Monocular 3D Scene Understanding with Explicit Occlusion Reasoning

Christian Wojek[1] Stefan Walk[2] Stefan Roth[2] Bernt Schiele[1]

[1]MPI Informatics
Saarbrücken, Germany

[2]Department of Computer Science
TU Darmstadt, Germany

Abstract

Scene understanding from a monocular, moving camera is a challenging problem with a number of applications including robotics and automotive safety. While recent systems have shown that this is best accomplished with a 3D scene model, handling of partial object occlusion is still unsatisfactory. In this paper we propose an approach that tightly integrates monocular 3D scene tracking-by-detection with explicit object-object occlusion reasoning. Full object and object part detectors are combined in a mixture of experts based on their expected visibility, which is obtained from the 3D scene model. For the difficult case of multi-people tracking, we demonstrate that our approach yields more robust detection and tracking of partially visible pedestrians, even when they are occluded over long periods of time. Our approach is evaluated on two challenging sequences recorded from a moving camera in busy pedestrian zones and outperforms several state-of-the-art approaches.

1. Introduction

The goal of this paper is to enable reliable multi-object tracking from a moving platform in challenging real-world scenes (see, *e.g.*, Fig. 1) even in cases when the objects are partially occluded for extended periods of time. Though by no means limiting the applicability, we focus on multi-people tracking, which is particularly challenging due to the large variability of human pose and appearance. The impressive progress in human detection and long-term tracking has allowed to detect and track several people simultaneously in complex scenes. Yet, state-of-the-art systems are still severely challenged by partial and full occlusions, which occur frequently in scenes of realistic complexity.

Typical multi-people tracking systems employ a Bayesian approach that relies on the robustness of both the human detection model and the tracking module. Without any explicit occlusion model such approaches have shown some robustness w.r.t. partial occlusions [1, 6, 10].

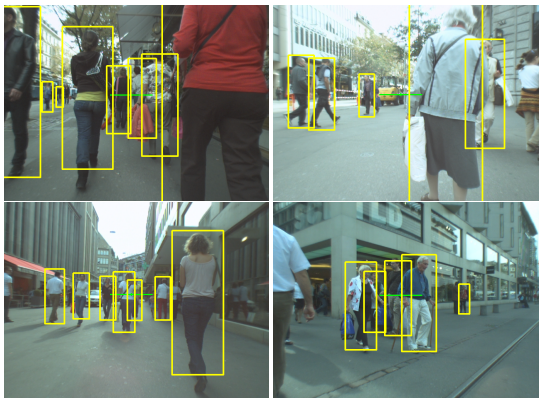


Figure 1: Scene tracklets obtained with our explicit 3D occlusion model.

Elaborate association schemes have been proposed to enable recovery from partial and even full occlusions [7, 11, 14, 24]. However, these approaches are limited by the ability of their respective human detection model to detect and re-detect people before and after the occlusion, which limits their applicability to cases where people are sufficiently visible before and after the occlusion. In contrast, we explicitly address the problem of detecting and tracking people even when they are never fully visible or when they are significantly occluded over long periods of time. Fig. 1 gives several examples in which our system is able to find and track people that are partially occluded over extended periods.

Drawing on successful prior work, we propose a new approach for multi-people tracking in the presence of challenging occlusions. The first important component is to track the complete scene rather than an assembly of individuals. This idea has been shown to enable robust 2D tracking of multiple people in surveillance scenarios [12, 17]. We adopt and extend this idea to *3D scene tracking using a monocular camera* [2, 22]. This is in contrast to other 3D tracking work [6] that uses stereo camera setups, yet is outperformed by our monocular system (see Sec. 4). In order to enable detection and tracking of people even when they are never fully visible, we directly extend successful detection approaches such as HOG [3] and DPM [8] to enable the detection of partially visible humans for a variety

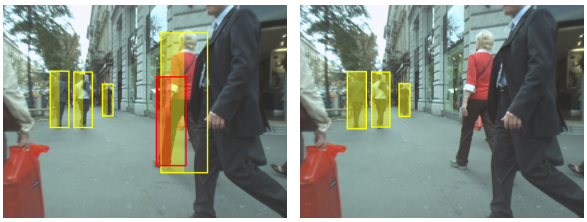
To appear in the Proc. of the *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 1993–2000, Colorado Springs, Colorado, June 2011.

of visibility scopes (see Fig. 3) and integrate them into our 3D scene model. Having a full 3D scene model allows us to determine the visibility of each individual in the scene, which in turn enables to predict which parts of the body are sufficiently visible and thus detectable. This allows us to define a novel complete 3D scene likelihood that tightly integrates full and partial human detectors within a 3D scene tracking framework. Quantitative experiments on publicly available data demonstrate that our model outperforms previous approaches and allows to associate and track people even in the presence of long-term partial occlusions.

Related work. Partial and full occlusions are a major error source for tracking and have been addressed in a variety of ways. Thus, reviewing the complete state-of-the-art is beyond the scope of this paper. Nonetheless, we note that relatively little work has attempted to tightly integrate explicit occlusion and scene modeling in a tracking-by-detection framework, *e.g.*, through elaborate matching schemes.

To deal with partial occlusion, Isard & MacCormick [12] integrated human detection and tracking more tightly and proposed a complete 2D scene likelihood for tracking humans in a surveillance setting by using local filter responses as image evidence. Similarly, Shet *et al*. [17] and Wu & Nevatia [24] employed a combination of body part detectors to obtain a 2D scene model from a static surveillance camera. In contrast, the goal of this paper is to understand scenes recorded from a moving monocular camera and to infer a 3D scene model to eventually support autonomous navigation. In this context, Choi & Savarese [2] recently proposed a 3D tracking approach for monocular semi-static cameras. Despite modeling track interaction, explicit occlusion reasoning is not performed. Wojek *et al*. [22] proposed a 3D scene tracking approach for moving monocular cameras, but also do not explicitly take occlusion into account. Other approaches with 3D scene models include work by Ess *et al*. [6] and Gavrila & Munder [10], but both use a stereo camera setup. [7] uses stereo disparity to extend tracks under occlusion using occlusion maps. Similarly, Kaucic *et al*. [14] leverage scene segmentation to predict occlusion. Common to these approaches is that they only allow to temporally extend tracks of fully visible objects, but do not allow to detect and track objects that are only partially visible. As our experiments show, the proposed, tightly integrated scene tracklet model with explicit occlusion handling outperforms these approaches [2, 6, 22], and is able to detect and associate also partially visible objects.

The basic detectors for most tracking-by-detection systems for 3D scene understanding are full-body detectors, which allow to deal with occlusions only when the person is visible enough to be detectable by the respective full-body detector. In many scenes, however, certain people are never fully visible making these approaches prone to fail. To address this, we propose to use multiple partial detectors that allow to also detect and track people that are never fully visible. To our knowledge, the only other work that uses partial detectors for tracking is by Xing *et al*. [25], which uses only two partial and one full detector without a scene model to infer expected occlusion. In contrast, we use six partial detectors in a tightly integrated scene model.



(a) Left-half detector (b) Full-body detector

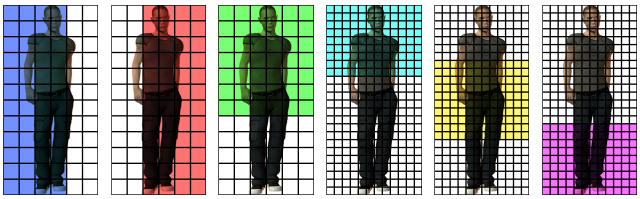
Figure 2: Sample detections for models trained on the left half of a pedestrian and for the full-body detector. While the models for partial views do not perform as well overall, they are able to provide the scene model with hypotheses for partially occluded pedestrians.

There is also an extensive literature on explicit occlusion handling for object detection. Enzweiler *et al*. [4] learned local head, torso and leg detectors combined in a mixture-of-experts framework and leveraged stereo and flow cues. Wang *et al*. [20] performed occlusion handling in a modified SVM framework, while Lin *et al*. [15] adapted a boosting cascade to cope with partial occlusion. Winn & Shotton [21] employed CRFs to couple segmentation and detection under partial occlusion. Finally, Vedaldi & Zisserman [19] employed structured output regression to detect partially truncated and misaligned multi-aspect objects. For human pose estimation, Sigal & Black [18] proposed an approach to model self-occlusion of body-parts. However, none of these approaches models the interaction of different objects. In contrast, the goal of this work is an explicit model for complex object-object occlusions.

2. Detectors

Our system uses seven detector components to provide the detection hypotheses. All components use HOG-like features [3], which have been proven to be a robust and effective feature for pedestrian detection.

The first detector component is the deformable parts model (DPM) by Felzenszwalb *et al*. [8]. It uses a combination of a global HOG template and several higher-resolution templates for parts that are allowed to vary in position relative to the position of the global template. This component performs best for fully visible large-scale pedestrians, but cannot handle small pedestrians and occlusions well. All other detectors are obtained by training an SVM on various parts of the HOG block grid of the detector window. They differ from [3] by an intersection kernel SVM instead of a linear SVM [16], using multiple rounds of retraining to make the training procedure stable [8], and an improved



(a) Left (b) Right (c) Top (d) Top-HR (e) Mid-HR (f) Bot-HR

Figure 3: Detector regions of object part detectors.

non-maximum suppression scheme [23].

The SVM for three of the detectors is trained on the upper, left and right halves of the block grids (rounded up). The upper-body detector uses the top 8×7 blocks, and the left- and right-half body detectors the 15×4 left and right blocks, respectively (for an illustration see Fig. 3(a)–(c)). We also employ three models using a higher resolution detection window (256×128 pixels, resulting in a grid of 31×15 blocks), trained using only rows 3–12 (Top-HR), 11–20 (Mid-HR) and 20–29 (Bot-HR) (see Fig. 3(d)-(f)). These are motivated by the fact that in crowded scenes pedestrians are often quite close to the camera, but it is still desirable to detect them. All seven detectors are trained on the INRIA Person dataset [3].

It is important to note that training an SVM on parts, *i.e.* subsets of blocks, is different from using the model learned for the full body and only evaluating on the "visible" subset (which would be theoretically possible for additive kernels), especially because during the bootstrapping phase of training the detector finds hard samples for the partial-view models instead of hard samples for the full-body model. Even though a bank of detectors increases the computational load, we stress that the low-level feature representation can be shared among detectors and therefore only the classifiers need to be evaluated. To further reduce the load it may also be possible to adapt the DPM formulation to allow a tighter integration of our partial detectors; this will remain future work. Fig. 2 shows detection examples for the left-half and full-body models. The full-body detector (Fig. 2(b)) is good at spotting fully visible pedestrians, but has problems finding pedestrians that are partially occluded. For these cases partial-view detectors can be beneficial, as they can spot partially occluded pedestrians (Fig. 2(a)). However, they typically also produce more false positives (as in Fig. 2(a)). As these tend to be inconsistent with the 3D scene model, our method can discard them.

3. 3D Scene and Occlusion Model

Before being able to introduce our explicit occlusion reasoning scheme based on 3D information, we first describe the basic 3D scene model. The 3D scene model is based on our recent work [22], which aimed to combine image evidence from detectors, geometric constraints and priors, as well as temporal reasoning to infer the 3D position of all objects in a scene from monocular video alone. For simplicity, we follow the notation of [22] and denote image coordinates in lower case, 3D world coordinates in upper case, and other vectors in bold.

For now assuming that only a single frame is given (the frame index is omitted for clarity), we take a Bayesian approach and define the posterior for the 3D scene $\mathbf{X}$ given image evidence $\mathcal{E}$ as

$$P(\mathbf{X}|\mathcal{E}) \propto P(\mathcal{E}|\mathbf{X}) \cdot P(\mathbf{X}), \tag{1}$$

where $P(\mathcal{E}|\mathbf{X})$ describes the observation model and $P(\mathbf{X})$ the prior assumptions about the 3D scene. The state of the 3D scene $\mathbf{X}$ is comprised of the individual objects $\mathbf{O}^i$, whose state is given by their 3D position $(O_x^i, O_y^i, O_z^i)^\top$ relative to the observer and their height H^i. The scene state $\mathbf{X}$ also includes the intrinsic and extrinsic camera parameters $\mathbf{K}$ and $\mathbf{R}$ (rotation only, see [22, Fig. 2]).

Prior. We make the same basic assumptions as in [22], which apply to a variety of robotics and automotive scenarios: We assume that the camera is rigidly mounted to a platform, which along with all objects stands on a common ground plane ($O_z^i = 0$), and has been calibrated off-line. The camera is furthermore assumed to undergo no roll and yaw w.r.t. the platform; odometer readings are used to determine the speed and turn rate of the platform itself. Hence the observer-centric coordinate system is fully specified by the pitch angle Θ, which may vary slightly as the platform accelerates or slows down.

Due to the low viewpoint in the sequences employed in this paper, the correct estimation of distant objects is difficult requiring reliable estimates of the camera pitch. We also aim to avoid detecting background structures that stand on the ground but have incorrect height. To address these issues, we integrate prior knowledge. Specifically, the camera pitch Θ is assumed to follow a Gaussian distribution $\mathcal{N}(\Theta; \mu_\Theta, \sigma_\Theta)$ around the resting pitch μ_Θ. In addition, the height of each scene object H^i is assumed to follow a Gaussian distribution $\mathcal{N}(H^i; \mu_H^{c_i}, \sigma_H^{c_i})$, where $\mu_H^{c_i}$ denotes the mean height of the respective object class c_i. Consequently, the 3D scene prior can be written as

$$P(\mathbf{X}) \propto \mathcal{N}(\Theta; \mu_\Theta, \sigma_\Theta) \cdot \prod_i \mathcal{N}(H^i; \mu_H^{c_i}, \sigma_H^{c_i}). \tag{2}$$

Next, we turn to the observation model $P(\mathcal{E}|\mathbf{X})$. The image evidence $\mathcal{E}$ in this work is comprised of a set of candidate full object detections and a set of candidate object part detections. We will first describe our model for a single full body detector and then extend it to a setup with multiple part detectors. As we will see in the experiments, the combination of different detectors is beneficial for handling partial object-object occlusion as well as object truncation at the image boundary. Full object detectors return more reliable

hypotheses than part detectors, but are limited to entirely visible objects. Frequently the detection confidence drops severely even when the object is only partially occluded or outside the image. Part detectors on the other hand allow to detect objects based on partial appearance, but also tend to produce a higher number of false positive detections.

Single detector likelihood. In case of a single full body detector we define the likelihood following [22] as

$$P(\mathcal{E}|\mathbf{X}) \propto \prod_i \Psi_D\big(\mathbf{d}^{a(i)}\big) \cdot \Psi_G\big(\mathbf{O}^i, \Theta; \mathbf{d}^{a(i)}\big). \quad (3)$$

Herein every 3D object hypothesis $\mathbf{O}^i$ is associated with one of the candidate detections $\mathbf{d}^{a(i)}$ via an association function $a(i)$. The appearance potential Ψ_D maps the detector's appearance score for the associated detection $\mathbf{d}^{a(i)}$ into the positive range. In practice we perform hard clipping of the SVM margin at zero (for negative scores). The potential Ψ_G models geometric constraints imposed by the ground plane, which is governed by the camera pitch Θ. In particular, a Gaussian in x-y-scale-space measures how well the projection of the object $\mathbf{O}^i$ to the ground plane, $\mathbf{o}^i$, matches the associated detection $\mathbf{d}^{a(i)}$:

$$\Psi_G(\mathbf{O}^i, \Theta; \mathbf{d}^{a(i)}) = \mathcal{N}(\mathbf{o}^i; \mathbf{d}^{a(i)}, \boldsymbol{\sigma}_G + \bar{\boldsymbol{\sigma}}_G). \quad (4)$$

The kernel bandwidth is split into a constant component $\boldsymbol{\sigma}_G$ and a scale-dependent component $\bar{\boldsymbol{\sigma}}_G$ to account for the sliding-window detector's discrete scanning stride.

3.1. Multi-detector likelihood with occlusions

We now extend the above observation likelihood of [22] to include multiple detectors as evidence. To incorporate local part detections robustly we perform occlusion handling by explicitly leveraging 3D scene information. For each part p (we also refer to the full object detector as a *part* in the following) we compute its projection's expected visibility v_p^i based on the global 3D scene model. Assuming that the camera views the scene along the x-axis and that the objects are sorted with increasing depth, we can formally express a part's visibility as:

$$v_p^i = \text{AREA}\big(\mathbf{o}_p^i \setminus \bigcup_{j<i} \mathbf{o}^j\big)/\text{AREA}(\mathbf{o}_p^i), \quad \text{s.t. } \forall_j O_x^j < O_x^i \quad (5)$$

where $\text{AREA}(\mathbf{o}_p^i)$ denotes the image area in pixels covered by the projection of $\mathbf{o}_p^i$. Alg. 1 gives an efficient algorithm for obtaining v_p^i for rectangular projections $\mathbf{o}_p^i$. As detectors tend not to respond for parts with low visibility due to the lack of occluded samples in the training data, we discard part detections when the visible area v_p^i is below a certain threshold v_{min} (in practice, $v_{min} = 0.75$). We define our multi-detector observation likelihood with explicit occlusion handling as a mixture of experts [13] where the experts are the part detectors and the weights are proportional

Algorithm 1 Efficient visible area computation for rectangular regions: $\mathbf{r}$ - rectangle for which the number of visible pixels is computed; m - maximum tested object depth.
Since the intersection and AREA can be computed quickly for rectangles, this algorithm is faster in practice than a dense pixel-wise occlusion map, which is often used for arbitrary shapes.

Require: $\mathbf{O}^1, \ldots, \mathbf{O}^m$ sorted in increasing depth
function VISIBLEAREA($\mathbf{r}$, m)
  $v^r \leftarrow$AREA($\mathbf{r}$)
  for $k = 1 \ldots m-1$ **do**
    $\mathbf{o}^k \leftarrow$PROJECT($\mathbf{O}^k$)
    if $\mathbf{r} \cap \mathbf{o}^k \neq \varnothing$ **then**
      if $k \neq 1$ **then**
        $v^r \leftarrow v^r -$VISIBLEAREA($\mathbf{r} \cap \mathbf{o}^k, k$)
      else
        $v^r \leftarrow v^r -$AREA($\mathbf{r} \cap \mathbf{o}^k$)
      end if
    end if
  end for
  return v^r
end function

$v_p^i \leftarrow$ VISIBLEAREA($\mathbf{o}_p^i, i$)/AREA($\mathbf{o}_p^i$)

to the visible area v_p^i of those parts:

$$P(\mathcal{E}|\mathbf{X}) \propto \prod_i \frac{1}{\sum_p \delta[v_p^i > v_{min}] \cdot v_p^i} \cdot \quad (6)$$
$$\Big[\sum_p \delta[v_p^i > v_{min}] \cdot v_p^i \cdot \Psi_D\big(\mathbf{d}_p^{\mathbf{a(i)}}\big) \cdot \Psi_G\big(\mathbf{O}^i, \Theta; \mathbf{d}_p^{\mathbf{a(i)}}\big)\Big]$$

Here, $\mathbf{a(i)}$ denotes the association function that assigns candidate detections $\mathbf{d}_p^{\mathbf{a(i)}}$ (at most one for each part p) to every 3D object hypothesis $\mathbf{O}^i$. In case a detector is not firing despite a sufficiently large estimated visibility ($v_p^i > v_{min}$) we use a minimum appearance score to compensate missing evidence. Ψ_D and Ψ_G are defined as for the single detector likelihood, but use the associated part detector's estimate for the full object extent instead of the full body detector. Regarding the comparability of detector scores we found empirically that SVM margins on true positive detections tend to be larger for better performing detectors. This is probably due to the fact that we train all detectors on the same training set, and thus scores are implicitly normalized by scaling the SVM margin to 1. Therefore an implicit detector weighting is learned during SVM training and no further provision to balance SVM scores is required.

Multi-frame model. In video streams it is possible to leverage evidence from adjacent frames. To that end we extend our likelihood to entire "scene tracklets" [22, Sec. 4] and define the multi-frame observation likelihood as:

$$P(\mathbf{X}_t|\mathcal{E}_{-\delta t+t:t+\delta t}) \propto \prod_{r=t-\delta t}^{t+\delta t} P(\hat{\mathbf{X}}_r|\mathcal{E}_r), \quad (7)$$

where $\hat{\mathbf{X}}_r$ denotes the scene configuration that has been extrapolated from $\mathbf{X}_t$ based on the camera's estimated egomotion and assuming that object positions as well as the camera pitch vary only slowly in successive frames.

3.2. Inference

Hypotheses clustering. To enable efficient inference we cluster $\mathbf{a}(\mathbf{i})$ agglomeratively into groups of possible associations. Starting from an association function $\mathbf{a}(\mathbf{i})$ that only associates full object detections, we iteratively add associations to part detections when those overlap sufficiently for the respective object part. In each iteration we add the part detection with the highest overlap that has previously not been matched. Part detections that cannot be matched to an existing cluster lead to an additional, new cluster.

RJMCMC inference. Inference in our model is performed by Metropolis-Hastings MCMC sampling [22, Sec. 3.1-3.2], which employs reversible jumps in order to cope with a varying number of objects in the scene. Our framework employs diffusion, add and remove proposal moves. Add proposals are adapted from the agglomeratively clustered object hypotheses, which are selected with a probability proportional to its maximum part detector score. Finally, we perform projective 3D to 2D marginalization [22, Sec. 3.3] to compute a score for each object.

4. Experimental Results

We evaluate our models on two publicly available datasets: *ETH-Linthescher* and *ETH-PedCross2* (see Fig. 6 for sample images). Both were recorded with a moving stereo camera in densely populated pedestrian zones and originally published by Ess *et al*. [5]. The videos are recorded at a frame rate of $\sim$14Hz and a resolution of 640×480 pixels. We only use the left camera's images as input to our monocular system[1]. ETH-Linthescher is comprised of 1209 stereo image pairs with a total of 2606 annotated pedestrians. As our system with explicit occlusion reasoning is capable of detecting severely occluded pedestrians that are not contained in the original annotation, we manually extended the annotations (by all pedestrians which are at least 20% visible) to a total of 3018 pedestrians. ETH-PedCross2 consists of 840 frames recorded at a pedestrian crossing and along a rather narrow sidewalk with frequent occlusions among pedestrians. As the dataset comes without annotations, we annotated pedestrians in every 4^{th} frame similar to ETH-Linthescher, and included instances that are truncated by the image boundaries. Overall our annotations contain 1635 pedestrians[2]. We used the same set of parameters throughout all experiments and follow the evaluation protocol of Ess *et al*. [6] to consider only pedestrians with an annotation height of at least 60 pixels.

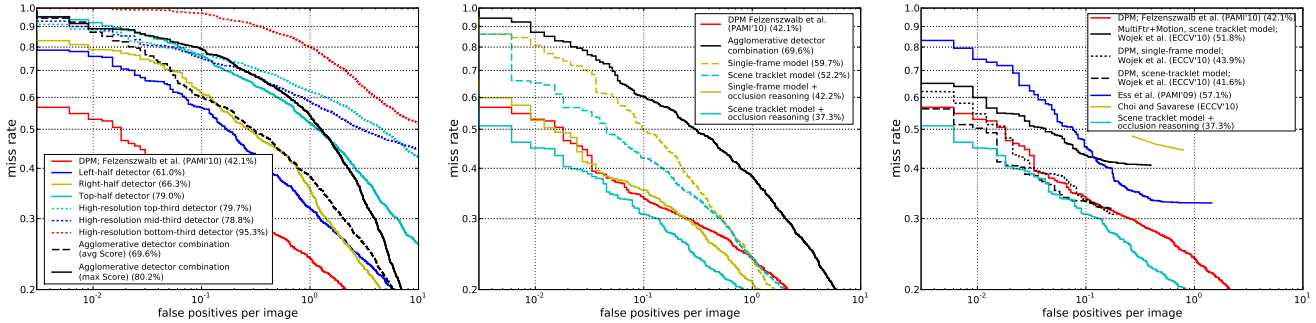
Due to the lack of 3D ground truth we project the estimated 3D models to the image plane and employ detection metrics to report full image performance as miss rate vs. false positives per image (FPPI) (see Fig. 4). Moreover, we use the *log-average miss rate* (LAMR) for an assessment across a large range of false positive rates. We define it as the average miss-rate sampled from the lowest false positive rate to a false positive rate of 1 FPPI. Missing samples for high FPPI rates are filled in with the minimum miss rate of the highest false positive rate on the curve. We use equally distant samples in log-space and therefore the log-average miss rate stresses low miss rates at high precision, which is preferable for the systems' output.

ETH-Linthescher. We start by evaluating the performance of the different human detectors on the ETH-Linthescher sequence (see Fig. 4(a)). Firstly, we observe that the part-based full body DPM detector [8] performs best as expected with an LAMR of 42.1%. When we only use left- and right-half detectors the performance drops to 61.0% for the left detector and to 66.3% for the right detector, respectively. The performance for the upper body top-half detector is even worse with an LAMR of 79.0%. This drop in performance may be explained by the missing discriminative evidence of the legs as lower object boundary. The high-resolution top-third pedestrian detector and mid-third detector roughly perform at the same level and achieve an LAMR of 79.7% and 78.8%, respectively. The missing recall is mostly due to pedestrians that appear at small scales and cannot be scanned by this detector. The bottom-third (feet) detector performs worse than the top-third and mid-third detectors and achieves an LAMR of 95.3%. When we combine all detectors by agglomerative clustering as described in Sec. 3.2, the combination achieves an LAMR of 69.6% when we use the average score of all detectors for the cluster, and 80.2% when we use the maximum score. While the clustering of detector hypotheses yields an unsatisfying false positive rate, the achieved minimum miss rate (11.2%) is promising and lower than for all stand-alone detectors. We thus use this detector combination as input and baseline for our models that employ explicit occlusion reasoning. As we will see they successfully improve the performance over systems with full-body detectors only.

Next we analyze our model's performance with explicit occlusion handling (Fig. 4(b)). As a baseline we run our model with evidence from a single frame and without occlusion reasoning. With this setup only a small improvement over agglomerative clustering can be achieved (LAMR 59.7%). When extending the evidence over multiple frames and performing scene tracklet inference, we achieve an LAMR of 52.2%. Most importantly, however, an even larger performance gain to an LAMR of 42.2% is

[1] We simulate yaw and speed sensor readings based on SfM results kindly provided by the authors of [5].

[2] Annotations are available at `http://www.d2.mpi-inf.mpg.de`.



(a) Detector performance (ETH-Linthescher) (b) Scene models (ETH-Linthescher) (c) Comparison to state-of-the-art (ETH-Linthescher)

(d) Detector performance (ETH-Pedcross2) (e) Scene models (ETH-Pedcross2) (f) Comparison to state-of-the-art (ETH-Pedcross2)

Figure 4: The first row shows results for the ETH-Linthescher sequence, the second row for the ETH-PedCross2 sequence. Percentages in the legends indicate the log-average miss rate (see text for definition). The left column shows the standalone detector performance. DPM [8] performs overall best, but detectors for body parts are able to achieve a lower miss rate at the cost of lower precision. The middle column depicts our models' performance for different configurations. Single-frame inference does not perform much better then agglomerative detector combination. Adding occlusion reasoning to the single-frame model improves more than adding scene tracklet inference. However, when both are combined the best overall performance is achieved. The right column compares this performance to other state-of-the-art methods. Our models outperform the stand-alone DPM detector [8], the stereo system by Ess *et al*. [6], our previous monocular single full-body detector system [22] and the monocular system by Choi & Savarese[2].

accomplished when using our newly proposed explicit 3D occlusion reasoning scheme. We note that this model already outperforms the standalone full body detector. When additionally using scene tracklets and thus the full model, we achieve a further improvement to an LAMR of 37.3%.

Finally, Fig. 4(c) compares our model to other state-of-the-art approaches[3]. The modular stereo system of Ess *et al*. [6] performs at an LAMR of 57.1%. Our previously proposed system with a single full body MultiFtr+Motion detector [22] achieves an LAMR of 51.8%. The same system in combination with DPM and the single-frame model achieves an LAMR of 43.9% and in combination with the scene-tracklet model an LAMR of 41.6%. We note that for all systems the minimum miss rate saturates at 30%-40%. Our model on the contrary achieves an LAMR of 37.3% with a substantially lower minimum miss rate of only 16.0%, while also improving the false detection rate for all miss rates. Choi and Savarese [2] report three points on the recall vs. FPPI curve, which are not competitive compared to our model; the miss rate is about 20% higher at the same error rates. Moreover, we note that [2] reports performance for the original annotations, which do not include all occluded pedestrians. Hence, the performance on the modified annotation set may be slightly worse. Overall, our full model with explicit occlusion reasoning and scene tracklets outperforms three state-of-the-art approaches. In particular it achieves the highest recall and reduces the error rate along the entire curve. Fig. 6 compares our full model on some sample scenes to competing state-of-the-art approaches.

ETH-PedCross2. Next, we turn to the more difficult ETH-PedCross2 sequence, which contains more occluded pedestrians. Again, we start by analyzing the detectors' performance alone (Fig. 4(d)). Similar to ETH-Linthescher, DPM [8] yields the best standalone detector performance with an LAMR of 65.2%. The next best performance again is achieved by left- and right-half detectors with 78.6% and 79.1% LAMR, respectively. The top-half detector performs at 86.0% LAMR. For the high-resolution detectors the top-third (83.0% LAMR) and mid-third detectors (84.1% LAMR) again perform better than the bottom-third detector (90.9% LAMR). Interestingly, the agglomerative detector combination with average scores (88.3% LAMR) performs worse than when the maximum score is used (76.2%). We conjecture that low scores of the full-body detector on the occluded samples lower the average score on this dataset.

[3]The authors of [6] kindly provided us with their latest results (for our annotations).

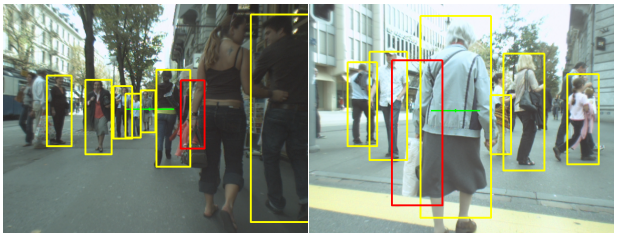


Figure 5: First (left) and second (right) false positive detection on the ETH-PedCross2 sequence for our proposed model with occlusion handling. Both false positives (red bounding boxes) are due to false detections of a left- or right-half detector and are strengthened by occluding true positives. The false detection on the left actually detects an occluded true pedestrian, but with a too large scale. This detection also suppresses the true detection on the pedestrian in the close range.

When we apply our single-frame 3D scene model (Fig. 4(e)), the performance of the agglomeratively clustered detector combination is improved only slightly to 78.3%. Again, explicit occlusion reasoning improves the results more (59.6% LAMR) than the scene tracklet formulation (68.2% LAMR). However, as for the previous sequence, the best performance is achieved when both tracklet inference and explicit occlusion reasoning are performed (56.5% LAMR).

We finally compare our model to our previously proposed approach [22], which only uses a single full body MultiFtr+Motion object detector (Fig. 4(f)). In terms of LAMR our full model outperforms this segmentation supported model (LAMR 66.9%). When we replace the detector with DPM the single-frame model achieves an LAMR of 66.1% and the scene tracklet model an LAMR of 64.1%. However, note that our new model with explicit occlusion reasoning achieves a substantially higher recall. It is also instructive to go back to the DPM full body detector, which is also clearly outperformed by our model and in addition is able to estimate 3D positions. In particular our full model achieves a substantially lower minimum miss rate (26.0%) compared to the two baselines, which achieve 39.0% [8] and 50.7% [22]. For our full model we have also analyzed false positive failures that occur with a high score. Fig. 5 shows the two highest scoring detections that are due to false partial detections and incorrectly supported by occlusion reasoning.

For this dataset we additionally analyzed the performance on partially occluded pedestrians. To that end we annotated all partially occluded pedestrians and performed the evaluation restricted to these instances. Overall 1052 pedestrians were marked as partially occluded out of which DPM [8] detected 40.7%. Our previous approach without explicit occlusion reasoning [22] was able to detect only 19.2%. This low recall compared to the standalone detector is mostly due to the tracklet formulation, which tends to drop detections that are partially occluded in at least one frame of a tracklet. The proposed model, on the other hand, can solve this shortcoming using an explicit 3D occlusion reasoning and achieves a recall almost three times better (55.0%). Our current C++ implementation runs about 2sec on average per frame on recent hardware; the run time is dependent on the number and density of objects in a scene.

5. Discussion and Conclusion

We introduced a model for multi-object tracking from a moving platform that combined 3D scene tracking, full object and object part detectors, and explicit 3D object-object occlusion reasoning to also handle objects that are partially occluded for long durations of time or never fully visible at all. As our experiments with multi-people tracking have shown, our model is capable of robustly detecting occluded and truncated pedestrians by strengthening weak evidence obtained from partial human detectors through the accumulation of geometric scene constraints and by evidence obtained over multiple frames. The proposed model outperforms similar monocular approaches [2, 22] without occlusion reasoning, as well as a stereo-based system [6], and is able to obtain a substantially higher recall than these competing approaches. Also, our approach outperforms state-of-the-art part-based detectors [8].

For future work incorporating a stronger spatial part model [9] may prove beneficial. Other possible avenues toward even more robust scene understanding for robotics and automotive applications may include incorporating a stereo formulation to make use of disparity discontinuities for occlusion reasoning and detection, or to integrate information from other, complementary sensors such as laser scanners.

Acknowledgements. We would like thank Andreas Ess for making datasets and results available.

References

[1] M. Bajracharya, B. Moghaddam, A. Howard, S. Brennan, and L. H. Matthies. A fast stereo-based system for detecting and tracking pedestrians from a moving vehicle. *The International Journal of Robotics Research*, 28, 2009.

[2] W. Choi and S. Savarese. Multiple target tracking in world coordinate with single, minimally calibrated camera. In *ECCV*, 2010.

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[4] M. Enzweiler, A. Eigenstetter, B. Schiele, and D. M. Gavrila. Multi-cue pedestrian classification with partial occlusion handling. In *CVPR*, 2010.

[5] A. Ess, B. Leibe, K. Schindler, and L. Van Gool. A mobile vision system for robust multi-person tracking. In *CVPR*, 2008.

[6] A. Ess, B. Leibe, K. Schindler, and L. Van Gool. Robust multi-person tracking from a mobile platform. *PAMI*, 31(10):1831–1846, 2009.

[7] A. Ess, K. Schindler, B. Leibe, and L. Van Gool. Improved multi-person tracking with active occlusion handling. In *ICRA Workshop on People Detection and Tracking*, 2009.

[8] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *PAMI*, 32:1627–1645, 2010.

[9] M. A. Fischler and R. A. Elschlager. The representation and matching of pictorial structures. *IEEE Trans. on Computer*, 22(1):67–92, Jan. 1973.

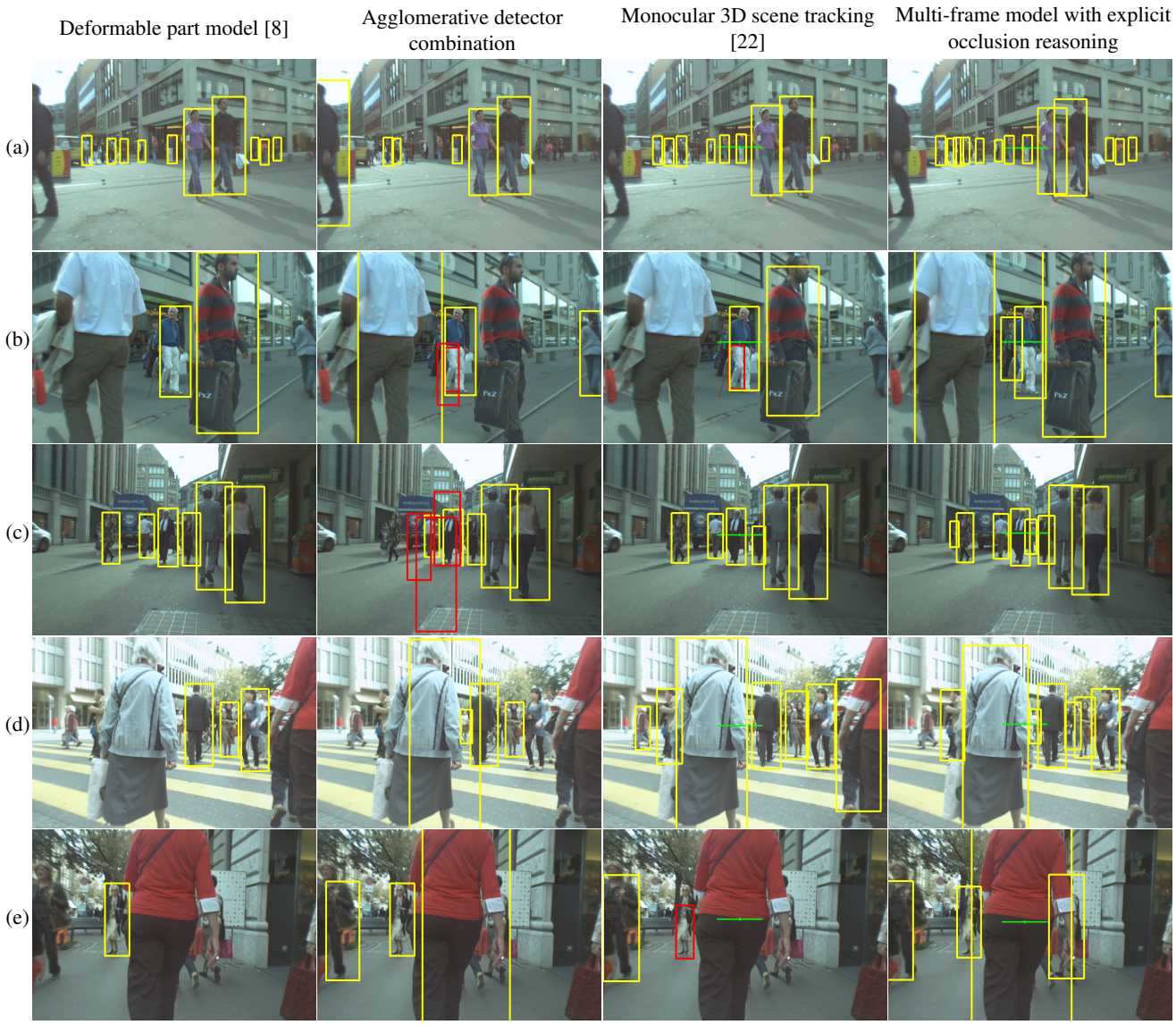


Figure 6: The first three rows show sample scenes for ETH-Linthescher, the last two rows for ETH-PedCross2. All results are depicted at a constant error rate of 1 FPPI. Yellow bounding boxes are true positives, red boxes are false positives. Both DPM and our previous work without explicit occlusion handling [22] are not able to cope with occlusion and object truncation at the image boundary well. On the contrary the model proposed here and shown in the rightmost column is able to detect occluded pedestrians as well as pedestrians very close to the camera.

[10] D. M. Gavrila and S. Munder. Multi-cue pedestrian detection and tracking from a moving vehicle. *IJCV*, 73:41–59, 2007.

[11] C. Huang, B. Wu, and R. Nevatia. Robust object tracking by hierarchical association of detection responses. In *ECCV*, 2008.

[12] M. Isard and J. MacCormick. BraMBLe: A Bayesian Multiple-Blob tracker. In *ICCV*, pages 34–41, 2001.

[13] R. Jacobs, M. Jordan, S. Nowlan, and G. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991.

[14] R. Kaucic, A. G. Perera, G. Brooksby, J. Kaufhold, and A. Hoogs. A unified framework for tracking through occlusions and across sensor gaps. In *CVPR*, 2005.

[15] Y.-Y. Lin, T.-L. Liu, and C.-S. Fuh. Fast object detection with occlusions. In *ECCV*, 2004.

[16] S. Maji, A. Berg, and J. Malik. Classification using intersection kernel SVMs is efficient. In *CVPR*, 2008.

[17] V. Shet, J. Neumann, V. Ramesh, and L. Davis. Bilattice-based Logical Reasoning for Human Detection. In *CVPR*, 2007.

[18] L. Sigal and M. Black. Measure locally, reason globally: Occlusion-sensitive articulated pose estimation. In *CVPR*, 2006.

[19] A. Vedaldi and A. Zisserman. Structured output regression for detection with partial truncation. In *NIPS*, 2009.

[20] X. Wang, T. X. Han, and S. Yan. An HOG-LBP human detector with partial occlusion handling. In *ICCV*, 2009.

[21] J. Winn and J. Shotton. The layout consistent random field for recognizing and segmenting partially occluded objects. In *CVPR*, 2006.

[22] C. Wojek, S. Roth, K. Schindler, and B. Schiele. Monocular 3D Scene Modeling and Inference: Understanding Multi-Object Traffic Scenes. In *ECCV*, 2010.

[23] C. Wojek, S. Walk, and B. Schiele. Multi-cue onboard pedestrian detection. In *CVPR*, 2009.

[24] B. Wu and R. Nevatia. Detection and segmentation of multiple, partially occluded objects by grouping, merging, assigning part detection responses. *IJCV*, 82(2):185–204, 2009.

[25] J. Xing, H. Ai, and S. Lao. Multi-object tracking through occlusions by local tracklets filtering and global tracklets association with detection responses. In *CVPR*, 2009.