part, sometimes referred as omega detector.

Several methods to detect head-and-shoulders are investigated. They are all implemented as sliding window detectors. The ideas were adopted from some of the most cited and efficient methods of people detection. Namely Histograms of Oriented Gradients (HOG) [3] and channel features [4] which are combined with Support Vector Machines (SVM) and decision forests trained by boosting classifiers.

In order to reduce the search space and the number of false positives, a strong assumption about expected head size is done. The space of all possible head sizes is quantized and several detectors for different sizes are trained. With this approach, all features can be computed only once for the regions of interest (ROI) framed by the tracker objects and no feature scaling is needed. Each window of a ROI is then classified by one detector from the detector pool with the most similar size.

An investigation over possible parameter settings is done for each of the used detectors. The performance of head detectors is compared with some people detectors, namely HOG [3], Deformable Part Models [5] and Aggregated Channel Features [6].

1.4 Thesis structure

The thesis is structured as follows: An overview of the related work is given in *Chapter 2*. As the problem of people detection is very popular and the number of publications is enormous, only the relevant papers are included. The used methods are described in *Chapter 3*. The chosen features and classifiers along with the size handling are introduced. The testing methods and results are shown in *Chapter 4*. At the end, the conclusion and proposals for future work are stated in *Chapter 5*.

2 Related work

People detection, counting and tracking under various conditions have been heavily explored in the last years. Mainly due to its potential value foreseen in the security related applications, the interest from both research community and commercial companies is very high.

The solutions are usually tailored to a constrained problem so that the performance and reliability can be guarantied. The specified constraints include expected people density and camera position. Environment constraints include restrictions on where a person can appear and what is the expected size as well as it gives a hint on what are the possible occlusion patterns.

2.1 People detection

The question of how to detect people in images has been thoroughly investigated in the last decades. However, the problems still remains challenging due to the high variability of detected objects and of the scene itself. People can appear in various clothing, hairstyles and body proportions which together with varying lighting and other scene conditions makes the detection very difficult. Most of the people detectors are restricted to upright standing persons, but the number of possible poses is still quite high.

People detection is well defined problem with well established benchmarks and evaluation metrics. The largely used datasets are Inria [3], Caltech [7] and Kitti [8]. All of them include upright people in various poses (standing and walking), however Inria dataset is more diverse and smaller. The Inria dataset includes people at various indoor and outdoor places (including ski area for example) and the camera height is not fixed. On the other hand, Caltech and Kitti datasets are taken by a camera placed on a moving car. All of these datasets are well annotated. Evaluation technique described in [9] is widely accepted for evaluation of the performance of the people detection algorithms.

Accuracy of the proposed methods increased over time, however the pay off is usually the computational time. Many of the best methods do not perform in real time and are therefore not suitable for many surveillance application. A variety of features and classification methods have been used for people detection. According to [10], there are three main solution families, namely DPM variants, Decision Forests and Deep Networks.

Histograms of Oriented Gradients (HOG) proposed by Dalal & Triggs in [3] is one of the most cited papers in the history of people detection. The method is based on evaluating normalized local histograms of image gradient orientations in a dense grid. The basic idea is that local object appearance and shape can often be characterized rather well by the distribution of local intensity gradients or edge directions, even without precise knowledge of the corresponding gradient or edge positions. In practice this

is implemented by dividing the image window into small spatial regions ("cells"), for each cell accumulating a local 1-D histogram of gradient directions. To achieve better invariance to illumination, shadowing, etc., a measure of local histogram "energy" is accumulated over larger spatial regions ("blocks") and used to normalize all of the cells in a particular block. All normalized histograms from an image window are then concatenated and classified by SVM. This method was initially trained and tested on Inria dataset that was established along with the method in [3].

A real-time human detection system based on Viola & Jones face detection framework [11] and HOG features is presented in [12]. They treat each bin of the histogram as an individual feature, i.e. each feature is defined by its owing block position, its cell position and the orientation bin. A feature can be evaluated in 8 look-ups using integral images. They use dimensionality reduction of the feature pool based on haar like features. A cascade of weak classifiers is trained so that negative detections can be discarded at early stages.

People detection algorithm based on a hierarchical tree of HOG was proposed in [13]. They seek to find the most dominant cells i.e. the areas best describing human features. A dimension reduction of the HOG features is performed by extracting the most dominant orientation from the histograms. Histograms over all dominant orientations of the corresponding cells are computed on the training images. The trained most dominant cell edge orientation is given by the histogram maximum probability of occurrence, which defines the top tree node. Sub-nodes are in turn initiated from each node using the maximum cell's dominant edge orientation unused in the parent nodes. To limit a high memory requirement, the tree is divided into smaller successive trees. The detection is coupled with independently trained body part detectors to enhance the detection performance.

Deformable Part Models (DPM) [5] enriched the Dalal & Triggs model [3] using a star-structured part-based model and become the state of the art in both person and object detection. The idea was to naturally deal with different poses and partial occlusion by allowing particular parts to move from its most probable position and then calculating with their level of displacement. In terms of implementation, the image is first searched by a root filter analogous to the HOG detector proposed in [3] but on a lower resolution. The contribution of particular parts is calculated using HOG features too, however on twice higher resolution. The score of a star model at a particular position and scale within an image is the score of the root filter at the given location plus the sum over parts of the maximum, over placements of that part, of the part filter score at its location minus a deformation cost measuring the deviation of the part from its ideal location relative to the root. Both root and part filter scores are defined by the dot product between a filter (a set of weights) and a sub-window of a feature pyramid computed from the input image. The discriminative training of classifiers makes use of latent information, in particular by using latent SVM which is well defined in [5] too. This is useful as only partially labeled data can be used for training (i.e. part positions are estimated during training). The concept of DPM was further explored in [14, 15, 16].

Channel Features are a baseline for many of the fastest and most efficient algorithms [6, 17, 18]. The idea of *Integral Channel Features* (ICF) was first formally described

in [4]. The major idea is that the most expensive computation of a sliding window detector should not be computed for each window separately but only once for the entire image instead. One channel is an image obtained by transforming the original image where each channel pixel is a feature describing the primary pixel while preserving the spatial information. The feature vectors for each detector window then can be obtained by simple look-ups to either original or integral channels. This can represent any color space, histogram of gradient orientation (one channel for each histogram bin) and gradient magnitude among others. They used Viola & Jones framework [11] to compute integral images and Haar wavelets over the channels.

The ICF framework starts a new family of detectors that are an extension of the old ideas from Viola & Jones [19]. Sums of rectangular regions are used as input to decision trees trained via Adaboost. Both the regions to pool from and the thresholds in the decision trees are selected during training. The crucial difference from the pioneer work [19] is that the sums are done over feature channels other than simple image luminance.

The authors further improved the idea of channel features in [20, 21, 6, 22]. In [20], features computed at one scale are used to approximate features at nearby scales, increasing detector speed with little loss in accuracy. Work [21] coupled cascade evaluation at nearby positions and scales to exploit correlations in detector responses at neighboring locations and further increased speed of the ICF detector. As single-scale square Haar wavelets have shown to be sufficient in the ICF framework, they propose *Aggregate Channel Features* (ACF) [6] where, instead of computing integral images and Haar wavelets, the channels are simply smoothed and downsampled. The features are single pixel lookups in the "aggregated" channels. *Locally Decorralated Channel Features* [22] show that filtering the channel features with appropriate data-derived filters can remove local correlations from the channels. Given decorrelated features, boosted decision trees generalize much better giving a nice boost in accuracy.

Informed Haar-like Features [17] enhanced the idea of channel features by a basic idea of incorporating common sense and everyday knowledge into the design of simple and computationally efficient features. Basically they came with not so revolutionary yet innovating idea of using the knowledge how an upright human body looks like. They therefore employed a statistical model of the up-right human body where the head, the upper body, and the lower body are treated as three distinct components. A pool of rectangular templates that are tailored to such a shape model was systematically designed. By incorporating different kinds of low-level measurements, the resulting multi-modal & multi-channel Haar-like features represent characteristic differences between parts of the human body yet are robust against variations in clothing or environmental settings. Their approach avoids exhaustive searches over all possible configurations of rectangle features and neither relies on random sampling.

Filtered Channel Features proposed in [18] were inspired by the ideas from the above described publications [17] and [6]. They observed that these top performing pedestrian detectors can be modeled by using an intermediate layer filtering low-level features in combination with a boosted decision forest. Based on this observation they proposed a unifying framework and experimentally explored different filter families. Sum-pooling can be re-written as convolution with a filter bank (one filter per rectangular shape) followed by reading a single value of the convolution's response map. This "filter + pick" view generalizes the integral channel features [4] detectors by allowing to use any

filter bank (instead of only rectangular shapes).

Most of the published methods are using features based on gradient orientation and magnitude, as it gives useful information about the body shape. Although the appearance of people is diverse, color has shown to be an effective feature and hence multiple color spaces have been explored (both hand-crafted and learned) [4, 23, 24]. The LUV color is also included as channels used in [6, 17, 18]. Other investigated features are for example LBP [25, 26, 27], local structure [28, 29, 30, 31, 32] or covariance [33, 27]. Even methods using deep networks use some gradient and color features [34, 35, 36, 37, 38] with exception of [39] which shows promising results using RGB data only.

2.2 Occlusion handling

State-of-the-art people detectors perform well in scenes with relatively few people, however more crowded environments with frequent partial occlusions remain to be problematic. There are different types of occlusion we can observe in a scene. People can occlude each other or can be partially hidden behind an object. Also, the occlusion patterns differ with camera position. Data captured from a person or car view offer much smaller variety of how two people can occlude each other as their heads will be always in a similar height and persons far behind cannot be visible at all. On the other hand, a typical surveillance camera view offers a much larger set of possible patterns but the occluded person is usually more visible.

A traditional approach to this problem is to focus entirely on the occluded object and treat the occluder as a noise. There have been attempts to tackle the occlusion problem by integrating detection with segmentation [40] and latent variables for predicting truncation [41, 25] resulting in improved recognition performance, all these attempts have been tailored to specific kinds of detection models, and not been widely adopted by the community.

The fact that typical occlusions are due to overlaps between people is investigated in [42], where authors proposed a people detector tailored to various occlusion levels. Instead of treating partial occlusions as distractions, they leverage the fact that person-person occlusions result in very characteristic appearance patterns that can help to improve detection results. They proposed a new double-person detector that allows to predict bounding boxes of two people even when they occlude each other by 50% or more, and described a new training method for this detector. Also, they propose a joint person detector, that is jointly trained to detect single- as well as two-people in the presence of occlusions. The detector builds on the DPM approach; the detector shares the deformable parts across two people which belong to the same (two-person) root filter. For training, they synthetically generated two-people samples by cropping unoccluded people and overlapping them in order to create new samples with different levels of occlusion.

The approach in [42] directly inspired a work presented in [43]. However, they do not use synthetic data, the well annotated data of Kitti dataset [8] are used instead. They use annotations in the form of 3D object bounding boxes and readily available projection matrices to define a joint feature space that represents both the relative layout of two objects taking part in an occlusion and the viewpoint from which this

arrangement is observed by the camera. Clustering on this joint feature space is then performed, resulting in an assignment of object pairs to clusters that are used as training data for components of mixture models. The introduced method is also based on DPM approach, although the training is done using the structured SVM formulation as done for the DPM in [44]. They compare models with and without joint root. The method is designed primary for detection of occluded car, however is tested as a people detector as well to show that it adapts well to a nonrigid objects.

Another work [45] inspired by [42] proposed a method of a joint-people detector. It is again based on DPM and uses the structured SVM formulation proposed in [44]. Synthetic data obtained in same way as in [42] are used for training. In [42], they focused on side-view occlusion patterns, but crowded street scenes exhibit a large variation of possible person-person occlusions caused by body articulation or their position and orientation relative to the camera. To address this, [45] explicitly integrate multi-view person-person occlusion patterns into a joint DPM detector. They introduce an explicit variable modeling the detection type, with the goal of enabling the joint detector to distinguish between a single person and a highly occluded pair.

2.3 Detection in crowds

Classic (full body) people detectors are not sufficient for analysis of a highly crowded scene as most of the body is not visible. For a person or car view, this problem can be overcame by improving robustness against occlusion as described in *2.2*, mainly because people far behind the front line are not visible at all. However, these "hidden" people can be seen by the top view cameras, although only a head or upper body is usually visible if the density is high. Head detection does not seem to be very robust as the head shapes, hairstyles and coverings can vary greatly. But addition of shoulders can introduce a very typical "omega" shape. Head and shoulders detector overcome the problem of the head appearance variability and can be even more reliable than a full body detector as it does not suffer from large pose variations.

There are many research works studying detection of head and shoulders. Wu *et al.* [46] applied the edgelet features. Li *et al.* [47] extended the approach of Dalal & Triggs [3] to head-shoulder detection by boosting local HOG features and showed good performance in crowded scenes. Zeng *et al.* [48] proposed a discriminative multilevel HOG-LBP feature and proved its superiority over the HOG feature. However, these methods are time-consuming when searching an entire image which limited their usage in practical applications.

Li *et al.* [49] largely sped up the head and shoulders detector previously proposed in [47]. A Viola & Jones type classifier and a local HOG feature based AdaBoost classifier are combined to detect head and shoulders rapidly and effectively. In [50], the speed has been improved by attention-based foreground segmentation method to extract regions of interest. A robust real-time multi-view detection cascade is used on the selected regions. In the first layer of the cascade, a linear classifier with very high detection rate and relatively high rate of false positives is used to eliminate the obvious non-head-shoulders windows rapidly while keeping true candidates as much as possible. Very few windows are left to the more precise multi-view models that can run in parallel in the second layer. Each view model is a classifier and a window is rejected only if all

view models classify it as negative.

In contrast to most of the detectors presented in *2.2*, all above head and shoulders detectors are trained on real data only. Yu *et al.* [51] proposed a new type of synthetic data for upper body detection. They created 3D human models, rendered them in various poses and then placed them over images with a real background. A combination of Haar-like features and HOG features was used to train weak classifiers. The influence of various choices of training data was tested. They show that the performance can be improved by using synthetic data, however synthetic data alone are significantly worse than real data only.

An approach combining head and shoulders detector with a full body detector was presented by Wang *et al.* [52]. They reformulated the score computation of body parts in the original DPM detector to enhance the head part of the deformable part-based model to make it more suitable to the crowded sequences and used the "online" learned dictionary to refine the detection responses. The experimental results on three benchmark sequences demonstrated the superiority and effectiveness of this approach in detecting pedestrians with occlusions handling in crowded scenes.

3 Methods

The stated problem, people counting in bounding boxes given by a tracker, can be solved in different ways. The baseline method is to do frame by frame people detection and estimation of number by counting detections within each box of the tracker. Once having good detections, the problem can be somewhat simplified by the fact that the given input is a video from a static camera. Therefore, the position of a body/head in the next frame can be predicted based on current position and its previous behavior. Given the fact that a good detection algorithm is needed as a baseline, this work covers a research over existing people and head detectors and discusses their suitability for detection in semi and highly crowded scenes.

The described procedure of detection is based on the idea of sliding window, therefore features have to be computed and extracted for each part of an image, then classified and finally thresholded do decide whether it is a person/head or not. To avoid representation of the same object by multiple detections, a non-maximal suppression is done in order to filter close similarly but less confident detections. The further explored methods are based on several published papers that have been highly influencing and seem to give the most promising results. With regard to *features*, the Histogram of Oriented Gradients (HOG) [3] and a related idea of Integral Channel features [4] are described in *3.1*. The mostly used methods of *classification*, such as Support Vector Machine (SVM) and Decision Forests (DF) trained by boosting, are discussed in *3.2*.

The final detection can take an advantage of several possible simplifications. First, there is no need to exhaustively search the entire image by a sliding window detector as regions of interests (ROI) are already given. The ROI are stated by the tracker bounding boxes. Regardless of how accurate they are, the information about people outside of these boxes is irrelevant to this problem. Second, all data is coming from a static camera that is at a particular height and observing the ground plane under a particular angle, this means that the size of a person can be predicted based on the position in the image plane. Therefore, not all scales of an input image have to be searched for detection at all positions. A closer description of *scaling* options is in *3.3*.

3.1 Features

The most popular approach for improving detection quality is to increase/diversify the features computed over the input image. By having richer and higher dimensional representations, the classification task becomes somewhat easier, enabling improved results [10]. In the last decade improved features have been a constant driver for detection quality improvement.

The choice of features can also significantly influence the classification speed. Considering a sliding window detector, the features can be divided into two families. The first group would represent features that can be computed only once for the entire image. Example representatives of this family are HOG [3] and channel [4] features.

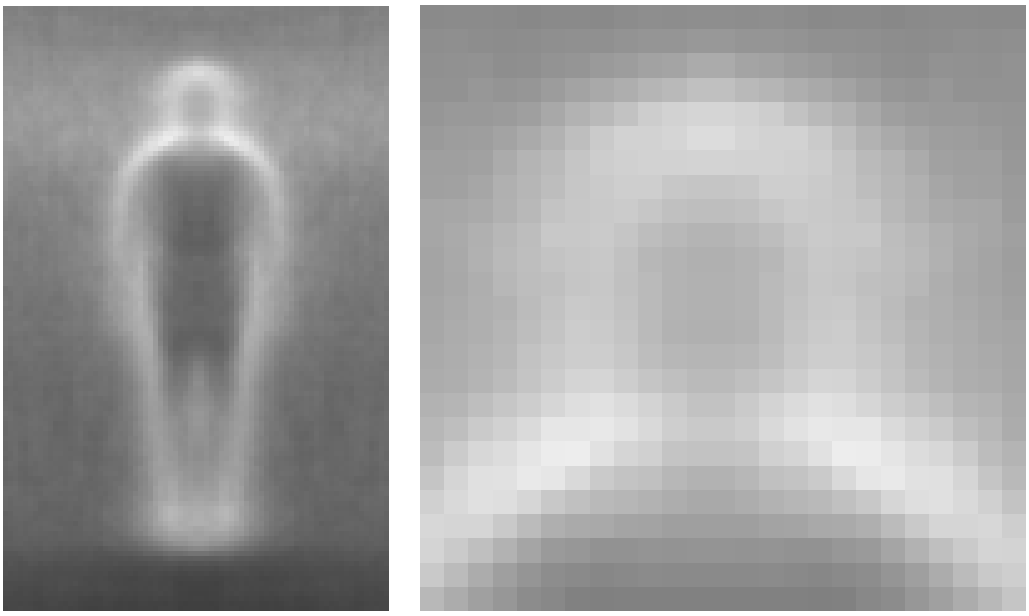


Figure 1 Average image of the gradient magnitude over all training samples of the INRIA dataset, full body and a cropped part of the head area

The other type includes such features that have to be computed for each window separately. This includes any features that are not invariant to translation. As one of the biggest concerns of the solution is speed, and the idea is to follow the trend of the most promising methods, the choice of features in this work is from the first family.

3.1.1 Histograms of Oriented Gradients (HOG)

One of the most influencing features are Histograms of Oriented Gradients (HOG) [3]. The proposed method become somewhat standard and implementations can be found in various computer vision libraries. For example OpenCV and Matlab Computer Vision Toolbox include pretrained HOG detectors and fast implementations of HOG features can be bound in VLFeat [53] and Piotr's Computer Vision Toolbox [54].

The method is based on the idea, that the most significant indicator is the human shape, and the best way to extract the shape is image gradient. An example of a gradient image can be seen in *Figure 2(b)*. The gradient expressiveness is outlined in *Figure 1* where average gradient images over all training examples from the INRIA dataset [3] are shown. It can be noted that one of the most stable body parts are shoulders.

Acquiring of HOG features involves several steps. The procedure details can vary from implementation to implementation, however the further described settings are used for testing in this work. First, a gradient magnitude and orientation images are obtained. The gradient for each pixel is computed using central difference, separately for each color channel and the one with maximum magnitude is used. HOG divides the input image into square cells and for each cell a histogram of gradients is computed. Each gradient is quantized by its angle and weighed by its magnitude and bilinear interpolation is used to place each gradient in the appropriate orientation bin.

Gradient strengths vary over a wide range owing to local variations in illumination

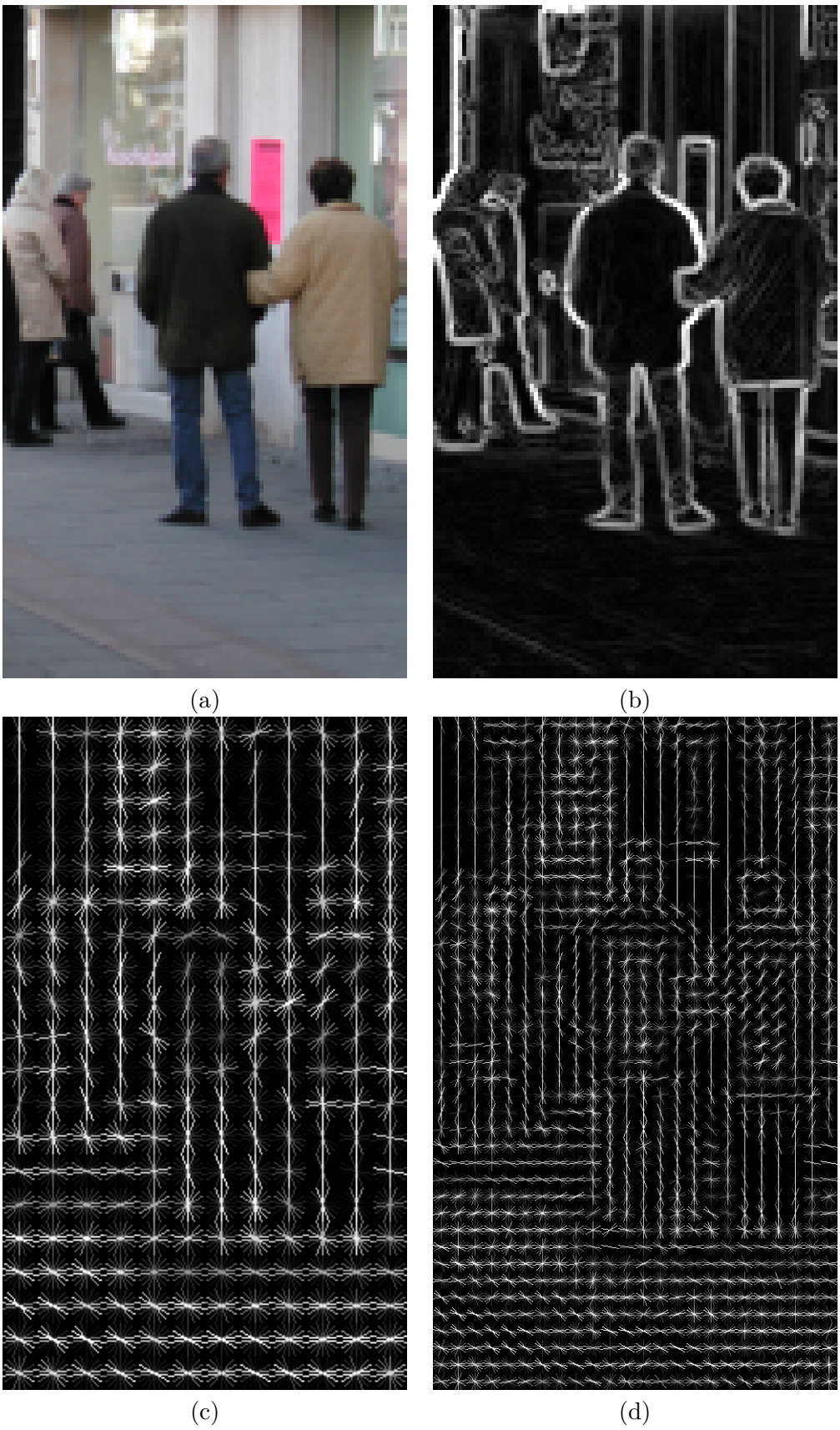
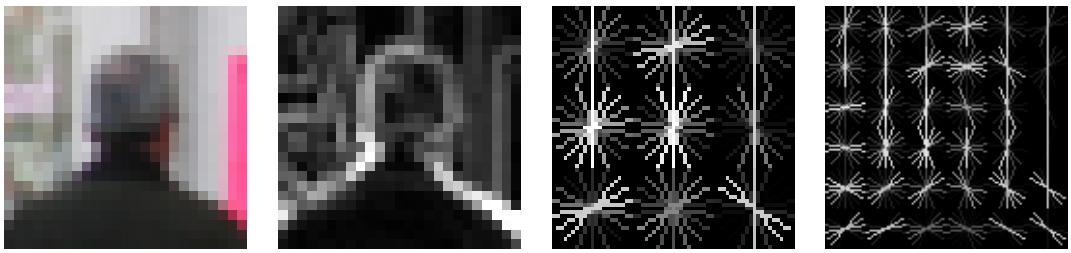


Figure 2 An example image from the INRIA dataset (a) and corresponding gradient magnitude (b) and HOG features with cell size 8 and 4 in (c) and (d) respectively



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

Figure 3 An example crop of the head-and-shoulders body from an example image of the INRIA dataset (a) and corresponding gradient magnitude (b) and HOG features with cell size 8 and 4 in (c) and (d) respectively

and foreground-background contrast, so effective local contrast normalization is essential for good performance [3]. For each resulting histogram of a cell, four different normalizations are computed using adjacent histograms. Cells are grouped into partially overlapping spatial blocks of four cells and L2 norm is computed for each block. Given a HOG cell, four normalization factors are obtained as the inverse of the norm of the four blocks that contain the cell. The originally used cell size in [3] was 8×8 with corresponding 16×16 blocks, the histograms of oriented gradients had 9 bins.

The implementation used in this work is from VLFeat open source Computer Vision library. There are two supported variants, the original from Dalal & Triggs [3] described above and UoCTTI [5]. The main difference is that the UoCTTI variant computes both directed and undirected gradients as well as a four dimensional texture-energy feature, but projects the result down to 31 dimensions (considering 9 orientation bins). Dalal & Triggs works instead with undirected gradients only and does not do any compression, for a total of 36 dimension. The dimension of a feature vector of each cell is given as four times the number of orientation bins, which is coming from the fact that each cell histogram has 4 distinct normalizations. Given an image of size $m \times n$, number of bins b and cell size c, the transformation to a HOG space results in a matrix of size $\lfloor m/c \rfloor \times \lfloor n/c \rfloor \times b * 4$. Examples of illustrative HOG feature visualizations can be seen in *Figure 2* and *3*.

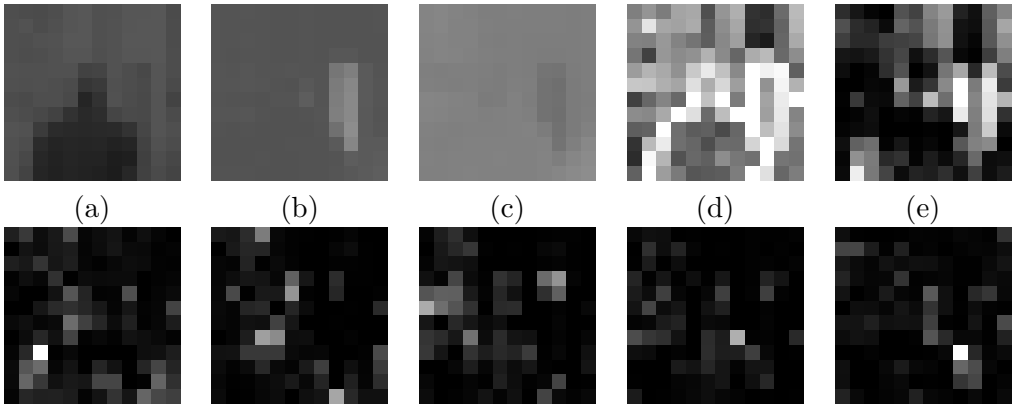
3.1.2 Channel Features

Channel features are very effective in sliding window object detection, both in terms of accuracy and speed. Channels are rather concept than a specific feature definition. Numerous feature types including histogram of oriented gradients (HOG) can be converted into channel features.

Given an input image I, a corresponding channel is a registered map of the original image, where the output pixels are computed from corresponding patches of input pixels (thus preserving overall image layout) [4]. A trivial channel is simply the input gray-scale image, likewise for a color image each color channel can serve as a channel. Other channels can be computed using linear or non-linear transformations of I, the only constraint is that channels must be translationally invariant. This allows for fast object detection, as the channels can be computed once on the entire image rather than separately for each overlapping detection window.

(a) (b)

Figure 4 Visualisation of the LUV color space on an example image from the INRIA dataset



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

(f) (g) (h) (i) (j)

Figure 5 10 channel features computed on the image from *Figure 3(a)*; 3 channels of LUV color (a-c), gradient magnitude(d) and gradient orientation (e-j). It was computed using Piotr's toolbox [54] with shrink parameter 4.

Usually, three channel types are used: color, gradient magnitude and gradient orientations. As color channels, LUV color space (see *Figure 4*)is typically used. Gradient orientations represent a histogram of oriented gradients where each histogram bin represents one channel Q_θ. In other words the channels are given by

$$Q_\theta(x, y) = G(x, y) \cdot \mathbf{1}[\Theta(x, y) = \theta], \tag{1}$$

where $G(x, y)$ and $\Theta(x, y)$ are the gradient magnitude and quantized gradient angle, respectively, at $I(x, y)$.

The channels itself are usually not used directly as features, however exceptions exist [6]. First order features (sums over rectangular parts of channels) and second order features (difference of first order features, such as Haar) are used instead. While various decision forest methods use 10 feature channels [4, 6, 55, 17, 18], some papers have considered up to an order of magnitude more channels [56, 57, 58, 59, 60]. Despite the improvements by adding many channels, top performance is still reached with only 10 channels (6 gradient orientations, 1 gradient magnitude, and 3 LUV color channels) [10]. An example of these 10 channels can be seen in *Figure 5*. The shown channels, as well as any further used channels, were computed by using Piotr's toolbox [54].

3.2 Classification

The choice of classification method, along with the choice of features, significantly influence detection performance in both accuracy and speed. For some applications, not only the classification speed but also the training speed plays an important role when an online learning is needed. Some algorithms tend to be more favorite than others, Support Vector Machines (SVM) and decision forests are among the most used ones. Also, the deep networks show interesting properties and fast progress.

Since the original proposal of HOG+SVM by Dalal & Triggs [3], linear and non-linear kernels have been considered. There is no conclusive empirical evidence indicating that whether non-linear kernels provide meaningful gains over linear kernels (for pedestrian detection, when using non-trivial features) [10]. Boosted decision trees seem particularly suited for pedestrian detection, reaching top performance [10]. In the work [56], it was argued that, given enough features, Adaboost and linear SVM perform roughly the same for pedestrian detection.

3.2.1 Support Vector Machines (SVM)

Support Vector Machines (SVM) are very important machine learning tool. The task is to find the best hyperplane that separates positive and negative samples. It maximizes the margin between the classes and finds a solution for data that are not entirely linearly separable. Also, by using various kernels, the separation does not have to be linear.

The classifier itself is a hyperplane defined by weight vector w and bias b, so that the classification is done by evaluating

$$y_i = wx_i + b \tag{2}$$

where $sign(y_i)$ is the predicted label for the i^{th} sample represented by a feature vector

x. Estimation of w and b requires to solve a quadratic optimization problem

$$\min_{\mathbf{w}\in\mathbb{R}^d} \left[\frac{\lambda}{2} \|\mathbf{w}\|^2 + \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(\mathbf{x}^\intercal \mathbf{w}_i) \right] \tag{3}$$

where $\mathcal{L}$ is a loss function for i^{th} datapoint. A hinge function is used as a loss function in this work. For an intended output $t = \pm 1$ and a classifier score y_i, the hinge loss of the prediction y_i is defined as

$$\mathcal{L}(y_i) = \max(0, 1 - t_i y_i) \tag{4}$$

Note that this does not include the bias term. The bias is incorporated by extending each data point $\mathbf{x}$ with a feature of constant value b_0, such that the objective becomes

$$\min_{\mathbf{w}\in\mathbb{R}^d} \left[\frac{\lambda}{2} \left(\|\mathbf{w}\|^2 + \left(\frac{b}{b_0} \right)^2 \right) + \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(\mathbf{x}^\intercal \mathbf{w}_i) \right] \tag{5}$$

The SVM solver used in this work is Stochastic Dual Coordinate Ascent proposed in [61] implemented in VLFeat library [53].

An important property of support vector machines is that the determination of the model parameters corresponds to a convex optimization problem, and so any local solution is also a global optimum. A great advantage during the testing is that features with preserved image coordinate system can be classified by simply convolving the vector w with the image/channel data.

3.2.2 Boosting

Boosting is one of the most popular learning techniques in use today. The idea is to combine many weak learners to form a single strong one. Boosting can give good results even if the weak classifiers have a performance that is only slightly better than random. Shallow decision trees are commonly used as weak learners due to their simplicity and robustness in practice. This powerful combination of Boosting and decision trees is the learning backbone behind many state-of-the-art methods across a variety of domains.

A boosted classifier (or regressor) has the form

$$\mathbf{H}(\mathbf{x}) = \sum_t \alpha_t \mathbf{h}_t(\mathbf{x}). \tag{6}$$

It can be trained by greedily minimizing a loss function $\mathcal{L}$; i.e. by optimizing scalar α_t and weak learner $\mathbf{h}_t(\mathbf{x})$ at each iteration t. Before training begins, each data sample $\mathbf{x}_i$ is assigned a non-negative weight w_i (which is derived from $\mathcal{L}$). After each iteration, misclassified samples are given greater weight when used to train the next classifier in the sequence. Each iteration requires training a new weak learner given the sample weights.

A decision tree $h_{TREE}(\mathbf{x})$ is composed of a stump $h_j(\mathbf{x})$ at every non-leaf node j. Trees are commonly grown using a greedy procedure, recursively setting one stump at a time, starting at the root and working through to the lower nodes. Each stump produces a binary decision; it is given an input $\mathbf{x} \in R^K$, and is parametrized with a

polarity $p \in \{\pm 1\}$, a threshold $\tau \in R$ and a feature index $k \in \{1, 2, ..., K\}$

$$h_j(\mathbf{x}) \equiv p_j sign(\mathbf{x}[k_j] - \tau_j), \tag{7}$$

where $\mathbf{x}[k]$ indicates the k^{th} feature/dimension of $\mathbf{x}$.

At each stage of stump training, the goal is to find the optimal parameters that minimize the weighted classification error ε

$$\varepsilon \equiv \frac{1}{Z} \sum w_i \mathbf{1}[h(\mathbf{x}_i) \neq y_i], \quad Z \equiv \sum w_i \tag{8}$$

where y_i is a sample label ($y_i \in \{1, -1\}$). For binary stump the error can be rewritten as

$$\varepsilon = \frac{1}{Z} \left[\sum_{\mathbf{x}_i[k] \leq \tau} w_i \mathbf{1}[y_i = +p] + \sum_{\mathbf{x}_i[k] > \tau} w_i \mathbf{1}[y_i = -p] \right]. \tag{9}$$

This error is minimized by selecting the single best feature k^* from all of the features.

$$\{p^*, k^*, \tau^*\} = \arg\min_{p,k,\tau} \varepsilon^{(k)}, \quad \varepsilon^* \equiv \varepsilon^{(k^*)} \tag{10}$$

Theoretically, other split criteria such as information gain, Gini impurity or variance can be used as well. However, the implementation described in [62] using classification error is used in this work. Boosting is very fast in testing but suffers from long training times. The method proposed in [62] focuses on speeding up the training time while the boosted tree offers identical performance to one with classical training.

The advantage of Boosting decision forest is that the number of weak classifiers tree depth can be set. Therefore, only the most reliable features from a potentially large pool can be chosen. By nature of the problem, it is very robust to overfitting. Also, the decrease of computational time during classification can be achieved by cascade evaluation of weak classifiers which discard negative samples at early stages.

3.3 Scaling

Images of the real world taken by a camera obey the rules of perspective projection. This means that people to be detected can appear in an image at various sizes depending on their position in the real world. Not only the size but also body proportions and angle with the image axes change due to the perspective.

The approach to scaling highly depends on the information about scene that the detector can rely on. The traditional approach is to scale an input image and to use the same detector over each scale and then choose the most reliable detections over all scales. Either the image itself can fe scaled and features are computed independently at each scale or the feature representation can be scaled directly. Multi-resolution image features can be approximated via extrapolation from nearby scales such as in [5]. This significantly reduces processing time and allow efficiently use multi-scaling in real time.

Figure 6 Illustration of the influence of the camera position on the predictability of position and size in various parts of the image. Heads are on the horizon for a car/person view and the size can vary (left). Size can be predicted for each part of the image when overhead camera view is used (right).

The problem of scaling can be simplified for a static overhead camera. An approximate scale for each position can be predicted if some additional information about the projection is known. The position and size of each individual strongly dependent on the camera position. As can be seen in *Figure 6*, if a camera captures a scene from the person/car view, the position of a head will be always very close to the horizon. On the other hand, if the camera is placed overhead, the head can appear anywhere but the size will differ.

There are several ways how to determine expected size for each position in an image. First, the projection can be computed from parameters of a calibrated camera. Another approach would be to estimate size from a known ground plane homography. The size prediction from homography is not accurate as it is a mapping between two planes, however it can provide a simple approximation. If neither camera calibration nor homography is known, the size map can be estimated from a sufficient amount of examples and constraints for perspective projection.

If a size map is available, the scaling approach can be reverted, i.e. the image does not have to be resized but will be detected by different detectors depending on the position in the image. This reduces the time needed for classification, although the pay off is the training time. However, the training time is not a big concern in this case.

The way how HOG and channel features are implemented are highly suitable for this method. As the idea is to compute features only once, the same cell size and number of orientation have to be used for each of the detectors. Therefore, during the training part, the detector size will start at the minimum expected person size and will increase by one cell size step until it reaches the maximum size. In order to preserve the window side length ratio, a head detector can be implemented as a square, i.e. the difference between scales will be always one cell on each side. The full body detector can preserve height twice longer than width. Therefore every other scale of a full body detector will have height two cells and width one cell shorter.

The size is not the only concern when detecting people in videos from surveillance cameras. The problem comes with a wide view camera. As can be seen on an illustra-

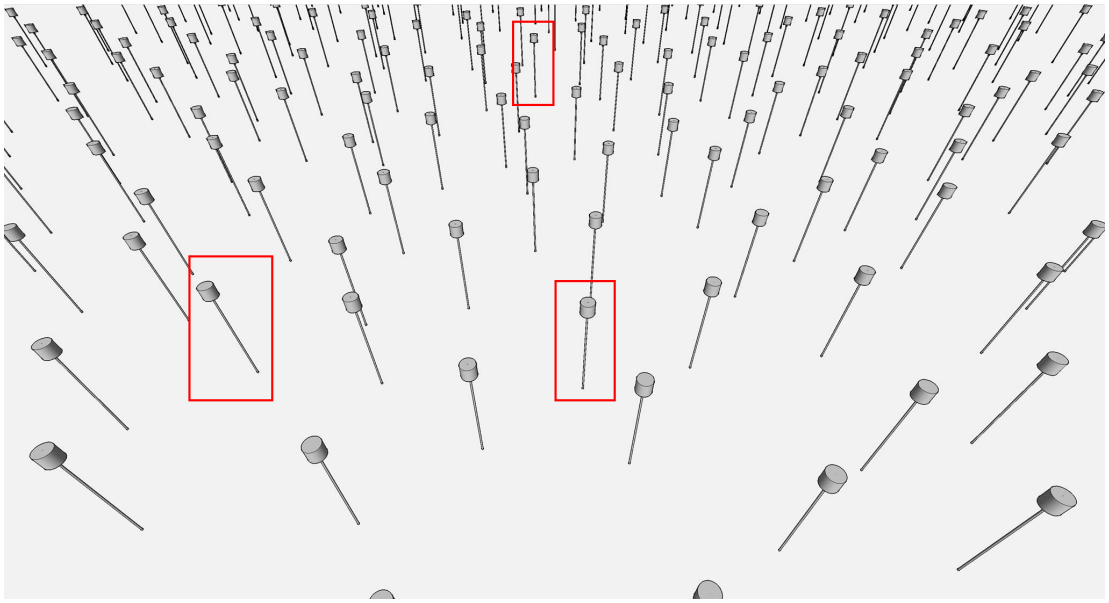


Figure 7 An abstract illustration of the human body appearance changes due to perspective projection in images captured by a camera with large field of view (a Google SketchUp illustration rendered by a camera with field of view 60 degrees)

tion in the *Figure 7*, people remain upright only in the central part of the image. On sides of a wide view, a common people detector usually fails due to the perspective. As this is not the main issue in the publicly available datasets used in this work, there is no proposed solution for this problem. However, it should be considered in a real application.

4 Experiments

The overall evaluation is based on comparison of several methods, their combination and parameter settings. The results are tested on two publicly available datasets which are described in *Section 4.1.* Different evaluation techniques are used based on available annotations, as either each individual was labeled or tracker bounding boxes were assigned a number of people visible inside of them. The evaluation metrics are described in *Section 4.3.*

The tested methods are designed to show whether counting by detection is a sufficient tool for a semi-crowded scene. To do so, two types of detectors are used, full-body and head-and-shoulders, which are for simplicity further referred as *person* and *head* detectors respectively. The detection methods are based on combinations of features and classifiers described in *Chapter 3.* Particularly, these detectors are used:

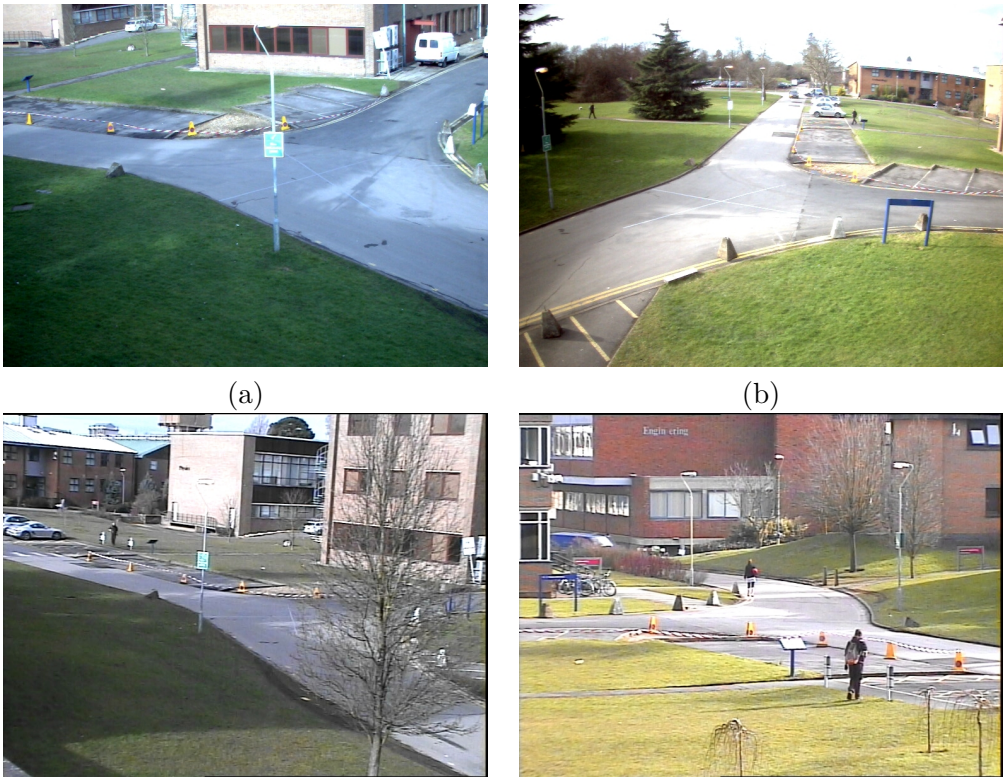
- **ACF** (Aggregated Channel Features proposed from [6])
- **DPM** (Deformable Part Models from [5])
- **HOG-SVM** (Histograms of Oriented Gradients from [3])
- **HOG-DF** (HOG features and decision forest trained by boosting)
- **Chns-SVM** (Channel features and SVM)
- **Chns-DF** (Channel features and decision forest trained by boosting)

The detectors were assigned one particular color for better orientation that is used for both graph comparisons and bounding boxes (the only exceptions are plots where results of only one detector are shown or there is a comparison of the same feature-classifier combination for head and person detector), the colors are for future reference listed in *Table 1.*

Different combination of C/C++ and Matlab code is used for each of the detectors. Implementation directly provided by authors is employed for ACF and DPM; HOG and SVM are used from VLFeat Computer Vision toolbox [53]; and channel features and boosting algorithm were from the Piotr's toolbox [54]. These methods were then combined in a Matlab code. Unfortunately, the combination of different implementations makes it hard to reliably measure the computational time.

detector type	color
ACF	yellow
DPM	cyan
HOG-SVM	magenta
Chns-SVM	red
Chns-DF	green
HOG-DF	blue

Table 1 Colors assigned to each detector type used in experiments. This color is consistent for all graphs and detector's bounding boxes in this Chapter.



(a) (b)

(c) (d)

Figure 8 Example images of the views from the Pets2009 [1] dataset. Views 1-4 are shown in (a-d) respectively.

(a) (b)

Figure 9 Example images of the views from the Caviar dataset [63]; corridor and front view in (a) and (b) respectively.

4.1 Datasets

All methods are tested on two datasets, Pets2009 [1] and Caviar [63] which have several views that capture the scene as shown in *Figure 8* and *9*. Both datasets contain data with high number of occlusions, mainly people partially occluded by other people. The Caviar dataset contain less crowded scenes focusing on occlusion of individual people. The crowd density of Pets2009 depends on a particular view.

The people in each view appear in various sizes, see *Table 2*. Also, the size of a frame is

Dataset and view	minimum person size	maximum person size
Pets2009 view 1	50	140
Pets2009 view 2	20	110
Pets2009 view 3	50	90
Pets2009 view 4	50	100
Caviar corridor	40	150
Caviar front	30	100

Table 2 The range of person sizes in which most of the detectable people from the particular dataset views appear

different for each dataset. The frame size is 384×288 pixels for Caviar and 768×576 for Pets2009. The Pets2009 includes several (9) videos that offer investigation of different level of occlusion, the length of each video varies roughly between 200 and 600 frames. Caviar dataset has 26 videos, mostly of 2 to 3 people passing each other. Only few (4) videos are suitable for testing of a semi-crowded scene.

The used type of annotations differs for each dataset and view. Pets2009 view 1 offers the most clearly visible individual persons within the whole frame and therefore each person could have been annotated, the annotations were taken from [64]. The other three views contain highly occluded persons in smaller sizes which would be difficult to annotate. For these three views, only bounding boxes of the tracker were annotated with number of people that are visible inside of them. The count is based on the number of visible persons not the real number of individuals that is in the 3D space framed by a particular bounding box in 2D. The Caviar dataset comes with its own annotations that include individual person bounding boxes.

4.2 Overview of detectors

Two types of detectors are compared in this work, head-and-shoulders (*head*) and full body (*person*). Both of them have some advantages and disadvantages over the other. Generally the person detectors do not have the capability to detect people under high occlusion and head detectors are sensitive to lack of contrast and produce more false positives.

As a baseline, the HOG detector [3] was tried. It is combination of HOG and SVM and implementations are available in OpenCV and Matlab. An example of HOG detection on full image is shown in *Figure 10*. This detector is advertised as a detector of unoccluded upright standing people and the fact that its performance significantly decreases when presented with occluded people was confirmed. On the other hand, alone standing individuals are detected reasonably well. A big concern of using the HOG detector is quite high number of false positives. The problem with the available implementations is that it has minimal person size either 128×64 or 96×48 pixels. The detected image is only downscaled, so only persons larger than the default size can be detected. This is a significant disadvantage for surveillance videos as the resolution is low and persons are usually much smaller.

The original HOG detector was reimplemented in order to see whether it has a potential to improve. The reimplementation was done by combining the HOG features

and SVM implementations from the VLFeat library [53]. The model was trained on INRIA dataset [3]. The new implementation deals with the person size variability as described in *Section 3.3*, i.e. several detectors for various sizes were trained. An example of the reimplemented HOG detector can be seen in *Figure 11*. There is no significant improvement in the detection accuracy with the exception that bounding boxes appear in appropriate sizes. A better comparison is shown in *Figure 12*, where all scales are appropriately detected. It produces less false positives and the bounding boxes are more accurate, however it does not provide much improvement in the detection performance.

Another more promising detectors were used to see whether full body detection is the right path for a scene with visible but partially occluded individuals. Publicly available implementations of DPM [5] and ACF [6] were used for this purpose. An example results of DPM and ACF detectors can be seen in *Figures 13* and *14* respectively. They both produce more accurate results that HOG detector and seem to work reasonably well. However, there are some common problems to these and perhaps most of the other person detectors when detection occluded people. First common type of error is shown in *Figure 15*. These situations occur when a person in the front occludes legs of another person in the back. This leads to detection of one "double person" which is a false detection in terms of detection-annotation matching and twice smaller count of detected people. Another common error is that an occluded person is simply omitted. Such examples can be seen in *Figure 16*. This case usually happen when legs of a person are hidden, but even a small occlusion and distraction of the usual body shape can cause that the person is "skipped". Even if legs and head of each individual are visible, the detection may not work. As shown in *Figure 17*, low contrast between bodies in a crowd is a difficult problem. Therefore even person in a front line of a bigger group are not detected properly and heads of people in the back of a group cannot be detected by a person detector in any case.

Most of the people have visible heads in the overhead camera views which means that having a quality head detector would solve the problem. Also, a head cannot appear in so many different poses such as limbs. However head itself is not a good object to be detected as the edges can vary a lot depending on the head position. Similarly, head only does not have enough strong features for a detector when the expected body size is small. On the other hand, head together with shoulders create a very typical "omega" shape that can be ideal for detection. For these reasons, several head-and-shoulders detectors were implemented. The idea of features and classifiers followed the idea of HOG and ACF detectors. Four combinations of features and classifiers were trained, namely HOG+SVM, HOG+DF, Channels+SVM and Channels+DF. All these detectors were trained on INRIA dataset and follows the scaling scheme from *Section 3.3*. The positive training examples were taken from the head area of person positives images; negative examples were randomly sampled from the provided negative images and other body parts from the positive images.

Head detectors significantly outperform person detectors in problems shown in *Figures 15* and *16*. Examples of head detection can be seen in *Figures 18* and *19*. The problem with head detectors is considerably high number of false positives. These are especially represented by some leg formations, personal things such as bags or surrounding are (especially road signs in the case of Pets2009 dataset). Also, when heads are too close to each other, they are hardly ever both detected. Although this might be just a problem of non maxima suppression is some cases. In terms of counting, head detectors

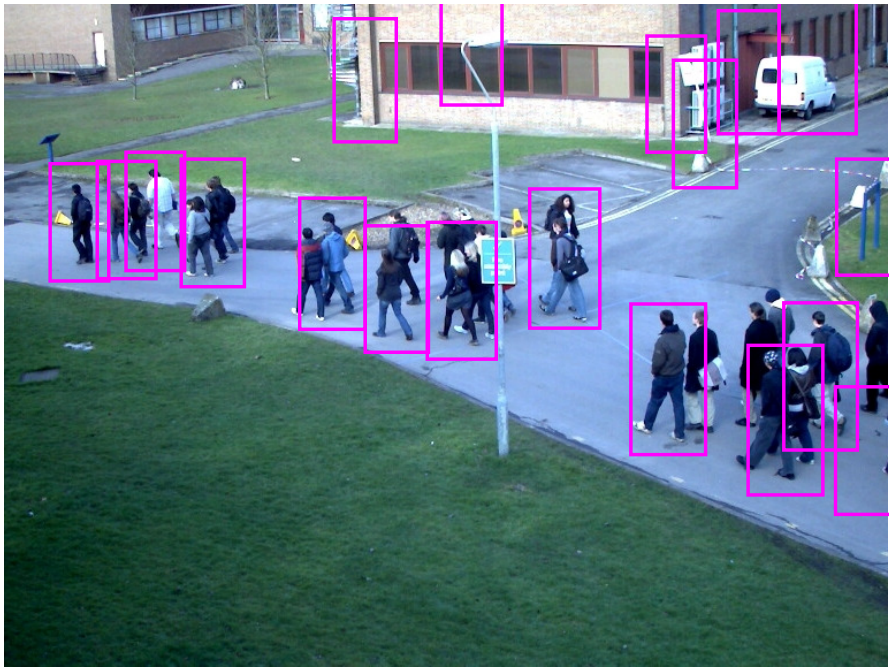


Figure 10 An example of the HOG detector result from Matlab Computer Vision toolbox.

tend to give correct or higher number of detections, in contrast to person detectors that almost always under valuate.

4.3 Evaluation metrics

Evaluation metrics depend on the type of annotation and detection. There are two types of annotations, bounding boxes around each individual and tracker bounding boxes labeled by count. Also, two types of detections are to be compared with the ground truth, head and person.

Person detection comparison with person annotation has well established evaluation metrics thoroughly described in [9]. It is based on modified version of the scheme laid out in the Pascal object detection challenges [65]. The expected output of a detector is a list of bounding boxes along with confidence score for each of them. A detected bounding box BB_{dt} and a ground truth bounding box BB_{gt} form a potential match if they overlap sufficiently. According to the Pascal measure, the area of overlap must exceed 50%, i.e.

$$a_0 \doteq \frac{area(BB_{dt} \cap BB_{gt})}{area(BB_{dt} \cup BB_{gt})} > 0.5 \tag{11}$$

As the main purpose of this work is not to get the most accurate detections but a people count, the used limits have been customized. The modification has two motivations. First, some detectors produce much bigger bounding box than the actual ground truth one. Second, overlapping persons are sometimes detected as one taller person (as shown in *Figure 15*). This is not accurate in terms of precise detection and the *Equation 11*

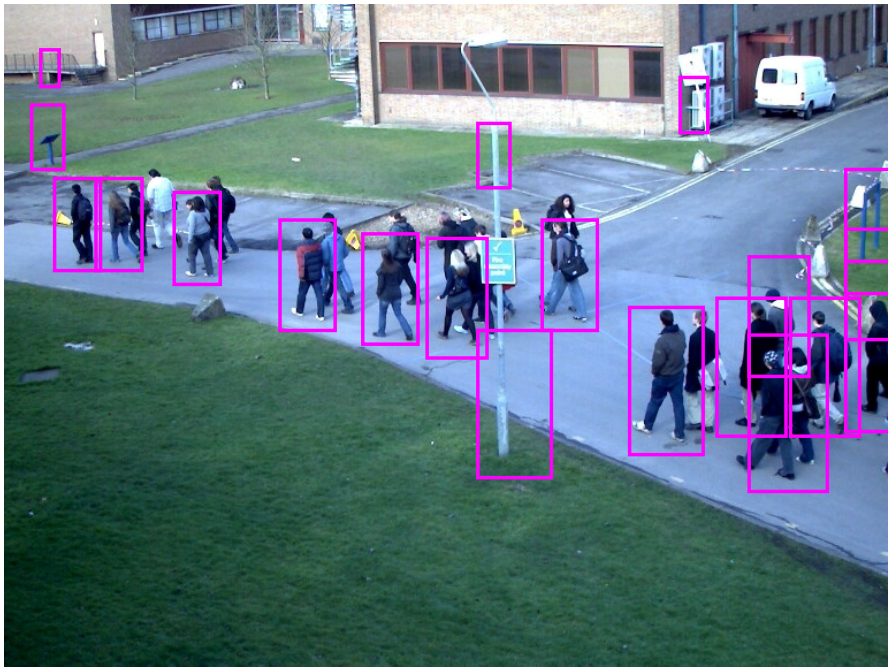


Figure 11 An example of the reimplemented HOG+SVM detector using size prediction.

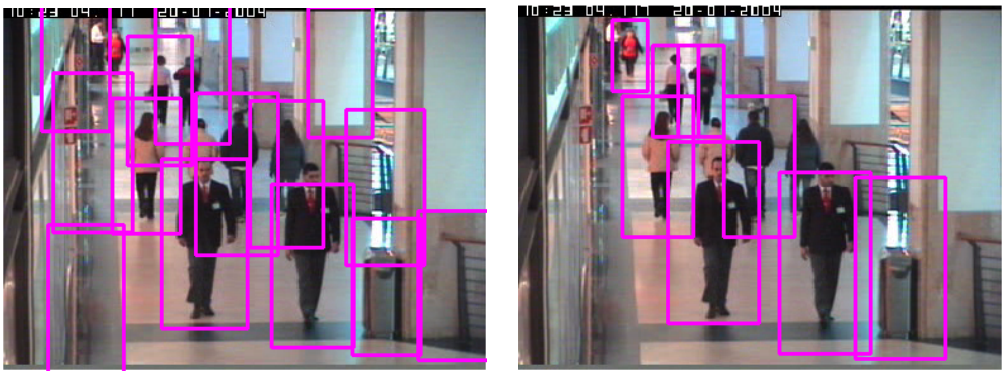


Figure 12 Comparison of the Matlab HOG (left) and HOG+SVM with size prediction (right) detection performance.

would reject it. However this detection is usually based on the real detected parts like visible head and legs, even though they belong to a different person. Therefore this type of detection is treated as one person detection. As a first step, the threshold from *Equation 11* is lowered to 0.3. If the result is smaller than 0.4, another condition is evaluated, in particular the distance between the compared bounding boxes' centers divided by the shorter side length of the two have to be higher than 0.4 in each direction. Also, the aspect ratio is unified for all bounding boxes to 0.41 as proposed in [9]. The modification is done so that the height and center remain unchanged, only the width is adjusted.

Each BB_{dt} and BB_{gt} may be matched at most once. Any assignment ambiguity is resolved by performing the matching greedily. Detections with highest confidence are matched first; if a detected bounding box matches multiple ground truth boxes, the

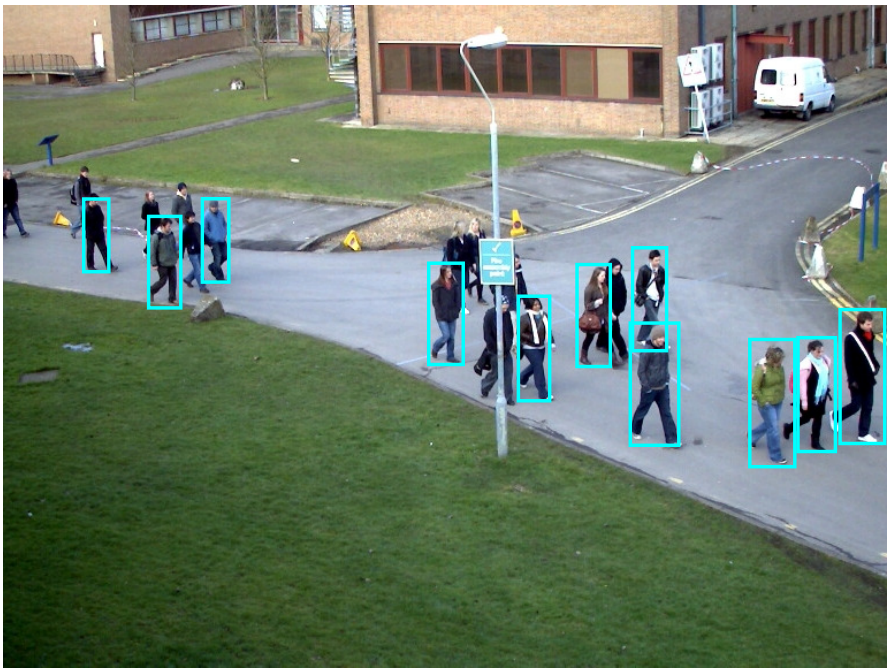


Figure 13 An example of the DPM detection result

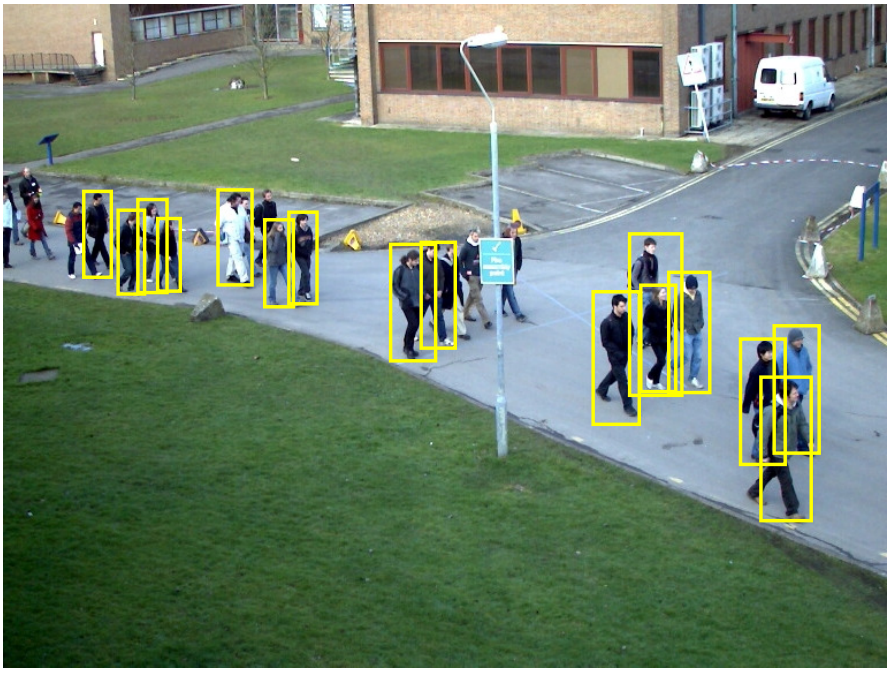


Figure 14 An example of the ACF detector result

Figure 15 Examples of fault detections produced by person detectors when presented with images containing frequent occlusion. Two people behind each other are detected as a "double person".

match with the highest overlap is used. Unmatched BB_{dt} count as false positives and unmatched BB_{gt} as false negatives.

The same scheme is used for comparison of *head detections*. With the difference that head detections are first extracted from the person ground truth annotations. The top quarter of such a annotation is taken in order to convert it to head annotation.

The same philosophy is also used for the *non maximal suppression*. If the expression in *Equation 11* for two detections is greater than the overlap threshold, then the bounding box with the lower score is suppressed.

The comparison of detector performance is done by evaluation of precision-recall (PR) and miss rate against false positives per image (ROC) curves. Both curves are obtained by varying the threshold on detection confidence. Precision and recall is obtained as

$$\text{precision} = \frac{\#\text{ correct detections}}{\#\text{ detections}} \tag{12}$$

Figure 16 Examples of fault detections produced by person detectors when presented with images containing frequent occlusion. "Skipping" of occluded people.

$$\text{recall} = \frac{\#\text{ correct detections}}{\#\text{ ground truth annotations}} \tag{13}$$

The number of correct detections is a number of detections that satisfied the condition 11 according to the procedure described above. However some videos have only ROI annotated by a corresponding person count. I this case the number of correct detections is given by simple comparison of the ground truth count $\#gt$ and the number of detections in the ROI $\#det$

$$\#\text{ correct detections} = \begin{cases} \#gt & \text{if} \quad \#gt < \#det \\ \#det & \text{if} \quad \#gt > \#det \end{cases} \tag{14}$$

The miss rate against false positives per image curve is a log-log plot. This type of curve is in most of the pedestrian detection publications preferred to precision recall curves as in certain tasks the acceptable false positives per image (FPPI) rate is independent of pedestrian density. A detail description of computing this type of curve is in [9].

The methods are tested on both full images and ROI given by the tracker objects. To minimize error caused by inaccurate detector and ROI size, the ROI bounding boxes

Figure 17 An example of the person detectors' performance on a crowd.

are always enlarged by 20 pixels so that the feature space is large enough to find a person even if the tracker object edges are very tight around the tracked person.

4.4 Parameter settings

Different combinations of features and classifiers are evaluated for both head and person detectors. Both types are trained on the INRIA dataset [3]. All tested types of features and classifiers have various parameters that can be set. In this section, different parameter settings are compared on ground truth annotation of view 1 from the Pets2009 dataset. The detection has been done on full frames in order to evaluate the overall tendency to produce false positives. The results for head detectors are shown in *Figures 20-23*. Person detectors trained by these methods are not further used for testing, however the results of parameter settings can be found in *Appendix B*.

HOG features have two parameters to be set, number of orientations NO and cell size cs. The channel features are used in the "classic" configuration 6 oriented gradients

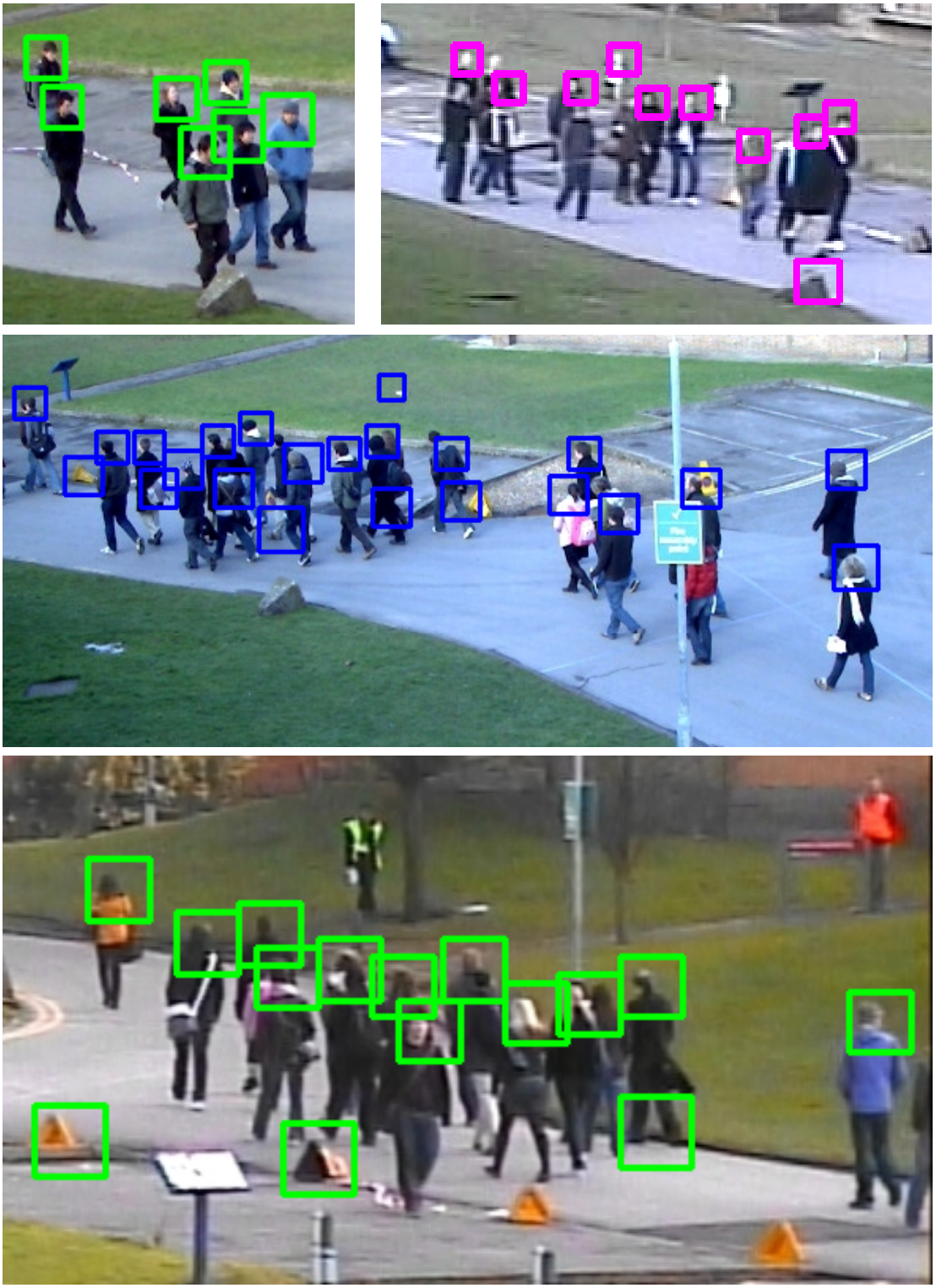


Figure 18 Examples of the head detectors' performance

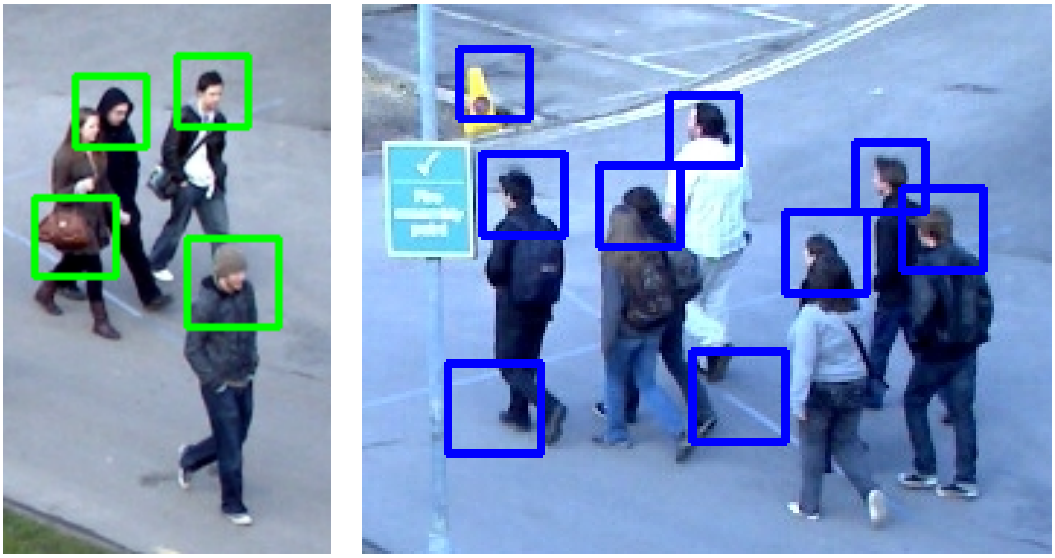


Figure 19 Examples of the typical false positives of head detectors.

channels, 1 gradient magnitude and 3 channels for LUV color. One pixel of the channel can represent an arbitrary large square region of the original image whose size is for simplicity also referred as cell size *cs* (for example cell size 4 means that the channels are 4 times smaller than the original image). Decision forests trained by boosting were tested with two adjustable parameters, the maximum numbers of weak classifiers w (i.e. the number of trees) and tree depth d. The SVM classifier has no variable settings.

Smaller cell size *cs* is generally better for head detection as the detector's window is relatively small and the shape does not vary a lot. Also the number of orientations *NO* does not have to be high in order to produce more accurate results. This covers the fact that higher *NO* leads to more sparse representation and is more sensitive to lower contrast. Generally the accuracy increases with the number of weak classifiers w and tree depth d. The difference is not that crucial for channel features as the number of features is lower however is quite significant for HOG features.

Only one detector from each "group" is selected for further testing. Based on the results in *Figures 20-23*, these detectors are selected: "*HOG-SVM-head NO 6 cs 6*", "*Chns-SVM-head cs 2*", "*Chns-Boost-head cs 4 w 256 d 3*" and "*HOG-Boost-head NO 6 bs 4 w 1024 d 2*".

4.5 Detection accuracy

The objective of this thesis is to count people in the tracker objects. In terms of detections, this means that only ROI framed by the detector objects' bounding boxes have to be searched for detections. This significantly reduces the number of false positives especially for head detectors. The provided ROI usually include a significant parts of a frame and any assumption about the bounding box cannot be made with a reasonable reliability. In order to reduce false negatives, the ROI are enlarged by 20 pixels so that larger detectors have enough data for detection.

Precision-recall and miss rate against false positives per image curves for all tested

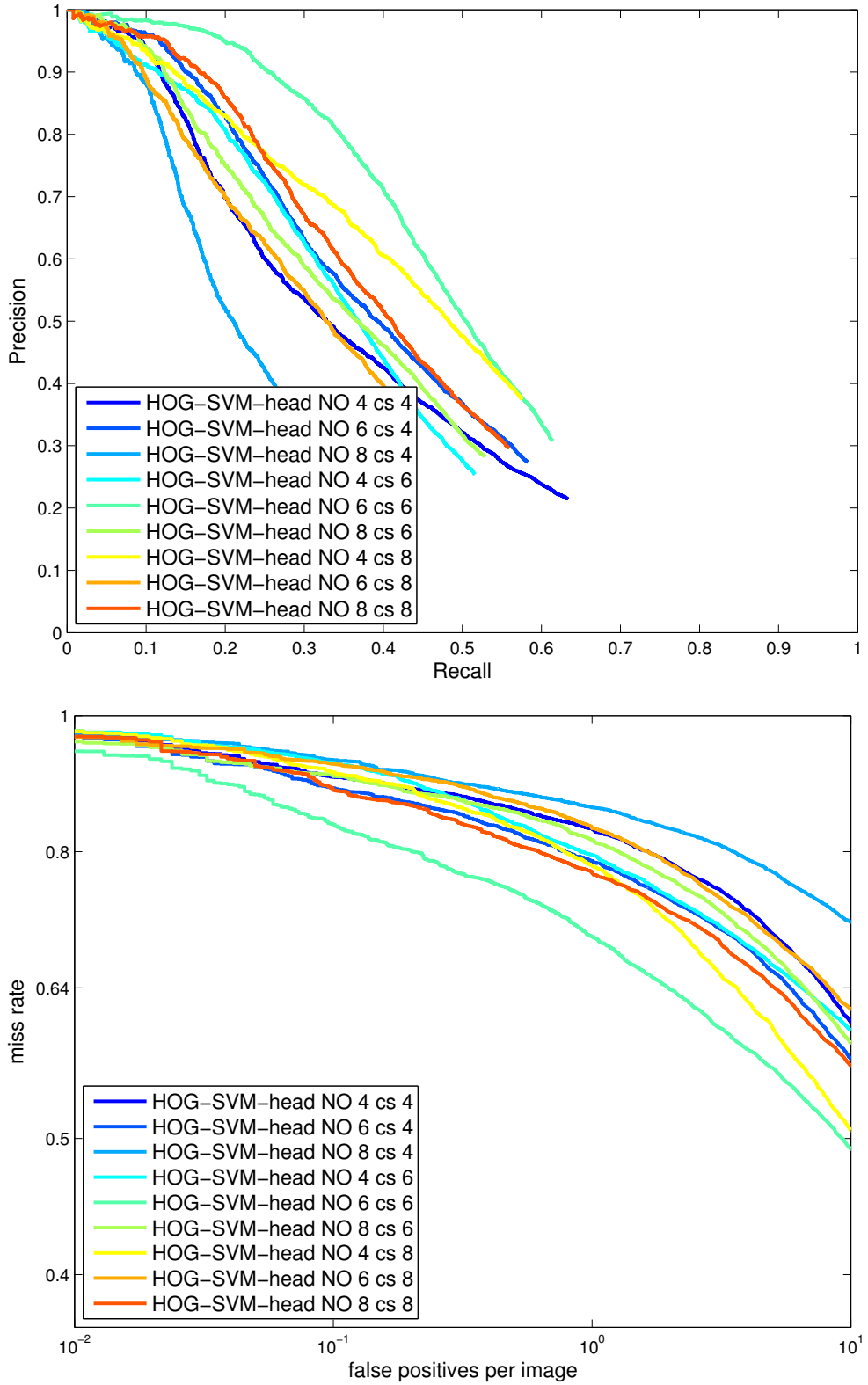


Figure 20 Precision-recall (top) and miss rate against false positives per image (bottom) curves for HOG+SVM head detector with different parameter settings

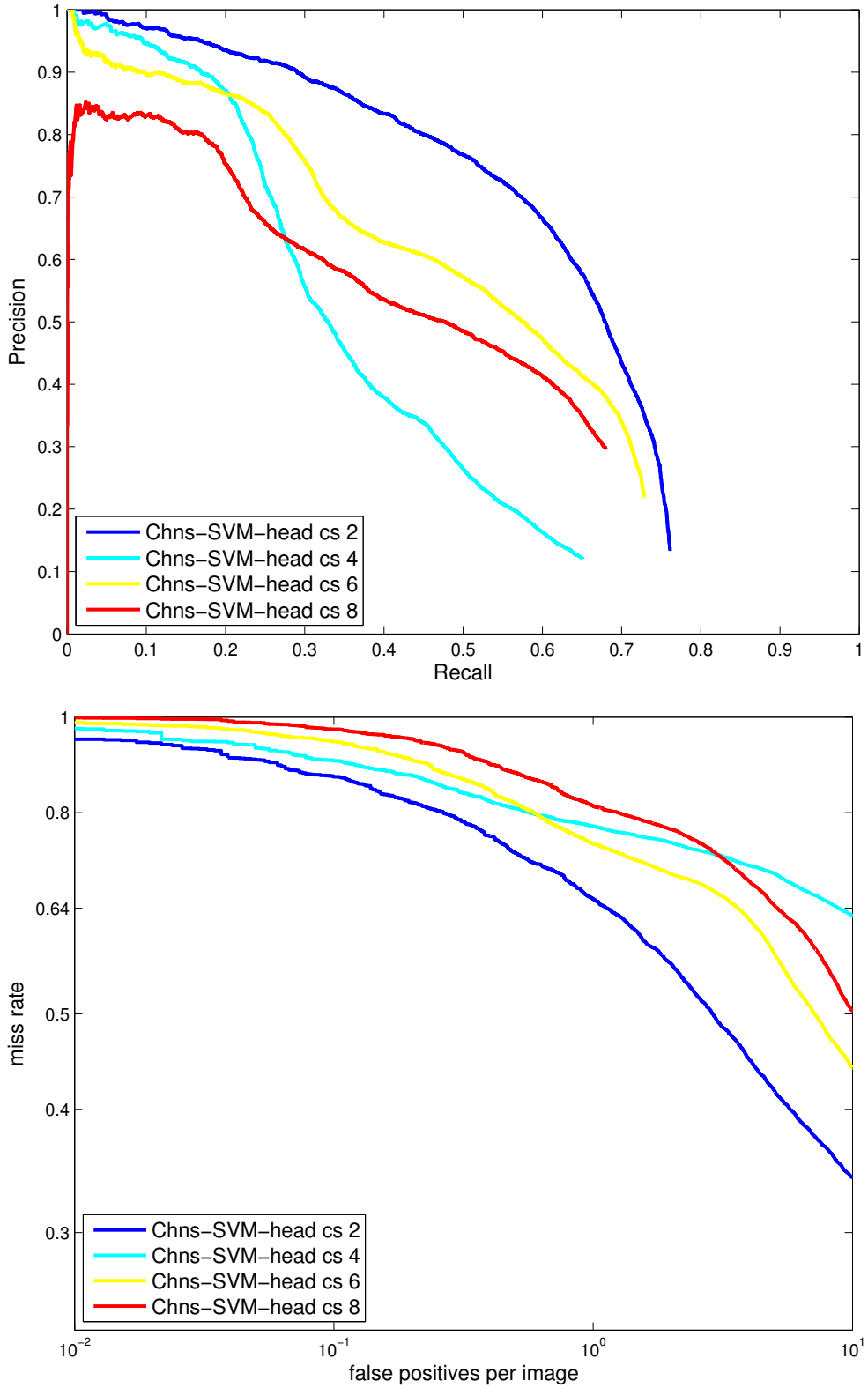


Figure 21 Precision-recall (top) and miss rate against false positives per image (bottom) curves for Chns+SVM head detector with different parameter settings

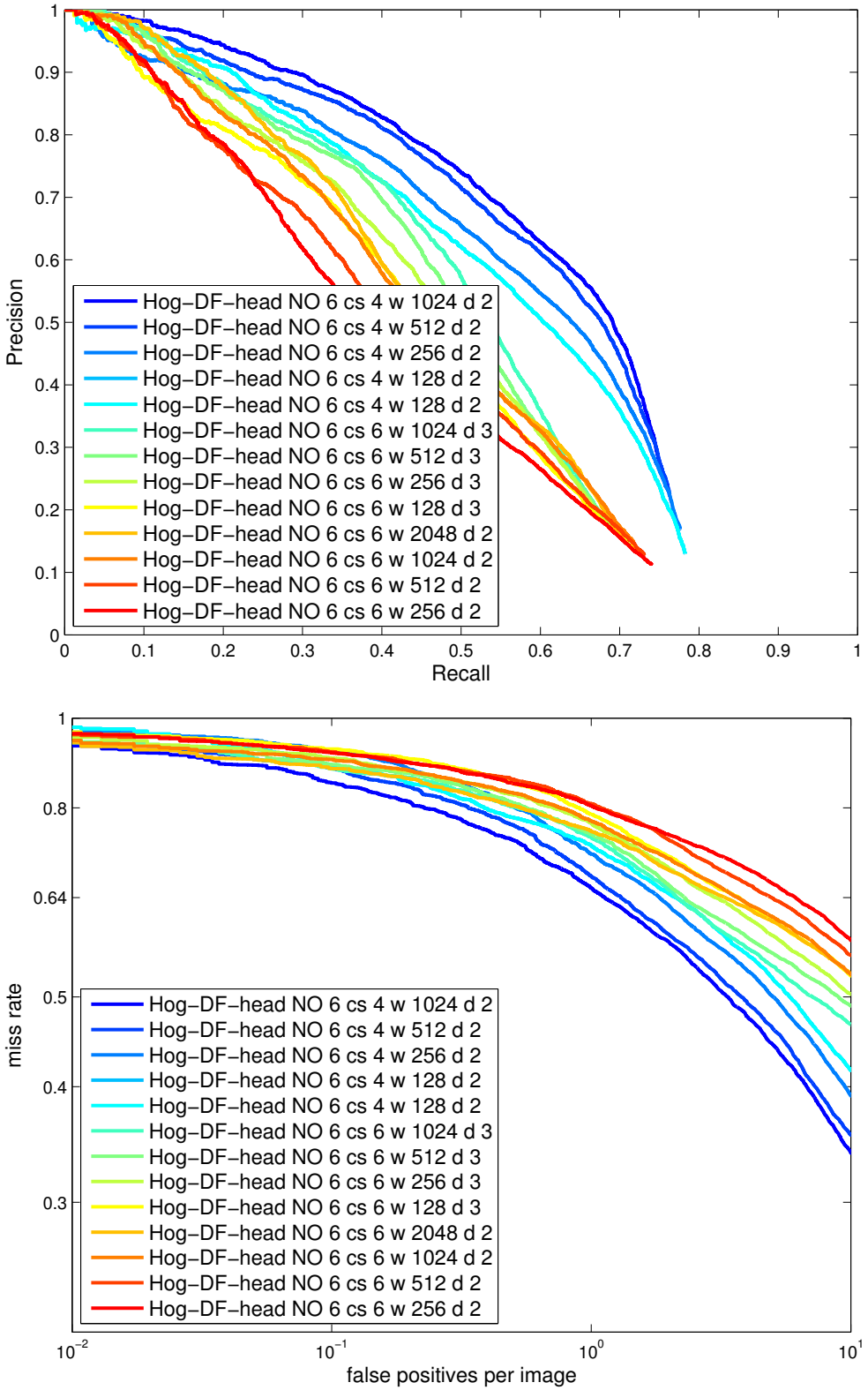


Figure 22 Precision-recall (top) and miss rate against false positives per image (bottom) curves for HOG+DF head detector with different parameter settings

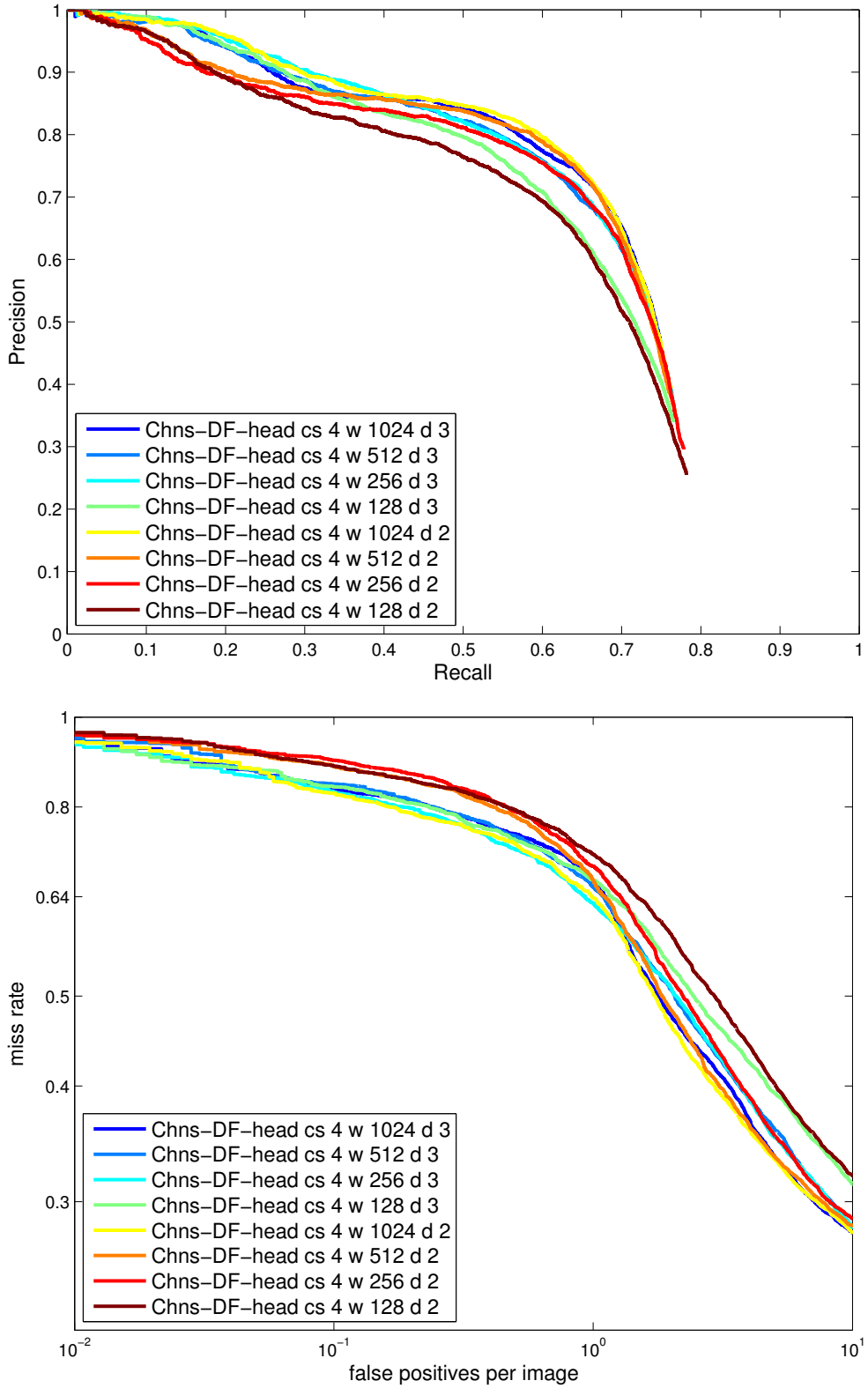


Figure 23 Precision-recall (top) and miss rate against false positives per image (bottom) curves for Chns+DF head detector with different parameter settings

person and head detectors are shown in *Figure 24*. These curves were computed on view 1 of the Pets2009 dataset. All considered detections and annotations have to be inside the enlarged ROI. The person detectors have the lowest number of false positives, especially the HOG detector. However, none of them is capable of detecting more that 60% of the annotated persons. Unfortunately, even head detectors are not exceeding 70%. Chns+DF detector is the most accurate out of all tested head detectors. The worst performance give the detectors that classify by SVM.

Visual comparisons are available in *Figures 25-27*, only the ROI are selected from the original frames for better visibility. The default threshold is 0 for all detectors, however the threshold has been increased for some of them, particularly Chns-SVM-head to 1.5, Chns-DF-head and HOG-DF-head to 4 (the DF detectors produce detections with higher confidence in general). As it is hard to sufficiently present video detections on paper, example videos with ROI and detection bounding boxes are included in the DVD that is a part of *Appendix A*.

Person detectors usually produce a lower number of detections but they are more accurate. On the other hand head detectors often produce more false positives and the optimal thresholding often depend on a particular scene (contrast, light conditions est.). For this reason DF classifiers are significantly better as they are much more accurate without finely tuned threshold. The least accurate is combination of Chns+SVM that uses cell size only 2. Therefore there might be a high number of features that are not stable and lead the detector to produce a vast number of false positives. The combination of Chns+DF is the best based on observations and the curves in *Figure 24*.

4.6 Count estimation

The count for each tracker object is estimated by counting the number of detected persons/heads inside of them. The count estimation is tested against ground truth count, i.e. bounding boxes of the detections are not matched with ground truth person/head bounding boxes. Two types of tests were done, plotting the number of detections over time and precision and recall values.

The number of people in the videos change over time as well as the number of detection. The evolution of number of poeple over time for some of the Pets2009 dataset videos can be seen in *Figures 28-29*, one figure for one of the four views. There are two graphs in each Figure, one for head detectors and the other for person detectors. The graphs show frame by frame detection results for two concatenated videos, namely S1-L1-Time13-57 and S1-L1-Time13-59. Each unit on y-axis represent one frame. The y-axis shows sum of people count over all detector objects in a particular frame. The threshold of each detector is set to 0 in order to see what is the largest number of detections produced by each detector. It is apparent that the detection quality depends on the view. The size of detected people as well as lighting conditions and contrast are different for each view . Also, the density of people is different for each view. In general, people detectors always under-valuate the result (the closes estimation is given by ACF). On the other hand, Chns+SVM is almost always overestimating. Among the head detectors, HOG+SVM seem to be the most stable detector as it never significantly over- or under- estimates. Both head detectors with DF provide very similar results and are the most accurate ones in most cases.

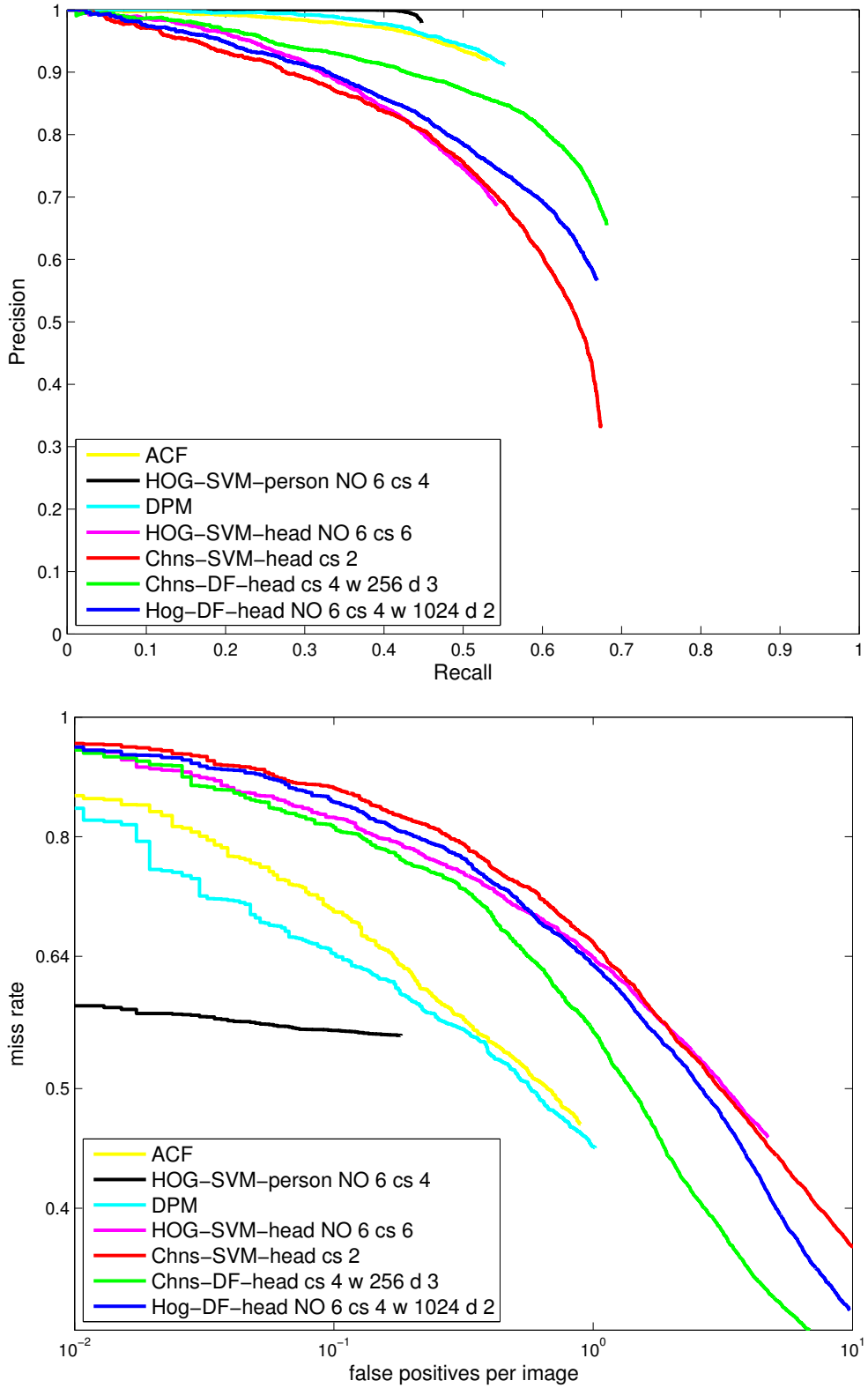


Figure 24 Precision-recall (top) and miss rate against false positives per image (bottom) curves for selected head and person detector

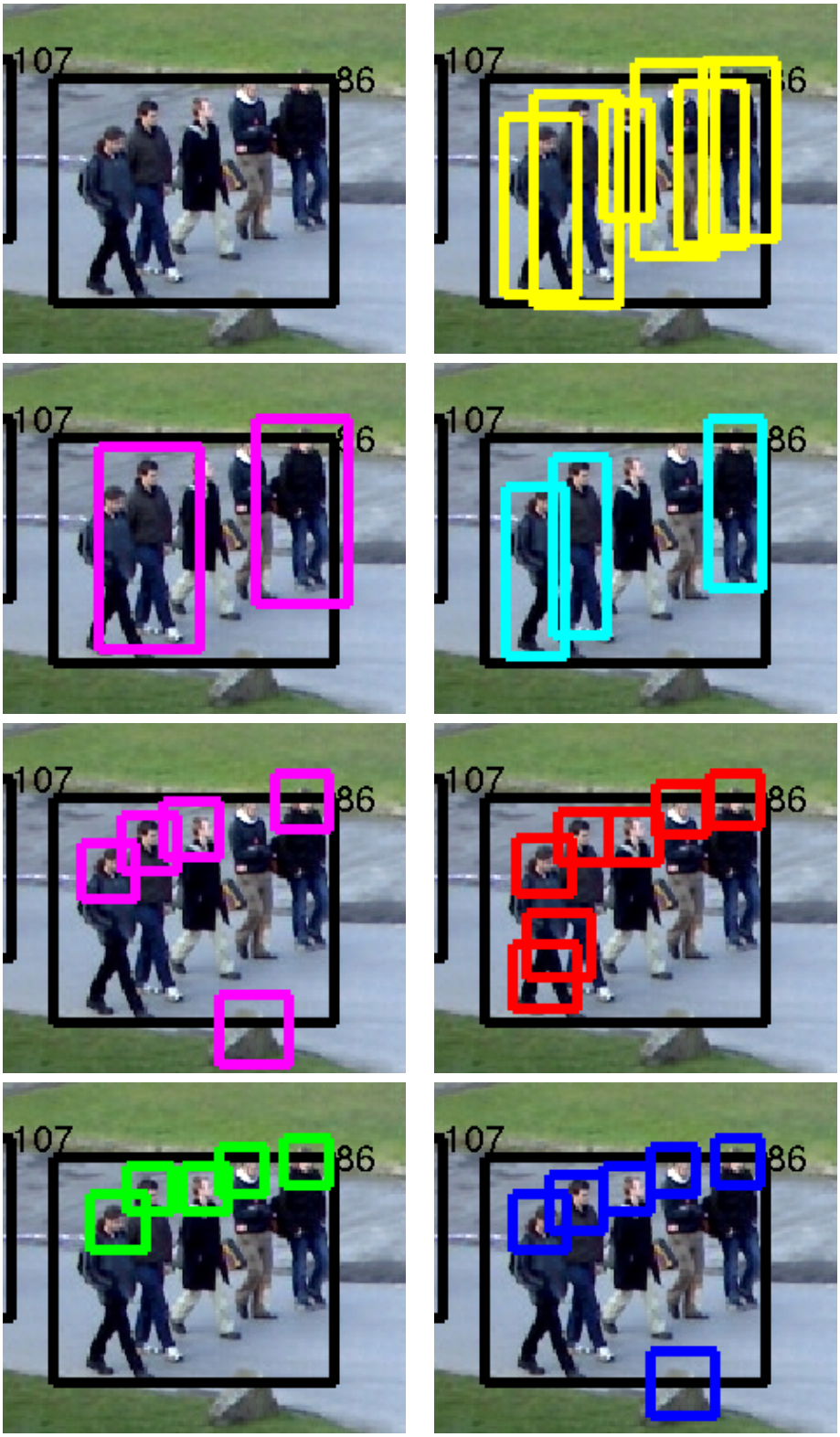


Figure 25 An example visualization of the performance of the head and person detectors (colors assigned as described in *Table 1)*

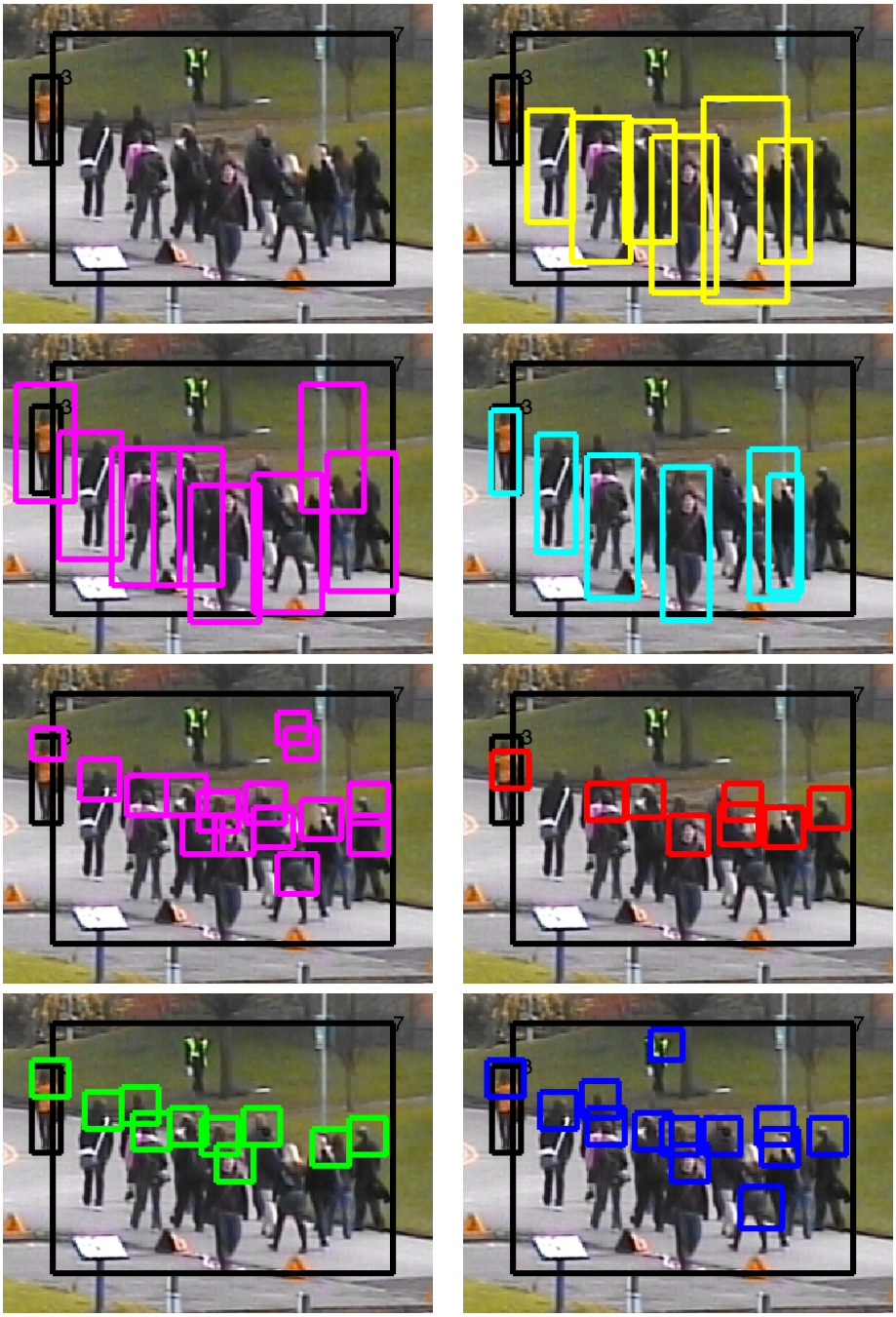


Figure 26 An example visualization of the performance of the head and person detectors (colors assigned as described in *Table 1)*

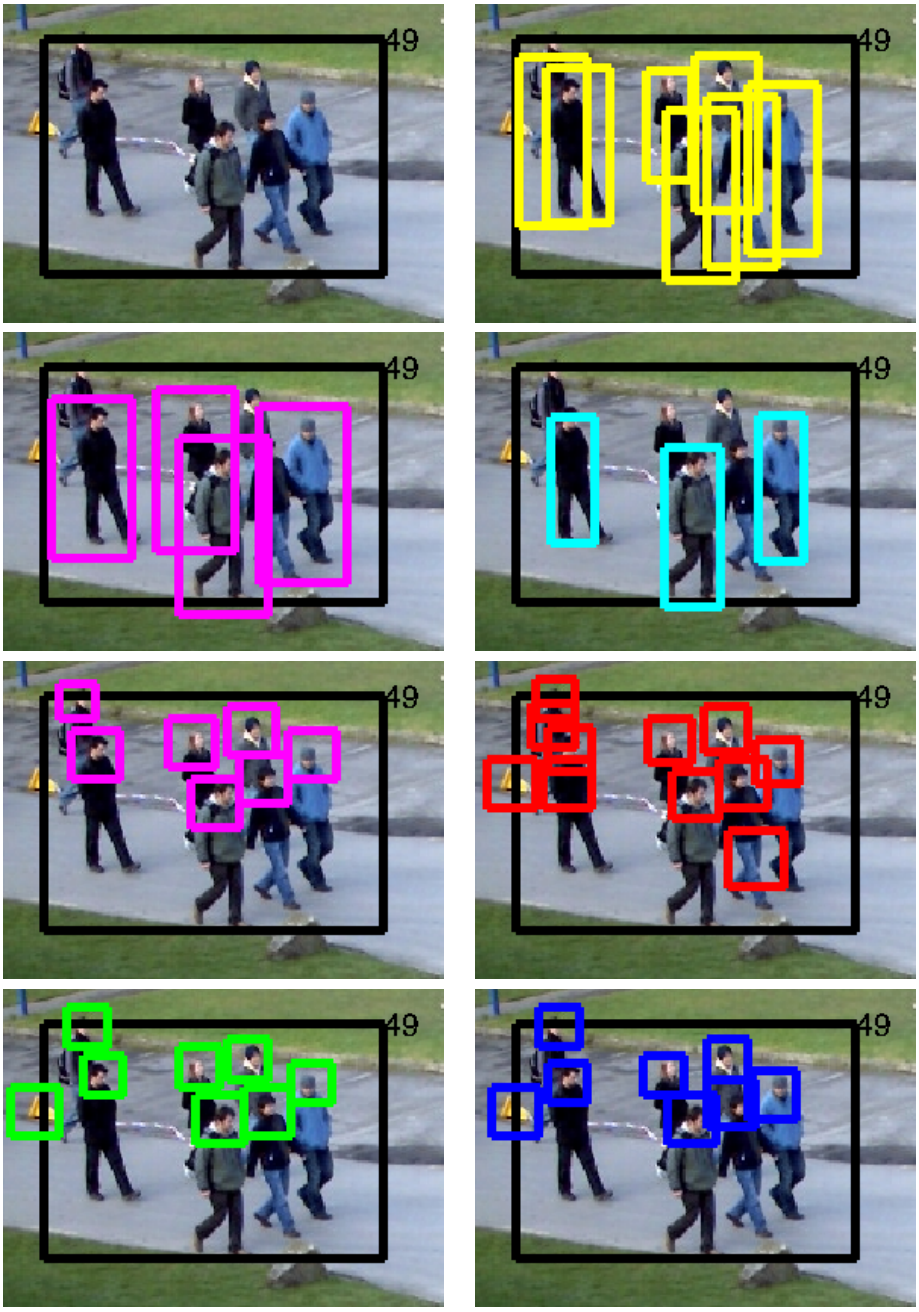


Figure 27 An example visualization of the performance of the head and person detectors (colors assigned as described in *Table 1)*

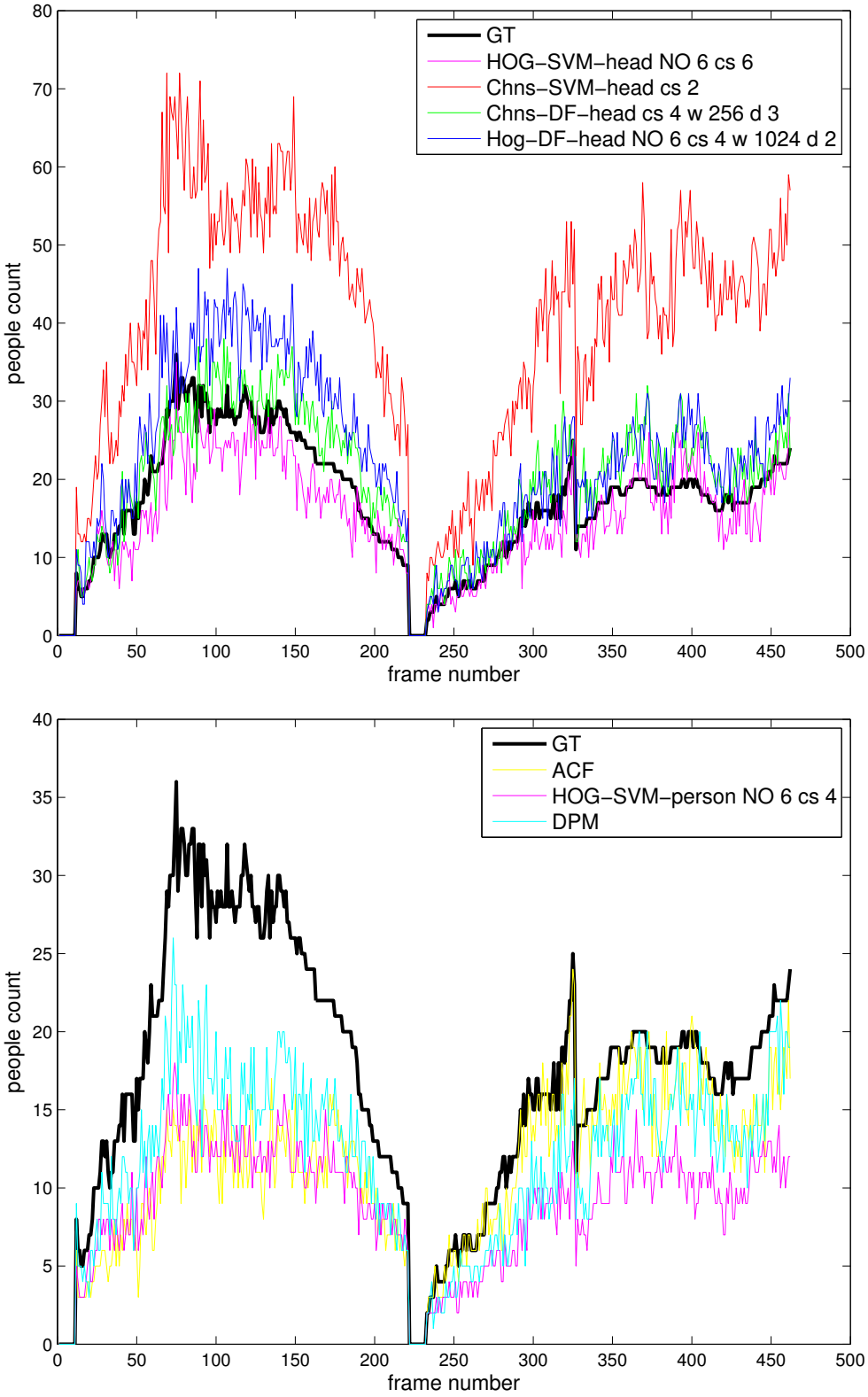


Figure 28 Count estimation results of the head (top) and person (bottom) detectors on the Pets2009 dataset VIEW 1

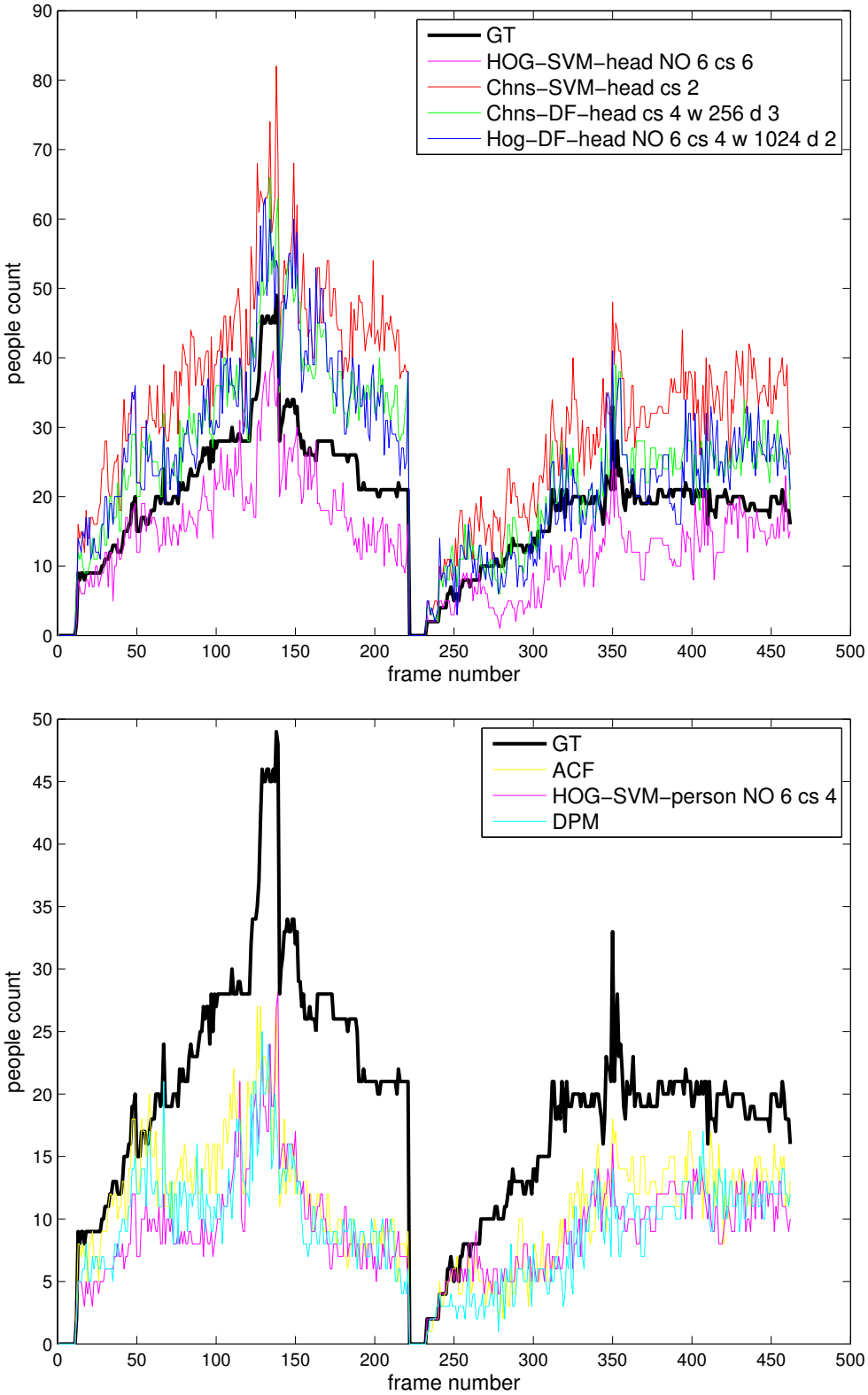


Figure 29 Count estimation results of the head (top) and person (bottom) detectors on the Pets2009 dataset VIEW 2

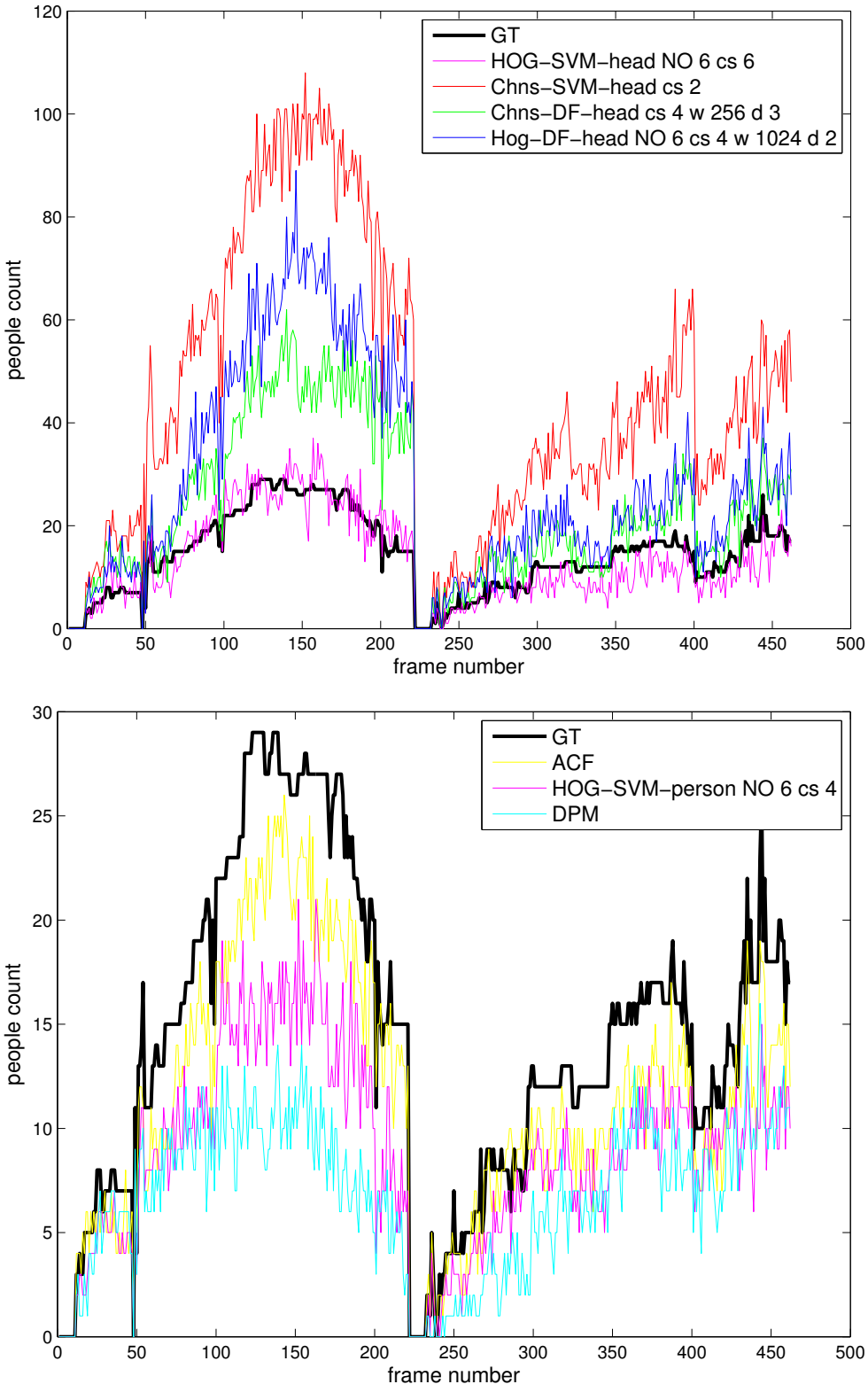


Figure 30 Count estimation results of the head (top) and person (bottom) detectors on the Pets2009 dataset VIEW 3

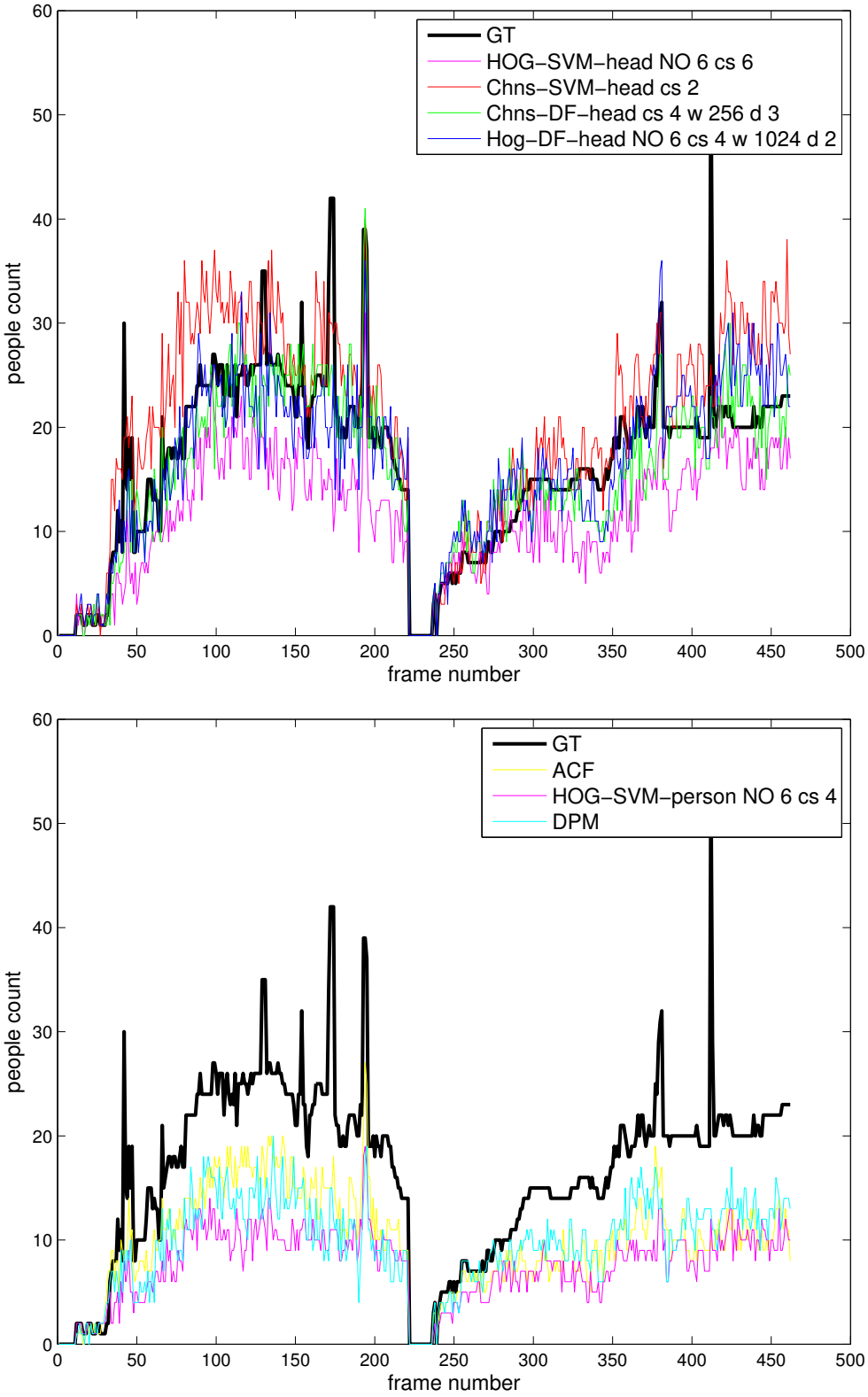


Figure 31 Count estimation results of the head (top) and person (bottom) detectors on the Pets2009 dataset VIEW 4

Precision and recall statistics were generated for each view separately and the results are shown in *Tables 3-5*. Thresholds were set to 0 for all detectors as in *Figures 28-29*. The precision an recall were computed by *Equations 12-14* for each tracker object separately. The shown values are then averages of results given for each tracker object. This way, incorrect estimations for bounding boxes with less people are penalized strongly (i.e. difference of 1 detection contributes differently than the same difference for an object with 30 people). HOG+SVM person detector gives the best precision with a comparable recall as ACF detector. The recall of DPM is rather insufficient. Also for the head detection, HOG+SVM detector has the best precision, however has significantly lower recall than the other head detectors. Considering the fact that the threshold was set to 0 for all detectors and can be increased in a practical implementation, the Chns+DF detector gives the most promising results.

	Precision			
Detector	view 1	view 2	view 3	view 4
ACF	0.9510	0.9843	0.9516	0.5344
HOG-SVM-person NO 6 cs 4	0.9856	0.9975	0.9939	0.9453
DPM	0.9004	0.8609	0.7387	0.8206
HOG-SVM-head NO 6 cs 6	0.9068	0.9559	0.9443	0.7866
Chns-SVM-head cs 2	0.4491	0.6507	0.3648	0.7985
Chns-Boost-head cs 4 w 256 d 3	0.7986	0.7944	0.7198	0.8271
HOG-Boost-head NO 6 bs 4 w 1024 d 2	0.7575	0.8260	0.6589	0.7790

Table 3 Average precision of the count estimation on tracker objects for each view of the Pets2009 dataset separately

	Recall			
Detector	view 1	view 2	view 3	view 4
ACF	0.7523	0.5983	0.8203	0.3784
HOG-SVM-person NO 6 cs 4	0.6242	0.5917	0.7127	0.6811
DPM	0.5122	0.2824	0.2825	0.4812
HOG-SVM-head NO 6 cs 6	0.8333	0.6994	0.8350	0.6582
Chns-SVM-head cs 2	0.9989	0.9998	1.0000	0.9609
Chns-Boost-head cs 4 w 256 d 3	0.9780	0.9840	0.9918	0.9130
HOG-Boost-head NO 6 bs 4 w 1024 d 2	0.9788	0.9638	0.9964	0.8392

Table 4 Average recall of the count estimation on tracker objects for each view of the Pets2009 dataset separately

Detector	Precision	Recall
ACF	0.8553	0.6373
HOG-SVM-person NO 6 cs 4	0.9806	0.6524
DPM	0.8301	0.3896
HOG-SVM-head NO 6 cs 6	0.8984	0.7565
Chns-SVM-head cs 2	0.5658	0.9899
Chns-Boost-head cs 4 w 256 d 3	0.7850	0.9667
HOG-Boost-head NO 6 bs 4 w 1024 d 2	0.7553	0.9446

Table 5 Combined average precision and recall of the count estimation on tracker objects for all views of the Pets2009 dataset

5 Conclusion and future work

Motivated by the need to improve tracking performance of legacy trackers, the main focus of the presented work was estimation of the number of persons in images/video sequences of a semi-crowded scene with frequently occurring occlusions. Model-based methods, which give the estimated number of people by counting detections, were investigated.

Two types of detectors, full body (*person*) and head-and-shoulders (*head*) were investigated. The Aggregate Channel Features (ACF) [6], Deformable Part Models (DPM) [5] and reimplemented HOG+SVM [3] were tested as person detectors. The head detectors are also based on the ideas from these methods. Four combinations of HOG or channel [4] features with SVM or decision forest classifiers, i.e. HOG+SVM, HOG+DF, Chns+SVM and Chns+DF, were tested as head detectors. Pretrained models provided by authors were used for ACF and DPM, other methods were implemeted in Matlab with help of VLFeat [53] and Piotr's [54] Computer Vision toolboxes.
The implemented methods use a custom approach to scaling, in contrast with ACF and DPM . Instead of resizing the input frame or the features, several detectors were trained for different sizes of expected detection. An approximate size for each position in the frame is estimated and an appropriate detector is selected for detection. Therefore, the features can be computed only once on the entire image or region of interest and then reused by different detectors.

The most appropriate settings were estimated for all of the implemented detectors. The experiments and evaluations showed that the person detectors are more accurate, however they miss a lot of partially occluded persons. The HOG detector gives the most precise results on the ROI given by the tracker, however both ACF and DPM are capable of retrieving more people on certain types of scenes. On the other hand, the head detectors tend to produce a lot of false positives and have to be properly thresholded. The combination of Chns+DF seems to be the most accurate one.

Based on the experiments and observations, I have concluded that head detectors perform better for the given task and I suggest a solution using the Chns+DF combination. It has been shown that a simple frame by frame detection can give a sufficiently accurate estimation of people count, even if the density of people is high.

I suggest several topics with potential to further improve the results of presented work. First step could be elimination of head false positives by checking the confidence of a person detector behind the head detection. Another improvement could be to update the number of people in each tracked object, which would significantly smooth the results. The solution could also take an advantage of the fact that the analyzed media is a video. Therefore, information from previous frames could be used to improve detection accuracy in a current frame.

Further, there are methods based on different paradigms that deserve to be investi-

gated as potential solutions to people count estimation. The fact that all individuals are being detected has a lot of disadvantages. The scene with frequent occlusions can be very complex and detection becomes a very hard problem. For example the map-based methods [2] may achieve comparable estimation accuracy while avoiding shortcomings of methods investigated in presented work.

Bibliography

[1] *PETS 2009 Benchmark Data*. `http://www.cvg.reading.ac.uk/PETS2009/a.html`.

[2] M. Hashemzadeh, G. Pan, and M. Yao. "Counting moving people in crowds using motion statistics of feature-points". In: *Multimedia Tools and Applications* 72 (2014), pp. 453–487.

[3] N. Dalal and B. Triggs. "Histograms of oriented gradients for human detection". In: *Conference on Computer Vision and Pattern Recognition (CVPR)* 1 (2005), pp. 886–893.

[4] P. Dollar et al. "Integral channel features". In: *British Machine Vision Conference (BMVC)* (2009).

[5] P. Felzenszwalb et al. "Object detection with discriminatively trained part-based models". In: *Transactions on Pattern Analysis and Machine Intelligence (PAMI)* (2010).

[6] P. Dollar et al. "Fast feature pyramids for object detection". In: *Transactions on Pattern Analysis and Machine Intelligence (PAMI)* (2014).

[7] P. Dollar et al. "Pedestrian detection: A benchmark". In: *Conference on Computer Vision and Pattern Recognition (CVPR)* (2009).

[8] A. Geiger, P. Lenz, and R. Urtasun. "Are we ready for autonomous driving? the kitti vision benchmark suite". In: *Conference on Computer Vision and Pattern Recognition (CVPR)* (2012).

[9] P. Dollar et al. "Pedestrian detection: An evaluation of the state of the art". In: *Transactions on Pattern Analysis and Machine Intelligence (PAMI)* (2011).

[10] R. Benenson et al. "Ten years of pedestrian detection, what have we learned?" In: *European Conference on Computer Vision (ECCV)* (2014).

[11] P. Viola and M. Jones. "Rapid object detection using a boosted cascade of simple features". In: *Conference on Computer Vision and Pattern Recognition (CVPR)* (2001).

[12] H. Jia and Y. Zhang. "Fast Human Detection by Boosting Histograms of Oriented Gradients". In: *Fourth International Conference on Image and Graphics (IGIG)* (2007), pp. 683–688.

[13] E. Corvee and F. Bremond. "Body Parts Detection for People Tracking Using Trees of Histogram of Oriented Gradient Descriptors". In: *International Conference on Advanced Video and Signal Based Surveillance (AVSS)* (2010), pp. 469–475.

[14] D. Park, D. Ramanan, and C. Fowlkes. "Multiresolution models for object detection". In: *European Conference on Computer Vision (ECCV)* (2010).

[15] W. Ouyang and X. Wang. "Single-pedestrian detection aided by multi-pedestrian detection". In: *Conference on Computer Vision and Pattern Recognition (CVPR)* (2013).

[16] J. Yan et al. “Robust multi-resolution pedestrian detection in traffic scenes”. In: *Conference on Computer Vision and Pattern Recognition (CVPR)* (2013).

[17] S. Zhang, C. Bauckhage, and A.B. Cremers. “Informed haar-like features improve pedestrian detection”. In: *Conference on Computer Vision and Pattern Recognition (CVPR)* (2014).

[18] S. Zhang, R. Benenson, and B. Schiele. “Filtered channel features for pedestrian detection”. In: *Conference on Computer Vision and Pattern Recognition (CVPR)* (2015).

[19] P. Viola, M. Jones, and D. Snow. “Detecting pedestrians using patterns of motion and appearance”. In: *Conference on Computer Vision and Pattern Recognition (CVPR)* (2003).

[20] P. Dollar, S. Belongie, and P. Perona. “The fastest pedestrian detector in the west”. In: *British Machine Vision Conference (BMVC)* (2010).

[21] P. Dollar, R. Appel, and W. Kienzle. “Crosstalk cascades for frame-rate pedestrian detection”. In: *European Conference on Computer Vision (ECCV)* (2012).

[22] W. Nam, P. Dollar, and J.H. Han. “Local Decorrelation For Improved Pedestrian Detection”. In: *Conference on Neural Information Processing Systems (NIPS)* (2014).

[23] F. Khan et al. “Color attributes for object detection”. In: *Conference on Computer Vision and Pattern Recognition (CVPR)* (2012).

[24] R. Khan et al. “Discriminative color descriptors”. In: *Discriminative color descriptors* 2013 (Conference on Computer Vision and Pattern Recognition (CVPR)).

[25] X. Wang, X. Han, and S. Yan. “An hog-lbp human detector with partial occlusion handling”. In: *International Conference on Computer Vision (ICCV)* (2009).

[26] X. Wang et al. “Regionlets for generic object detection”. In: *International Conference on Computer Vision (ICCV)* (2013).

[27] S. Paisitkriangkrai, C. Shen, and A. Van den Hengel. “Strengthening the effectiveness of pedestrian detection with spatially pooled features”. In: *European Conference on Computer Vision (ECCV)* (2014).

[28] S. Walk et al. “New features and insights for pedestrian detection”. In: *Conference on Computer Vision and Pattern Recognition (CVPR)* (2010).

[29] Y. Goto, Y. Yamauchi, and H. Fujiyoshi. “Color similarity-based hog”. In: *Korea-Japan Joint Workshop on Frontiers of Computer Vision* (2013).

[30] P. Ott and M. Everingham. “Implicit color segmentation features for pedestrian and object detection”. In: *Conference on Computer Vision and Pattern Recognition (CVPR)* (2009).

[31] D. Ramanan. “Using segmentation to verify object hypotheses”. In: *Conference on Computer Vision and Pattern Recognition (CVPR)* (2007).

[32] Y. Socarras et al. “Improving hog with image segmentation: Application to human detection”. In: *Advanced Concepts for Intelligent Vision Systems* (2012).

[33] O. Tuzel, F. Porikli, and P. Meer. “Pedestrian detection via classification on riemannian manifolds”. In: *Transactions on Pattern Analysis and Machine Intelligence (PAMI)* (2008).

[34] P. Sermanet et al. “Pedestrian detection with unsupervised multi-stage feature learning”. In: *Conference on Computer Vision and Pattern Recognition (CVPR)* (2013).

[35] W. Ouyang and X. Wang. “Joint deep learning for pedestrian detection”. In: *International Conference on Computer Vision (ICCV)* (2013).

[36] P. Luo et al. “Switchable deep network for pedestrian detection”. In: *Conference on Computer Vision and Pattern Recognition (CVPR)* (2014).

[37] W. Ouyang and X. Wang. “A discriminative deep model for pedestrian detection with occlusion handling”. In: *Conference on Computer Vision and Pattern Recognition (CVPR)* (2012).

[38] W. Ouyang, X. Zeng, and X. Wang. “Modeling mutual visibility relationship with a deep model in pedestrian detection”. In: *Conference on Computer Vision and Pattern Recognition (CVPR)* (2013).

[39] J. Hosang et al. “Taking a deeper look at pedestrians”. In: *Conference on Computer Vision and Pattern Recognition (CVPR)* (2015).

[40] T. Gao, B. Packer, and D. Koller. “A segmentation-aware object detection model with occlusion handling”. In: *Conference on Computer Vision and Pattern Recognition (CVPR)* (2011).

[41] A. Vedaldiand and A. Zisserman. “Structured output regression for detection with partial occulsion”. In: *Conference on Neural Information Processing Systems (NIPS)* (2009).

[42] S. Tang, M. Andriluka, and B. Schiele. “Detection and Tracking of Occluded People”. In: *International Journal of Computer Vision (IJCV)* 110(1) (2014), pp. 58–69.

[43] B. Pepik et al. “Occlusion Patterns for Object Class Detection”. In: *Conference on Computer Vision and Pattern Recognition (CVPR)* (2013).

[44] B. Pepik et al. “Teaching 3D geometry to deformable part models”. In: *Conference on Computer Vision and Pattern Recognition (CVPR)* (2012).

[45] S. Tang et al. “Learning People Detectors for Tracking in Crowded Scenes”. In: *International Conference on Computer Vision (ICCV)* (2013).

[46] B. Wu and R. Nevatia. “Tracking of multiple humans in meetings”. In: *Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)* (2006), pp. 143–143.

[47] M. Li et al. “Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection”. In: *International Conference on Pattern Recognition (ICPR)* (2008), pp. 1–4.

[48] C. Zeng and H. Ma. “Robust head-shoulder detection by pca-based multilevel hog-lbp detector for people counting”. In: *International Conference on Pattern Recognition (ICPR)* (2010), 2069–2072.

[49] Min Li et al. “Rapid and robust human detection and tracking based on omega-shape features”. In: *International Conference on Image Processing (ICIP)* (2009), pp. 2545–2548.

[50] J. Tu, Ch. Zhang, and P. Hao. “Robust real-time attention-based head-shoulder detection for video surveillance”. In: *International Conference on Image Processing (ICIP)* (2013), pp. 3340–3344.

[51] J. Yu et al. "Improving person detection using synthetic training data". In: *International Conference on Image Processing (ICIP)* (2010), pp. 3477–3480.

[52] B. Wang et al. "Pedestrian detection in highly crowded scenes using "online" dictionary learning for occlusion handling". In: *International Conference on Image Processing (ICIP)* (2014), pp. 2418–2422.

[53] A. Vedaldi and B. Fulkerson. *VLFeat: An Open and Portable Library of Computer Vision Algorithms*. `http://www.vlfeat.org/`. 2008.

[54] P. Dollar. *Piotr's Computer Vision Matlab Toolbox (PMT)*. `http://vision.ucsd.edu/~pdollar/toolbox/doc/index.html`.

[55] R. Benenson et al. "Seeking the strongest rigid detector". In: *Conference on Computer Vision and Pattern Recognition (CVPR)* (2013).

[56] C. Wojek and B. Schiele. "A performance evaluation of single and multi-feature people detection". In: *German Association for Pattern Recognition (DAGM)* (2008).

[57] J. Lim, C. L. Zitnick, and P. Dollar. "Sketch tokens: A learned mid-level representation for contour and object detection". In: *Conference on Computer Vision and Pattern Recognition (CVPR)* (2013).

[58] S. Paisitkriangkrai, C. Shen, and A. Van den Hengel. "Efficient pedestrian detection by directly optimize the partial area under the roc curve". In: *International Conference on Computer Vision (ICCV)* (2013).

[59] J. Marin et al. "Random forests of local experts for pedestrian detection". In: *International Conference on Computer Vision (ICCV)* (2013).

[60] A.D. Costea and S. Nedevschi. "Word channel based multiscale pedestrian detection without image resizing and using only one classifier". In: *Conference on Computer Vision and Pattern Recognition (CVPR)* (2014).

[61] S. Shalev-Shwartz and T. Zhang. "Stochastic Dual Coordinate Ascent Methods for Regularized Loss Minimization". In: *Journal of Machine Learning Research (JMLR)* (2013), pp. 437–469.

[62] R. Appel et al. "Quickly Boosting Decision Trees - Pruning Underachieving Features Early". In: *International Conference on Machine Learning (ICML)* (2013), pp. 594–602.

[63] R. Fisher. *CAVIAR: Context Aware Vision using Image-based Active Recognition*. `http://homepages.inf.ed.ac.uk/rbf/CAVIAR/`.

[64] A. Milan. *Ground Truth Annotations*. `http://www.milanton.de/data/`.

[65] M. Everingham et al. "The PASCAL visual object classes (VOC) challenge". In: *International Journal of Computer Vision (IJCV)* 88 (2010), 303–338.

Appendix A

Contents of the enclosed DVD

directory	content
thesis	This thesis in PDF
code	Matlab code for the proposed detectors and testing
videos	videos with detection results

Table 6 Content of the attached DVD

Appendix B

Parameter settings of person detectors

Test results for parameter settings of person detectors.

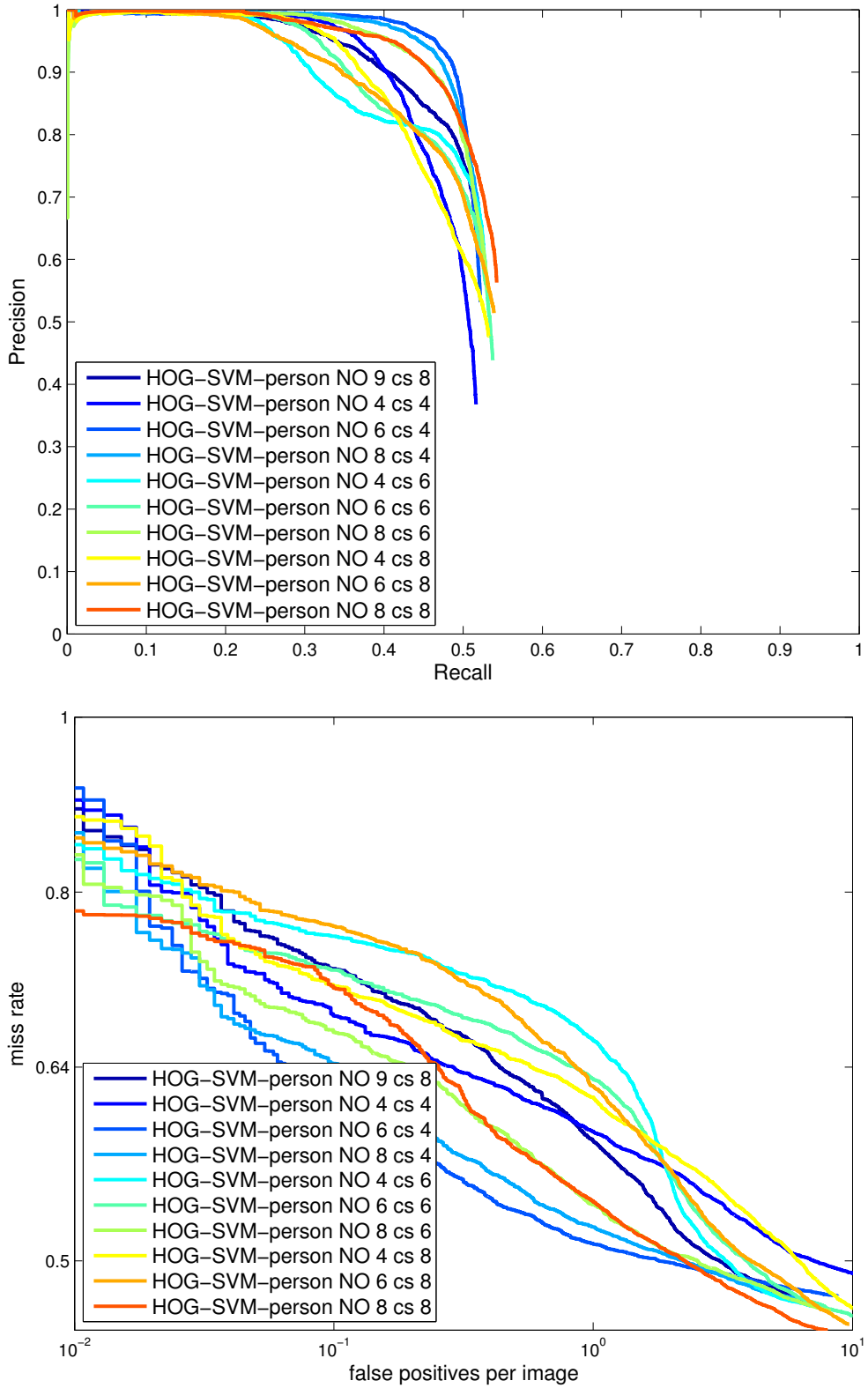


Figure 32 Precision-recall (top) and miss rate against false positives per image (bottom) curves for HOG+SVM person detector with different parameter settings

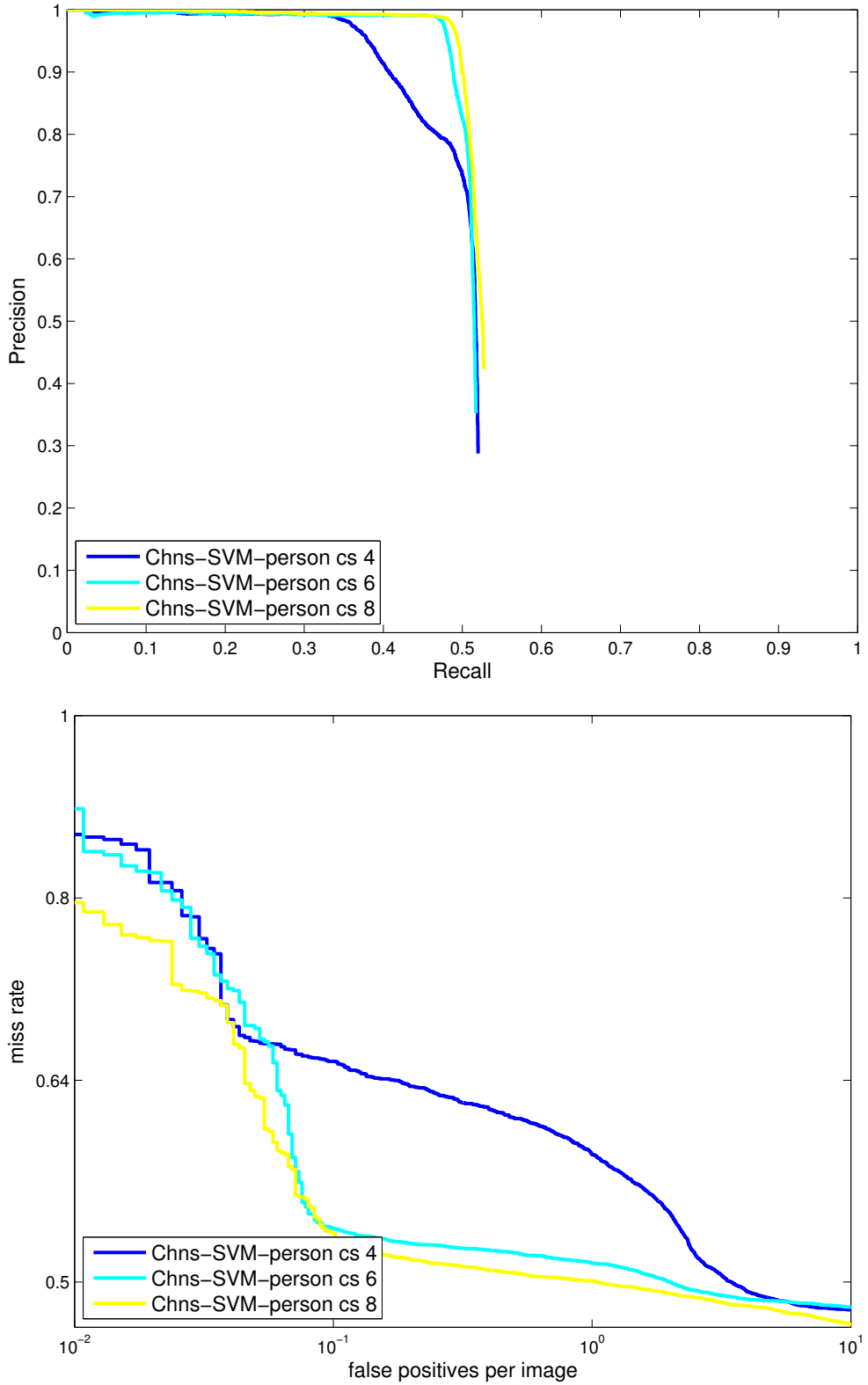


Figure 33 Precision-recall (top) and miss rate against false positives per image (bottom) curves for Chns+SVM person detector with different parameter settings

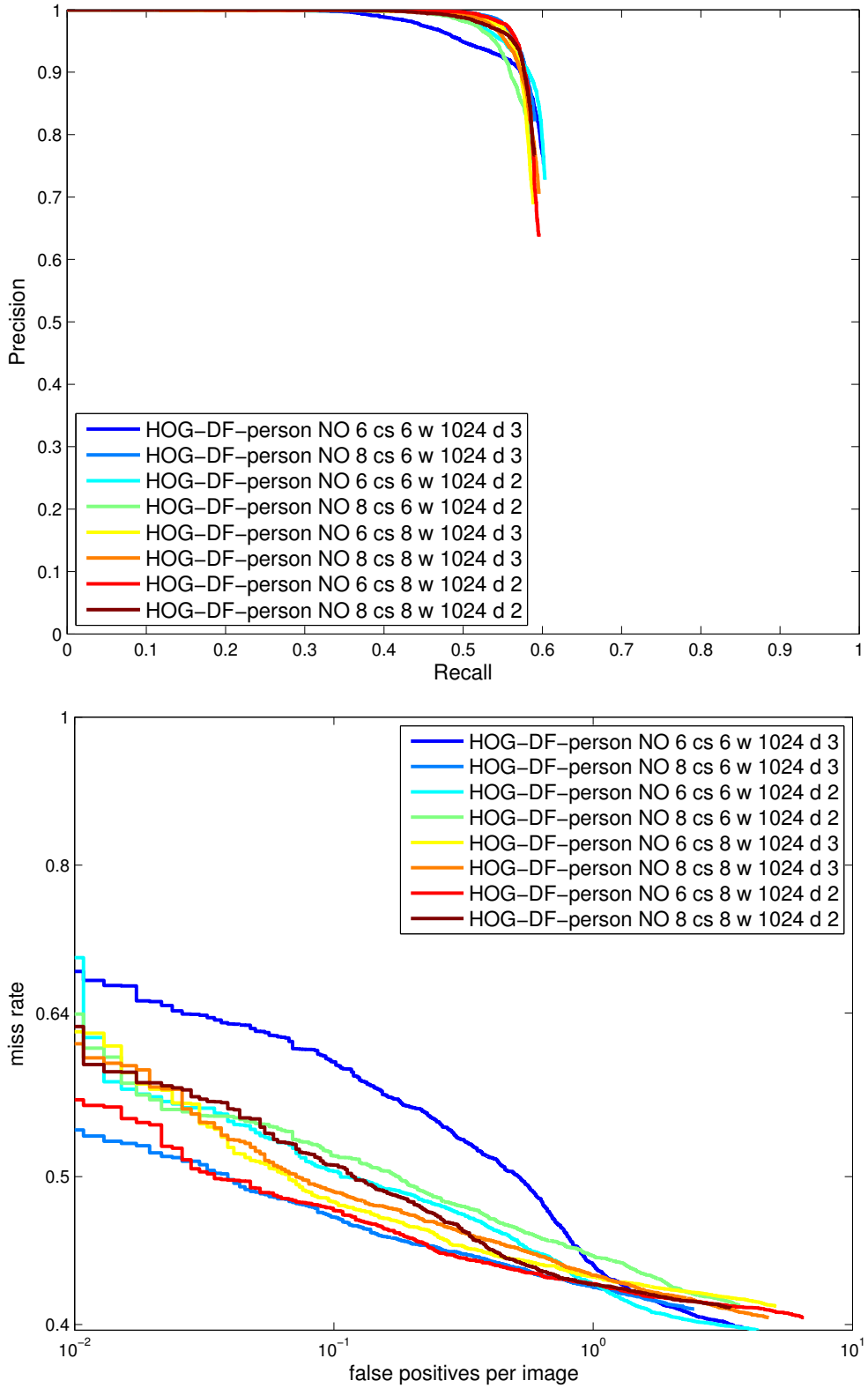


Figure 34 Precision-recall (top) and miss rate against false positives per image (bottom) curves for HOG+DF person detector with different parameter settings

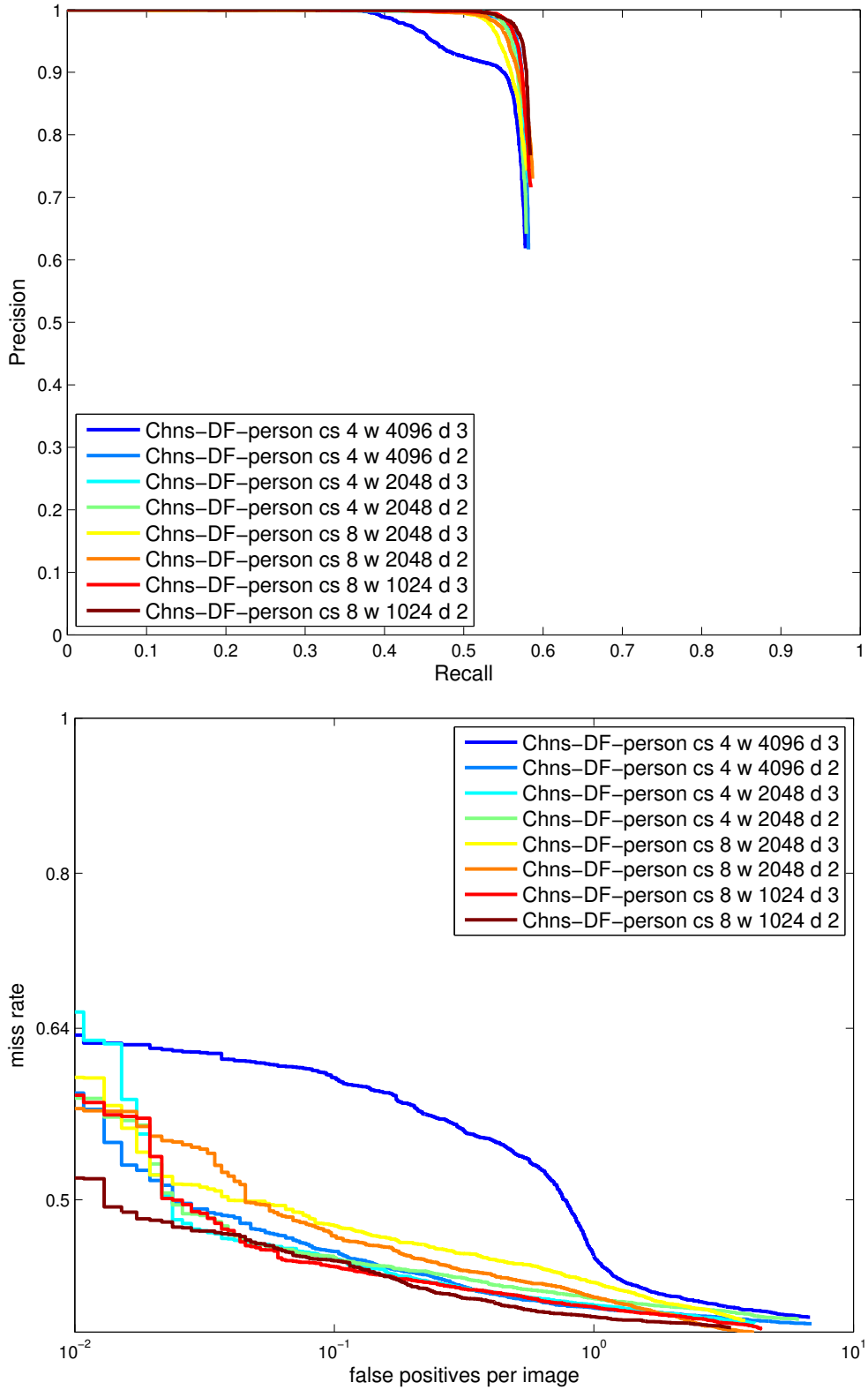


Figure 35 Precision-recall (top) and miss rate against false positives per image (bottom) curves for Chns+DF person detector with different parameter settings