Late Fusion for Person Detection in Camera Networks

Martin Hofmann, Martin Kiechle, Gerhard Rigoll
Institute for Human-Machine Communication
Technische Universität München
martin.hofmann@tum.de, martin.kiechle@mytum.de, rigoll@tum.de

Abstract

In this paper, we present a novel method to detect multiple partially occluded persons in multi-view camera networks. We present a new fusion scheme to integrate the output of part-based object detectors from multiple camera views. This is achieved using subtle and precise modeling of detection and projection uncertainties as well as a fusion method based on probabilistic kernel density estimation. Using a multi-view setup also allows to incorporate additional real-world prior knowledge about person appearances, which not only speeds up processing, but also increases detection rates. Experiments show that this multi-camera approach outperforms methods based on a single perspective, particularly in occlusion-intense scenarios.

1. Introduction

Detecting people in images is a key task of computer vision. There exists a multitude of applications ranging from smart rooms, driver assistance systems to private security and automated surveillance systems. For all these applications, using robust software algorithms to automatically find and locate people in images has high potential to greatly improve surveillance strategies.

Robustly detecting people in these scenarios is however still a challenging task and research in this field is far from being completed. Varying body poses, different clothing and variations in appearance impose great challenges. Most of all, full or partial occlusions remain a major issue which is still difficult to handle with current algorithms.

Using a network of cameras and thus integrating information from multiple views from different angles can greatly improve recognition quality. Therefore, the motivation behind this work is to suggest an approach for person detection by smart fusion of all available distributed data sources.

The proposed fusion scheme makes use of the early stage output from an arbitrary part based person detection method, which gives the likelihood of a person detection for each position. In our experiments we make use of the widely known histogram of oriented gradient detector (HOG detector), but our method is not limited to this particular detector. Detection distributions for each camera are integrated to a common coordinate system, where kernel density estimation is used for information fusion. The central idea of this approach is the fact that non-maximum suppression is delayed to the latest possible stage, where all information is available in a common domain. It has turned out that in order to postpone the decision making, a precise modeling and propagation of both detection and projection uncertainties is necessary. In order to manage the masses of data, we use kernel density estimation to reduce large data distributions to a few sampling points.

2. Related Work

Previously, a multitude of different approaches for monocular person detection have been presented (see [10] for a recent summary). Global features like edge templates as well as local features like Haar wavelets and Sift-like orientation features (e.g. HOG) have been suggested and experimentally evaluated.

It has been shown several times [9, 10] that for monocular pedestrian detection, HOG descriptors outperform other methods in many scenarios.

Various approaches specifically address the occlusion problem by representing human appearance as a combination of multiple descriptors for different body parts or from different views [12, 14, 15].

The use of a multi-camera system has been proposed in several people tracking methods in the past to solve occlusion issues [2, 8, 11].

However these approaches are mostly based on blob detection and data fusion with a Bayesian framework. By contrast, this paper makes use of more elaborate HOG body part detections instead of blob detections. Data from various sources is gathered using kernel-density based fusion, which contrasts the usual Bayesian approaches. This approach greatly improves the detection results in occlusion-intense scenarios.

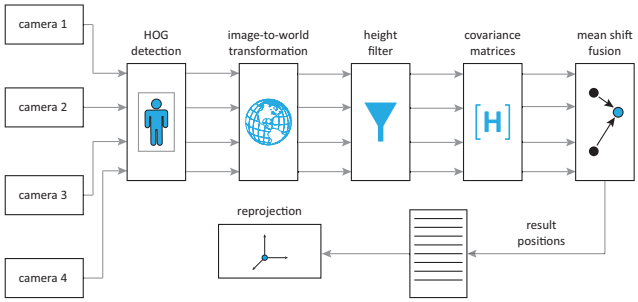


Figure 1: Overview of the MCMP feature extraction and pedestrian detection chain.

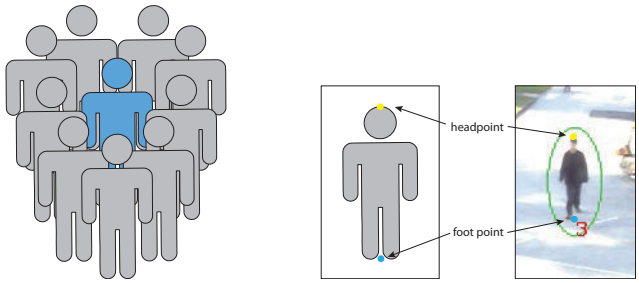


Figure 2: (a) in crowded scenes, part detectors become necessary (b) 'foot point' and 'head point' are relative positions within the detection window

3. The Multi-Camera Multi-Person Detector

Human body detectors, such as the HOG detector [7] perform very well when detecting fully visible pedestrians. However, using images captured by a network of multiple cameras can help to overcome limitations such as full or partial occlusions. In order to achieve a performance gain, however, it turns out to be extremely important to precisely model all relevant uncertainties, most notably detection and projection uncertainties and to have a well designed fusion scheme.

To this end, the following sections describe the concept of our Multi-Camera Multi-Person (MCMP) Detector.

3.1. System Overview

Figure 1 depicts the overview of the MCMP detector. The processing chain consists of part based detection, projection, modeling of uncertainties, filtering and fusion, detailed as follows:

(A) In a first step, an arbitrary person detection method (we use HOG) is used to generate a likelihood distribution (before non-maximum suppression) of person existence for each position in each camera. Part detectors are used to improve detection of occluded people. (B) Taking detection errors and projection errors into account, these likelihood maps are projected into a common world coordinate system. (C) Based on prior knowledge of camera location and orientation as well as on maximum person height assumptions, detections are pruned. (D) Detection and projection errors are captured in covariance matrices (E) Mean Shift on kernel density estimtion is performed to optimally fuse all detections and to get final detection results in world-coordinates. (F) Results are gathered and reprojected for visualization.

3.2. Part Based Body Detection

Standard person detectors have difficulties in crowded scenes due to occlusions as illustrated in Figure 2a. Using multiple cameras can leverage the problem to a certain degree. In addition, human part detectors become necessary for good recognition in crowded scenes.

Typically, surveillance cameras look down on the scene from an elevated position. Consequently, at least the head and shoulder part is visible. We therefore use a full body detector which has high detection rate and low false positive rate and in addition a head shoulder detector, which has a worse false positive rate, but helps in dense scenarios.

It is important to note that the standard HOG detector also includes a method for non-maximum suppression in image coordinates. We deliberately avoid this early non-maximum suppression step and instead only make use of the intermediate SVM output from the HOG detector. This way all (uncertain) information is preserved and non-maximum suppression in our approach is postponed to the fusion module (see Section 3.4).

For each position (x,y) and scale, HOG calculates the corresponding SVM score. We dramatically reduce the huge amount of data by only using points with a score grater than a rough threshold. This is a valid approach, because only a relatively small percentage of the detection windows will contain positive detections. Thus, the intermediate output consists of a relatively low number of x-y-scale-score quadruples.

3.3. Transformation from 2D to 3D and Filtering

The part based detection method described in the previous section results in multiple detections in each camera view. In order to locate pedestrians in the 3D space, a (non-linear) homography mapping based on the Tsai calibration [13] is used.

It can be assumed that people are standing on the ground plane. Since in the training data (INRIA full body), persons are very well alligned with the detection bounding boxes, we can assume that the 'foot point' of each detection is always located at a certain percentage relative to the detection window (see Figure 2b). The same is valid for the 'head

point'.

With this information, for each HOG detection quadruple x-y-scale-score (in image coordinates), a corresponding X-Y-height-score quadruple (in world coordinates) is generated.

These quadruples contain world-coordinate dimensions. Thus, false detections which occasionally occur on higher scale levels (see Figure 4) due to human-like shapes formed by the background, can be efficiently filtered out using a-priori assumptions on the person height.

3.4. Data Fusion by Kernel Density Estimation

The problem of finding the location of pedestrians as a maximum of several body part detections in a 3-D space is cast into a kernel density estimation problem. A mean shift mode detection procedure [4] is used and adjusted to the given requirements.

Mean Shift approach

The input to the fusion stage are X-Y-height-score quadruples, generated from various views, transformed to world coordinates and pruned as described in the previous sections. Let $\mathbf{z}_{i,j} = (X, Y)_{i,j}$, $i = 1 \dots l$, $j = 1 \dots n_i$ denote the coordinates of the j-th point in the i-th camera, and $w_{i,j}$ the corresponding SVM score weight. There is a total of $n = \sum_{i=1}^{l} n_i$ detection points. n_i detection points from the i-th camera. Also, let $\mathbf{H}_{i,j}$ be the 2×2 covariance matrices associated with each respective detection point $\mathbf{z}_{i,j}$ in which the bandwidth parameters can be adapted for every single detection point (see below). Note that in $\mathbf{z}_{i,j}$, the height is omitted. This greatly reduces the search space and is reasonable, assuming upright standing persons.

The detection goal is to find the modes of the distribution $\widehat{f}(\mathbf{z})$, which is approximated using kernel density estimation. The density estimate at a point $\mathbf{z}$ can be formulated as (c.f. [3, 4])

$$\widehat{f}(\mathbf{z}) = \frac{1}{n(2\pi)^{\frac{3}{2}}} \sum_{i=1}^{l}\sum_{j=1}^{n_i} |\mathbf{H}_{i,j}|^{-\frac{1}{2}} \; t(w_{i,j}) \; \exp\left(-\frac{D^2\left[\mathbf{z}, \mathbf{z}_{i,j}, \mathbf{H}_{i,j}\right]}{2}\right) \quad (1)$$

where $D^2\left[\mathbf{z}, \mathbf{z}_{i,j}, \mathbf{H}_{i,j}\right] \equiv \left(\mathbf{z} - \mathbf{z}_{i,j}\right)^\top \mathbf{H}_{i,j}^{-1} \left(\mathbf{z} - \mathbf{z}_{i,j}\right)$ is the squared Mahalanobis distance between $\mathbf{z}$ and $\mathbf{z}_{i,j}$. The term $t(w_{i,j})$ represents the clipped classification score of the detection point $\mathbf{z}_{i,j}$ returned from the SVM. The soft clipping function $t(w_{i,j}) = a^{-1} \log(1 + \exp(a\,(w_{i,j} + c))$ with parameters $a = 10$ and $c = 0$ as suggested by Dalal [6, p. 60] is used to provide the mean shift with the required positive values $t(w_{i,j}) > 0, \forall w_{i,j}$. The gradient of (1) then becomes the following

$$\begin{aligned}
\nabla\widehat{f}(\mathbf{z}) &= \frac{1}{n(2\pi)^{\frac{3}{2}}} \sum_{i=1}^{l}\sum_{j=1}^{n_i} |\mathbf{H}_{i,j}|^{-\frac{1}{2}} \; \mathbf{H}_{i,j}^{-1} \left(\mathbf{z}_{i,j} - \mathbf{z}\right) \\
&\qquad t(w_{i,j}) \; \exp\left(-\frac{D^2\left[\mathbf{z}, \mathbf{z}_{i,j}, \mathbf{H}_{i,j}\right]}{2}\right) \\
&= \frac{1}{n(2\pi)^{\frac{3}{2}}} \left[\sum_{i=1}^{l}\sum_{j=1}^{n_i} |\mathbf{H}_{i,j}|^{-\frac{1}{2}} \; \mathbf{H}_{i,j}^{-1} \mathbf{z}_{i,j} \right. \\
&\qquad t(w_{i,j}) \; \exp\left(-\frac{D^2\left[\mathbf{z}, \mathbf{z}_{i,j}, \mathbf{H}_{i,j}\right]}{2}\right) \\
&\qquad \left. - \left\{\sum_{i=1}^{l}\sum_{j=1}^{n_i} |\mathbf{H}_{i,j}|^{-\frac{1}{2}} \; \mathbf{H}_{i,j}^{-1} \right.\right. \\
&\qquad \left.\left. t(w_{i,j}) \; \exp\left(-\frac{D^2\left[\mathbf{z}, \mathbf{z}_{i,j}, \mathbf{H}_{i,j}\right]}{2}\right)\right\} \mathbf{z}\right] \quad (2)
\end{aligned}$$

If the weights $\varpi_{i,j}$ are defined as

$$\varpi_{i,j}(\mathbf{z}) = \frac{\sum_{i=1}^{l}\sum_{j=1}^{n_i} |\mathbf{H}_{i,j}|^{-\frac{1}{2}} \; t(w_{i,j}) \; \exp\left(-\frac{D^2\left[\mathbf{z}, \mathbf{z}_{i,j}, \mathbf{H}_{i,j}\right]}{2}\right)}{\sum_{i=1}^{l}\sum_{j=1}^{n_i} |\mathbf{H}_{i,j}|^{-\frac{1}{2}} \; t(w_{i,j}) \; \exp\left(-\frac{D^2\left[\mathbf{z}, \mathbf{z}_{i,j}, \mathbf{H}_{i,j}\right]}{2}\right)} \quad (3)$$

and $\sum_{i=1}^{l}\sum_{j=1}^{n_i} \varpi_{i,j} = 1$ then by dividing (2) by (1) using (3), the result is

$$\frac{\nabla\widehat{f}(\mathbf{z})}{\widehat{f}(\mathbf{z})} = \sum_{i=1}^{l}\sum_{j=1}^{n_i} \varpi_{i,j}(\mathbf{z}) \; \mathbf{H}_{i,j}^{-1} \mathbf{z}_{i,j} - \left(\sum_{i=1}^{l}\sum_{j=1}^{n_i} \varpi_{i,j}(\mathbf{z}) \; \mathbf{H}_{i,j}^{-1}\right) \mathbf{z} \quad (4)$$

Let

$$\mathbf{H}_h^{-1}(\mathbf{z}) = \sum_{i=1}^{l}\sum_{j=1}^{n_i} \varpi_{i,j}(\mathbf{z}) \; \mathbf{H}_{i,j}^{-1} \quad (5)$$

be the harmonic mean of the covariance matrices $\mathbf{H}_{i,j}$ computed at $\mathbf{z}$ and weighted by all detections. The variable bandwidth mean shift vector is then defined by (4) and (5) as

$$\mathbf{m}(\mathbf{z}) = \mathbf{H}_h \frac{\nabla\widehat{f}(\mathbf{z})}{\widehat{f}(\mathbf{z})} \equiv \mathbf{H}_h(\mathbf{z}) \left[\sum_{i=1}^{l}\sum_{j=1}^{n_i} \varpi_{i,j}(\mathbf{z}) \; \mathbf{H}_{i,j}^{-1} \mathbf{z}_{i,j}\right] - \mathbf{z} \quad (6)$$

The gradient becomes zero $\nabla\widehat{f}(\mathbf{z}) = 0$ at the mode location, implying $\mathbf{m}(\mathbf{z}) = 0$. Thus, the mode can be iteratively estimated for each detection point by starting from any $\mathbf{z}_{i,j}$ until the mean shift vector is zero $\mathbf{m}(\mathbf{z}) = 0$. In practice, the iteration will be terminated if the Euclidean norm of the mean shift vector is smaller than a certain threshold $|\mathbf{m}(\mathbf{z})| < \varepsilon$. In that case, $\mathbf{z}$ represents the wanted detection position with x-y coordinates and the final detection score. For further details on the theory of the mean shift procedure consult [3, 4].

Covariance Matrices $\mathbf{H}_{i,j}$

The uncertainties involved in the processing chain mainly originate from the following three sources: (1) Due to the camera angle and due to pixel quantization, areas in the background will have higher uncertainty than foreground regions. (2) The camera calibration is imprecise due to approximations in the camera model as well as potentially imprecise annotations of correspondence points. (3) The 'foot points' and 'head points' within the HOG detection window are uncertain because of variance in the alignment of the training data.

Thus, the covariance matrices $\mathbf{H}_{i,j}$ are a key parameter for the mean shift algorithm and model the occurring uncertainties. Because the uncertainties are scaled depending on the position and orientation between the detected object and the camera, the smoothing values of the covariance matrices are obtained from the individual detection position itself. In order to achieve this, besides the main detection position $\mathbf{p}_{i,j}$ in the image, e.g. the foot point, q further samples of image positions $\mathbf{p}_{i,j\,k}$, $k = 1 \dots q$ with the distinct distance vectors $\mathbf{d}_{i,j\,k}$ to the main detection position are transformed to the world space, resulting in the transformed positions $\mathbf{p}'_{i,j}$, $\mathbf{p}'_{i,j\,k}$ and the transformed distance vectors $\mathbf{d}'_{i,j\,k}$. Because the transformed positions in world coordinates are assumed to be on the ground plane with $z = 0$, $\mathbf{d}'_{i,j\,k}$ are 2-D vectors omitting the z-coordinate. The positive semidefinite diagonal covariance matrices $\mathbf{H}_{i,j}$ is then computed as

$$\mathbf{H}_{i,j} = \frac{1}{q} \left[\mathbf{d}'_{i,j\,1}, \mathbf{d}'_{i,j\,2}, \dots, \mathbf{d}'_{i,j\,q}\right] \left[\mathbf{d}'_{i,j\,1}, \mathbf{d}'_{i,j\,2}, \dots, \mathbf{d}'_{i,j\,q}\right]^\top \tag{7}$$

where q represents the number of additional samples. In order to invert $\mathbf{H}_{i,j}$ (see Equation (3.4)), $\mathbf{H}_{i,j}$ is required to be positive definite. To achieve this, the distance vectors must be chosen $|\mathbf{d}_{i,j\,k}| > 0$ which yields $|\mathbf{d}'_{i,j\,k}| > 0$, resulting in positive definite covariance matrices $\mathbf{H}_{i,j}$. Even though it is possible to choose a large number of additional samples to compute $\mathbf{H}_{i,j}$ in (7), we use only four additional samples $q = 4$. As depicted in Figure 3, the four positions centered on the main detection position are chosen. The Euclidean norm of the distance vectors between the main detection position and the samples to the left and the right is defined as σ_x and for the norm of the distance to the top and bottom samples as σ_y.

4. Evaluation and Results

4.1. Datasets

The experiments on the MCMP detector are carried out on the PETS 2009 benchmark data set [1]. To evaluate the performance on scenes with crowds of different density, the data sets S2_L1 (sparse), S2_L2 (medium dense) and S2_L3 (dense) have been selected. For S2_L2 and S2_L3 there are four perspectives (view001-004) of the scene available whereas S2_L1 could only be evaluated using three camera sources (view002 not available).

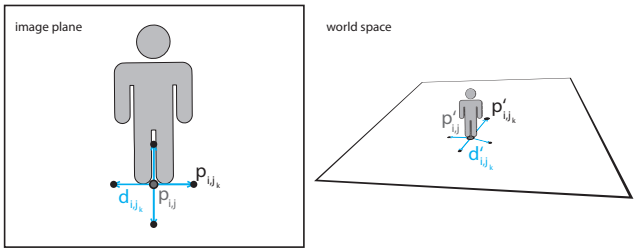


Figure 3: The transformed distance vectors $d'_{i,j\,k}$ between the main detection position $p'_{i,j}$ and further sample points $p'_{i,j\,k}$ are used to compute the covariance matrices $\mathbf{H}_{i,j}$

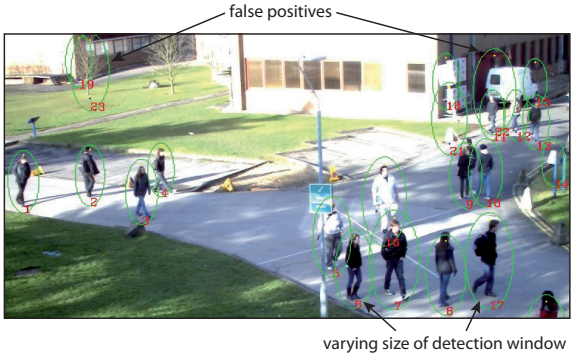


Figure 4: Additional knowledge about the expected detection result can be used to filter false detections like the tree.

4.2. Evaluation Methodology

First, the ground truth annotations were obtained using a multi-camera annotation tool. This allowed precise annotation of the feet positions of each person on the ground plane. To minimize dataset specific influence on the detection result caused by inaccurate camera calibration, areas with obvious uneven terrain or which cannot be seen by at least two cameras are excluded from evaluation.

For each of the three scenarios with sparse, medium dense and dense crowds, twenty different frames distributed over the entire dataset are randomly chosen for evaluation and their results averaged. Detection and annotation data are matched using a greedy, truncated matching algorithm based on euclidean distance. Starting from the best match, successively all pairs are matched until a tolerance threshold of $d_{\text{max}} = 60$cm is reached. Beyond this threshold remaining detections are considered false positives and the remaining ground truth objects are considered misses.

4.3. Overall Results

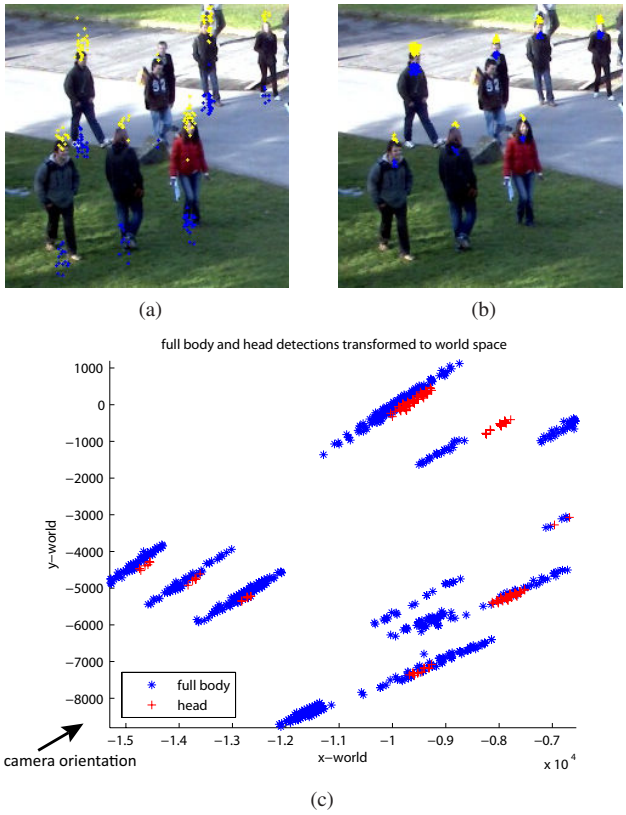
The performance of the MCMP detector is compared to the original HOG detector by Dalal and Triggs using the INRIA Object Detection and Localization Toolkit [5]. Both detectors are initialized with the same parameters for scanning and extracting feature vectors from input images.

As expected, the HOG detector - which is optimized for fully visible human appearances - is difficult to beat in low density scenarios (S2_L1). The MCMP detector can still reach a slightly lower miss rate than the HOG detector due to a few inter-object occlusions (see Figure 5a).

The more occlusions occur in a single perspective, the bigger the advantage of multi-camera analysis. The experiments on the medium dense crowd scenario S2_L1 show greatly increased detection performance of the MCMP detector (Figure 5b). In the dense crowd scenario, the HOG detector cannot reach a miss rate of 50% due to a large number of partially occluded pedestrians. The MCMP detector clearly outperforms the single camera approach in such scenarios (Figure 5c).

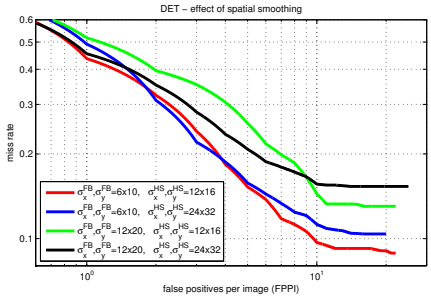
4.4. Performance

Several experiments have shown that certain parameters influence the detection results achieved with the MCMP detector to a large extent. Because σ_x and σ_y account for the magnitude of spatial smoothing of each detection position, they are key parameters for modeling the occurring uncertainties. If the values are too small, the uncertainty of all detections leads to many local maxima resulting in an increased number of false detections. If σ_x and σ_y are set too high, detections of different people standing closely together get fused, which increases the miss rate. Because the results of part detectors such as the head detector are additinally affected by the body height uncertainty, the spatial smoothing values σ_x and σ_y for part detectors have to be larger than for parts with a known detection point, e.g. the foot point for the full body detector. Figure (6) illustrates a typical HOG detection result. Obviously, the found foot and head point predominantly scatter along the optical axis between camera and objects (which corresponds to the y-axis in image coordinates), whereas the variance along the orthogonal axis (the x-axis) is rather small. This means that σ_y has to be chosen larger than σ_x to account for the greater uncertainty. For the used datasets, a factor of two gives optimal results.



(a) (b) (c)

Figure 6: HOG detection results of the full body (a) and the head detector (b) (upper point: yellow, lower point: blue). In world coordinates, (c) the main detection points predominantly scatter along the optical axis between camera and objects.

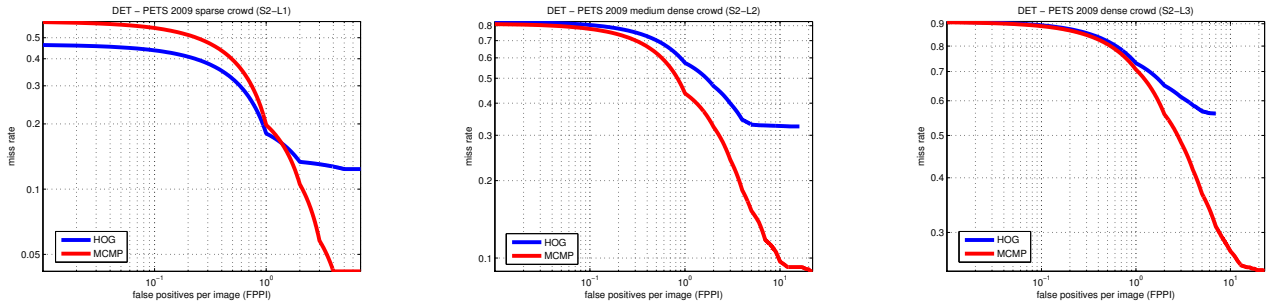


(a)

Figure 7: Effect of different values for spatial smoothing on the performance

5. Conclusion

In this work we have investigated an approach to extend single-view person detection to joint detection in a multi-camera network. Adding information from additional sensors naturally results in more information. However, it turns out to be a challenging task to bring this additional data together. In this work we have presented a precise modeling of known and estimated uncertainties and we have used a joint fusion and detection scheme based on kernel density estimation to make use of all available information. The most significant improvements were achieved in medium



(a) sparse crowd scenario (b) medium dense crowd scenario (c) dense crowd scenario

Figure 5: Performance comparison on the PETS 2009 dataset between the HOG and the MCMP detector in sparse (a), medium dense (b) and dense crowd (c) scenarios.

and densely crowded scenes. Using multiple part detectors and the described method of fusing detection results in a common world space helps handling full or partial inter-object occlusion to a great extent.

6. Outlook

Fusing information from multiple cameras is a promising approach, especially since future smart environments are likely to make intense use of multiple multi modal sensors.

The MCMP detector uses multiple part detections to find the most probable detection position. In the present version, the detection result from the head part detector is projected onto the ground plane and only its x-y position is used to accumulate the probability of a local maximum of detections on the ground plane. An improvement could be achieved by additionally taking the orientation and distance of multiple part detections towards each other into account. This would require either more complex kernels for the mean shift procedure or a preprocessing of detections with an alternative method.

In this work we have restricted our research to single frame detection. Of course, tracking and associating pedestrians over time would be a fundamental extension to the presented detector.

References

[1] PETS 2009 Benchmark Dataset. http://www.cvg.rdg.ac.uk/PETS2009/. 4

[2] T. Chang and S. Gong. Tracking Multiple People With a Multi-Camera System. *Proceedings, IEEE Workshop on Multi-Object Tracking*, pages 19–26, 2001. 1

[3] D. Comaniciu. An Algorithm for Data-Driven Bandwidth Selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(2):281–288, Feb. 2003. 3

[4] D. Comaniciu. Nonparametric Information Fusion For Motion Estimation. *IEEE Computer Society Conference on Computer*, 2003. 3

[5] N. Dalal. INRIA Object Detection and Localization Toolkit. http://pascal.inrialpes.fr/soft/olt/. 5

[6] N. Dalal. *Finding People in Images and Videos*. PhD thesis, Institut National Polytechnique De Grenoble, 2006. 3

[7] N. Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 886–893, 2005. 2

[8] S. Dockstader and A. Tekalp. Multiple Camera Fusion for Multi-Object Tracking. *Multi-Object Tracking, Proceedings IEEE Workshop on*, pages 95–102, 2001. 1

[9] P. Dollar, C. Wojek, B. Schiele, and P. Perona. Pedestrian Detection: A benchmark. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 304–311, June 2009. 1

[10] M. Enzweiler and D. Gavrila. Monocular Pedestrian Detection: Survey and Experiments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(12):2179–95, Dec. 2009. 1

[11] J. Kang, I. Cohen, and G. Medioni. Multi-Views Tracking Within and Across Uncalibrated Camera Streams. *First ACM SIGMM International Workshop on Video Surveillance*, page 21, 2003. 1

[12] B. Leibe, E. Seemann, and B. Schiele. Pedestrian Detection in Crowded Scenes. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 878–885, 2005. 1

[13] R. Y. Tsai. An Efficient and Accurate Camera Calibration Technique for 3D Machine Vision. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 364–374, 1986. 2

[14] B. Wu and R. Nevatia. Optimizing Discrimination-efficiency Tradeoff in Integrating Heterogeneous Local Features for Object Detection. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008. 1

[15] J. Xing, H. Ai, and S. Lao. Multiple Human Tracking Based on Multi-view Upper-Body Detection and Discriminative Learning. *20th International Conference on Pattern Recognition*, pages 1698–1701, Aug. 2010. 1