Improved Person Detection in Industrial Environments using Multiple Self-Calibrated Cameras

Roland Mörzinger, Marcus Thaler
JOANNEUM RESEARCH
DIGITAL - Institute for Information and Communication Technologies
Steyrergasse 17, 8010 Graz, Austria
{roland.moerzinger, marcus.thaler}@joanneum.at

Severin Stalder, Helmut Grabner, Luc Van Gool
ETH Zurich, Computer Vision Laboratory
Sternwartstrasse 7, 8092 Zurich, Switzerland
{stalder, grabner, vangool}@vision.ee.ethz.ch

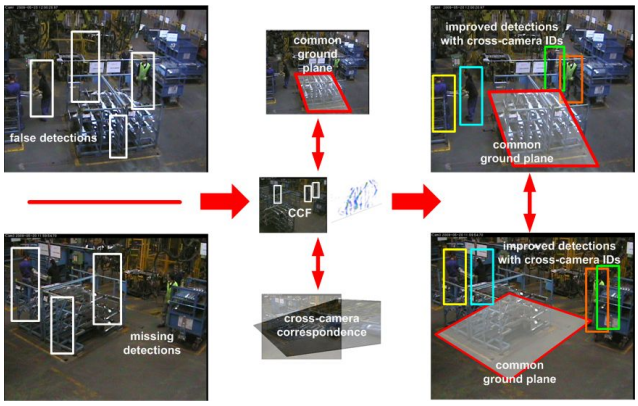
Abstract

Person detection is a challenging task in industrial environments which typically feature rapidly changing conditions of illumination and the presence of occluding objects and cluttered background. This paper proposes a series of algorithms for improving the robustness of person detection in such harsh industrial environments. Based on a state-of-the-art person detector, significant robustness and automation is achieved by introducing automatic ground plane estimation, confidence filtering, cross-camera correspondence estimation and multi-camera fusion. Detailed experiments made on an industrial dataset that captures an automotive assembly process show the stepwise improvement when combining the above mentioned techniques in a fully unsupervised manner.

1. Introduction

Safety concerns and information about quality and process efficiency are integral parts of enterprises like industrial plants or public infrastructure organizations. In large-scale manufacturing areas surveillance systems such as with multi-camera networks are often used for supervision. The goal is to ensure security and safety, i.e. the prevention of actions that may lead to hazardous situations, and quality, i.e. adherence to predefined procedures for production or services. Typically, this requires manual/human supervision attention by surveillance operators which is subjective and inefficient, especially in the presence of multiple simultaneous video streams.

Robust person detection (and tracking) is one of the first steps in order to achieve automated monitoring. Person detection from visual observations is highly challenging tasks especially in industrial environments [7]. Existing methods like color background modeling (e.g., [11]) and generic person detectors (e.g., [2]) perform dismally in such situations. The reasons are manifold: it is difficult to discern persons due to sparks and vibrations, occlusions (obscured by equipment), difficult structured background (upright racks) and other moving objects (welding machines and forklifts). Moreover the workers clothes have the same color as parts of the surroundings. The recordings suffer from rapid lighting changes (machinery in operation) and



(a) (b) (c)

Figure 1. Robust person detection in industrial environments. The performance of current/State-of-the-art person detection technology exhibits a high rate of missed and false detections (a) and needs to be improved by incorporating scene knowledge such as the ground plane, tracking and multiple cameras (b) for robust person detection with cross-camera correspondence (c).

camera shake (transport of heavy machinery).

The recently proposed Cascaded Confidence Filter (CCF) [10] combines a person detector with background modeling and short-term tracking. This algorithm significantly improves tracking, especially in the case of partial occlusions, changing backgrounds and distracting objects that look similar to the target. For that purpose, the temporal consistency of the detections is enforced through a trajectory filter. Additionally, the person detections are refined locally with respect to the appearance of the background and expected size of the person. The size of the persons needs to be known in advance and up to the present they need to be manually defined. This limitation and the expectation that the use of a multi-camera setup improves the performance of person detection serve as the motivation for this work.

This paper addresses the above mentioned challenges and aims at improving people detection in industrial environments as illustrated in Figure 1.

The contribution of this work is twofold:

- Automatic ground plane estimation and multi-camera correspondence estimation techniques overcome the limitation of manually setting up the scene specific knowledge and can directly be used as geometric filtering step in CCF.
- The CCF based person detection and tracking approach is extended for application in a multi-camera setup that observes the scene from two different partly overlapping camera views. Multiple camera systems are used in order to increase the efficiency of detection algorithms and to prevent the system from losing the target in case of object occlusions.

The remainder of the paper is organized as follows. Sec. 2 presents the series of techniques for automatic and robust person detection based on unsupervised scene calibration techniques. Detailed experiments as well as improved object detection results are shown in Sec. 3. Finally, Sec. 4 suggests further work and concludes the paper.

2. Approach

This section presents a set of fully automatic techniques for improving the robustness of person detections in industrial environments.

First, for two image sequences (of two partly overlapping camera views) a state-of-the-art approach is applied for initial person detection.

Second, the ground plane is automatically estimated in the two views individually. This information serves as geometric filter for removing detections at wrong scale.

Third, the cascaded confidence filter improves the detections by exploiting constraints on the geometry, the pre-dominant background and assumptions on the smooth movement of the objects in the scene.

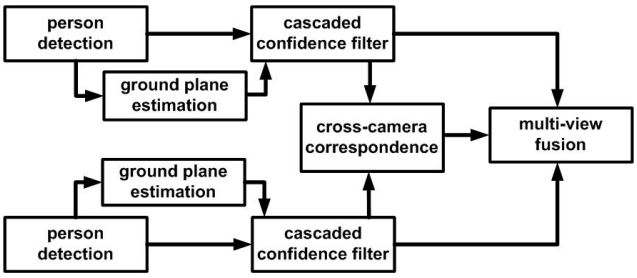


Figure 2. Overview of the approach. Person detection is improved by self-calibrated multi-view monitoring.

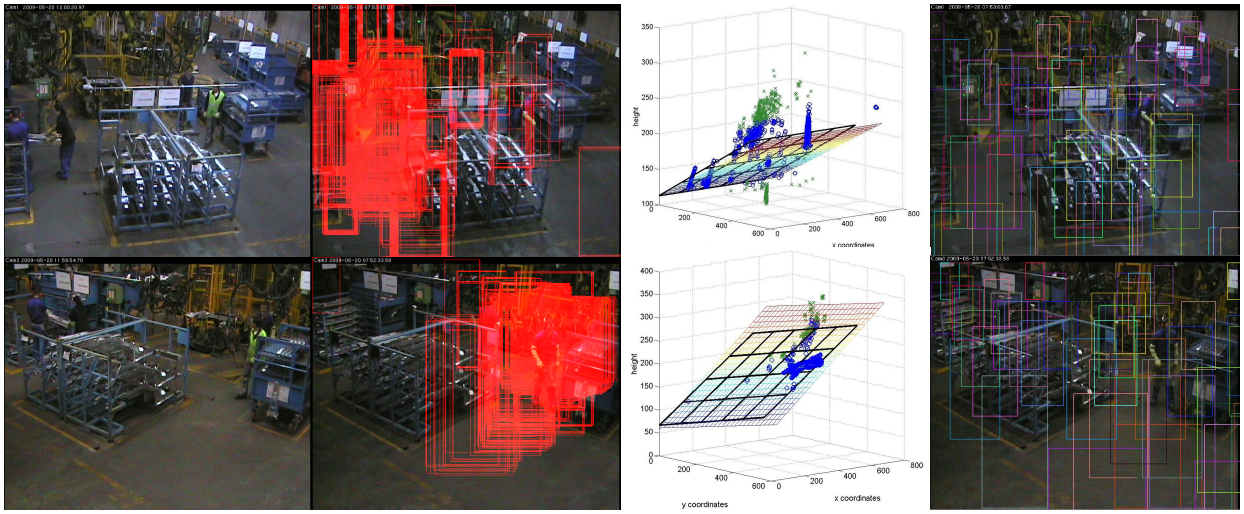
Fourth, the resulting improved detections enable the automatic estimation of the correspondence between the two camera views. Cross-camera correspondence is a prerequisite for linking overlapping parts in the individual views.

Fifth and finally, the detections in the individual views are fused in a common coordinate system. An illustration of the proposed approach is shown in Figure 2 and details to each of the used techniques are given in the following subsections.

2.1. Unsupervised Ground Plane Estimation

The task of object detection can be assisted by focusing the analysis on image regions where people are typically located, i.e. the ground plane. We propose a robust model for automatically estimating the ground plane from a few person detections. For that purpose we introduce assumptions that are valid in many surveillance scenarios, namely a static camera, a single planar ground plane and an approximately equal size of the objects of interest. In this case, the person height in the image varies linearly with its vertical position in the image.

The approach is summarized in the following, for details the reader is referred to [8]. First, the state-of-the-art detector [1] densely scans sample frames of the input video at multiple locations and scales and collects detection results with confidence scores. Second, based on these detection results the estimated person heights are calculated as a linear function of the feet locations (x and y image coordinates) in the presence of false-positive detections. In particular, outliers are removed by fitting a plane into the 3D point cloud using RANSAC [3]. Subsequently, the linear scale model is robustly fitted on the remaining inliers by taking into account their confidence scores, i.e. considering them as weights so to reduce the influence of an unreliable observation on the fit. We apply a linear regression using weighted least-squares for the 3D plane model [8]. The linear scale model is used in the following steps as geometric filter, i.e. for discarding (false positive) detections that do not fit to the estimated scene.



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

Figure 3. Estimated ground plane for cam1 (top row) and cam3 (bottom row). Sample images for each camera view are shown in (a), accumulated detections from original detector [1] (b), estimated scale model showing outliers in green circles and ground truth by a black mesh (c) and sampled likely detection windows for person detections (d). Best viewed in color.

Figure 3 illustrates the ground plane estimation based on 300 accumulated detections by means of two examples. The figures in the center plot the observations (height of the person detections) over the x and y image coordinates. The estimated (in fine colored mesh) and ground truth (solid black mesh) linear scale model are also shown.

2.2. Refined Detections by Confidence Filtering

The Cascaded Confidence Filter (CCF) [10] approach has recently be shown to significantly improve tracking-by-detection results on challenging datasets. Detection responses are put into their spatial and temporal context. In particular, a geometric filtering step, cf. Section 2.1 assumes object movement only on a single common ground plane and constrains object detections to appropriate candidate windows that satisfy the geometric constraints to allow for a certain variance of person sizes. Subsequently, based on the geometrically filtered confidence scores, a background filter models the distribution of the background that is assumed to be present more often than moving objects. The output can be seen as a location specific threshold adaptation for detection which suppresses static background structures. In addition, a trajectory filter similar to a vessel filter approach in medical imaging [4] is applied for detecting object that move smoothly (on the ground plane), i.e. detection confidences that change continuously over time. Specifically the smoothness of trajectories is ensured through a process analogous to vessel filtering in medical imaging. Figure 7 shows the input person detections in column (a) and the improved result after application of the proposed filters in column (b). The remaining sporadic missed detections (marked by green dashed boxes) are mainly due to occlusions. The use of multiple cameras attempts to solve this problem.

2.3. Unsupervised Cross-Camera Correspondence

Monitoring in industrial environments often faces the problem of occlusions and thus failed tracking when only a single camera is used to take the scene. We therefore make use of a multi-camera setup where two cameras simultaneously observe the same scene with partly overlapping views. It is obvious that these two camera views need to be linked and calibrated with respect to each other. Two partly overlapping camera views observing the same planar scene can be linked using an inter-image homography. Such a homography is typically estimated from at least 4 pairs of corresponding points [5]. Automatically obtaining the correspondence between the views by using matching image features is not promising since interest points located on the common ground plane are missing and brightness and appearance promixity constraints do not hold.

Therefore our approach to estimate the correspondences between the two views only builds on the detections resulting from the previous steps, strictly speaking the feet positions derived from the bounding boxes. This means that our approach does not require points with existing correspondence information as opposed to other approaches that use matching color [9] or SIFT features [6] for finding corresponding points between two views. Instead, the true correspondences are obtained by exploiting temporal and geometric information followed by an error counting procedure which maximizes the inlier support of the homography. De-

tails on this technique can be found in [12]. The approach is fully automatic and adaptive, it can be applied incrementally (the more detections the better) and it is suited for wide angles between the camera views.

It is worth to note that for successful correspondence estimation the approach needs a certain detection accuracy. It is essential that the same person is sufficiently often detected in both views at the same time (high precision). This is the case only after application of CCF. For details on the improved precision value see Section 3.2.

Figure 4 illustrates the overlapping camera views (a) and the resulting homography (b) and ground truth homography (c) when used for warping all pixels of cam3 to cam1.

2.4. Multiple Camera Fusion

After automatically extracting the cross-camera correspondence, i.e. the planar homography between the image planes of the overlapping camera views, the person detection results with cascaded confidence filtering obtained in the individual views need to be fused. One of the views is selected as reference (2D Euclidean) coordinate system and all detected persons (strictly speaking their feet positions) in the corresponding remaining view are projected to the reference view. For fusion we adopt a simple and thus generally applicable nearest neighbor distance clustering procedure. In particular, based on the 2D Euclidean distance between the feet positions, the detections are grouped into clusters so that the distances between the locations are minimized globally. In doing so the following two constraints are imposed. First, detections from the same view must be assigned to different clusters. Second, the minimum distance between two clusters is defined as half of the estimated person height on the particular location in the image/ground plane. This equates to the average estimated average person width derived from the unsupervised ground plane estimation. The mean positions of all feet locations of each cluster yields a final improved multi-camera person detection result.

3. Experimental Results

Our approach is applied on a real-world industrial dataset [1] recorded in the NISSAN Motor Iberica SA plant. It captures an automotive assembly process with 3 operators where one operator provides pieces to be assembled and two other ones are handling pieces to put them over an robotic assembly machine. For our experiments we chose 2 image sequences (cam1 and cam3) with partly overlapping camera views and a resolution of 704x576 pixels. The total duration of each sequence is approximately 15 hours (2 shifts on two consecutive days). Synchronicity between the camera views is ensured by time-stamped filenames. In the following the results of the ground plane and cross-camera correspondence techniques and the improvement of all proposed individual steps are evaluated.

[1] available at: http://scovis.eu/?q=node/37

(a) (b)

Figure 5. Cross-camera correspondence details when cam1 is projected to cam3 using the proposed approach (a) and the ground truth calibration (b). The mean homography estimation error of 16.2 pixels becomes apparent in the region with diverging yellow floor markings (a).

3.1. Ground Plane, Cross-Camera Correspondence

We apply ground plane estimation using 300 random person detection samples, see column (b) in Figure 3. The ground plane estimation provides the expected person height for a certain position in the image. For evaluation we compare this value to the real person height (ground truth) in pixels. Based on 50 manually extracted person samples at different locations, the mean error (between the estimated person heights and real person heights) is 11 pixels for cam1 and 6 pixels for cam3. This corresponds to an error of person height of 7.2% for cam1 and 5.4% for cam3. These numbers demonstrate the good accuracy of the approach, the error is equal to the natural error resulting from imprecise person detections, i.e. detections that are not exactly centered exactly around the person. The homography for the cross-camera correspondences (between cam1 and cam3) is estimated from 30 synchronized image pairs with 90 input detections in total. For comparison, both the estimated homography and a ground truth homography (derived from manual calibration using checkerboard patterns) are used for projecting all pixels of cam3 to cam1.

The results, shown in Figure 4 and 5, exhibit only a marginal difference (mean error of 16.2 pixels). The largest error can be seen in the center left image region, c.f. diverging yellow floor markings (left). The reason is that in this area no detections occurred and thus the quality of the homography decreases. The mean error (distance) between all points projected from cam1 to cam3 is also indicated.

3.2. Improved Person Detection

In order to evaluate the improvement for person detection when applying the individual proposed steps, we manually annotated all visible persons in 1000 frames (every 10^th frame in a 40 seconds image sequence recorded at 25fps) for the two camera views. A detection is correct if the intersecting area between its region and the ground truth region is larger than 50% of the union of the two re-

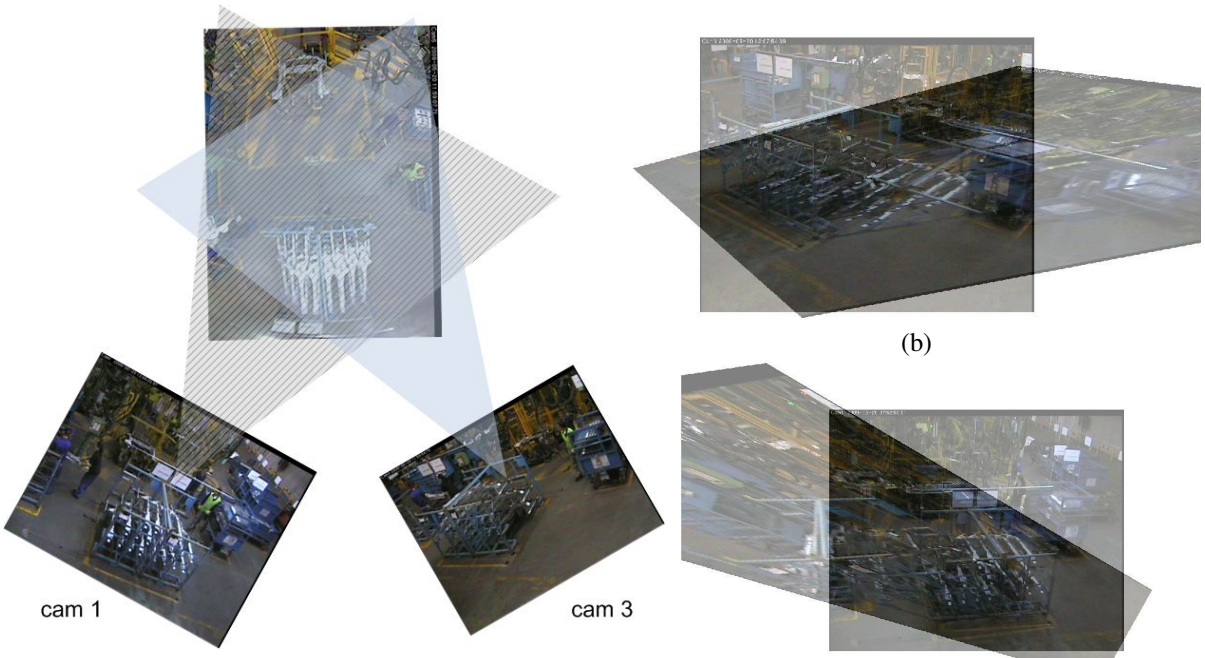


Figure 4. Estimated inter-image homographies. Two camera views (cam1 on the left and cam3 on the right) partly overlap and cover the welding place (a). The view of cam1 is projected to cam3 using the proposed approach (b) and the cam3 projected to cam1 (c).

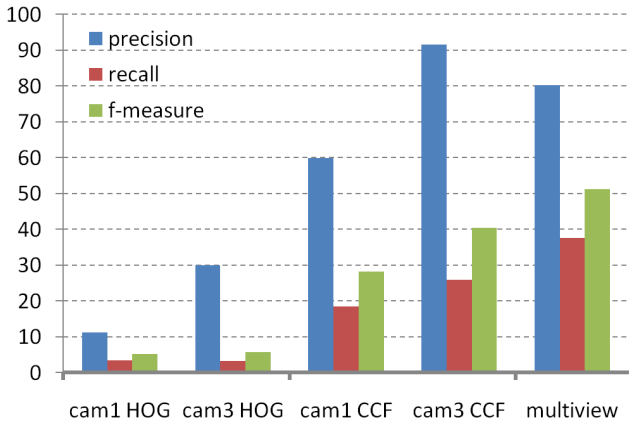


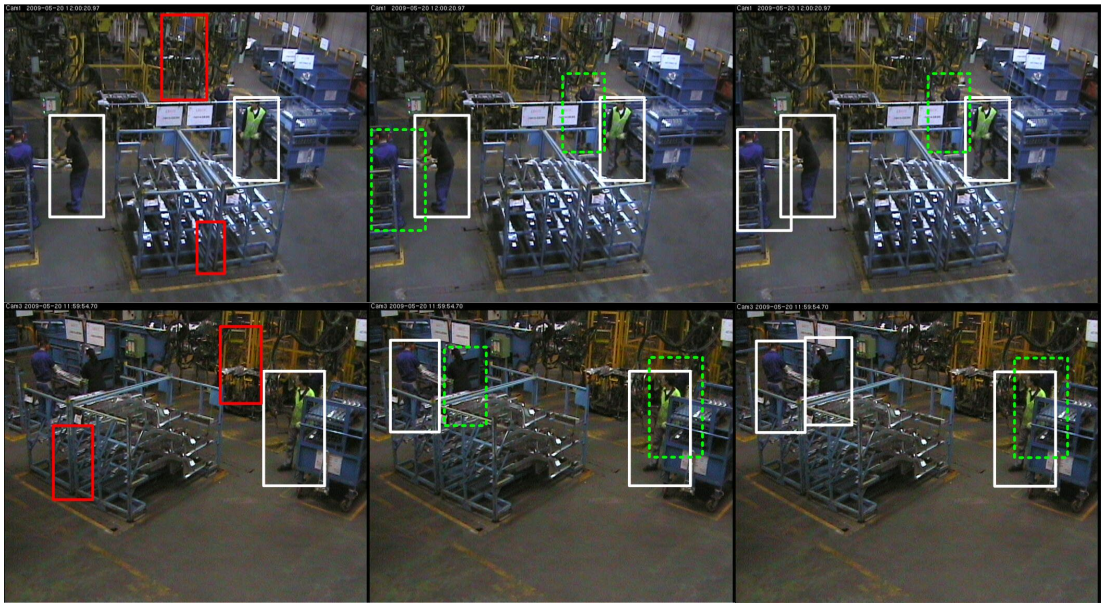
Figure 6. Comparison of the different techniques for improving the robustness of person detections. For two camera views (cam1 and cam3), the effect of applying a state-of-the-art person detector (HOG), the confidence filtering (CCF) and the multiple camera fusion is evaluated by means of precision, recall and F-measure.

gions. Figure 6 shows the values for precision (detection accuracy), recall (detection rate) and the harmonic mean thereof, the F-measure for the state-of-the-art person detector (HOG), the proposed ground plane estimation, the confidence filtering (CCF) and the multiple camera fusion. The notable general difference between the single views cam1 and cam3 are mainly due to the more challenging scene observed by cam1, i.e. more occluding objects, people at smaller scale.

The boost in precision (from 30% to 92% for cam3) when using CCF is the effect of its multiple constraints on the assumed scene scale, the background and the smoothness of trajectories. So it is possible to remove false positives leading to an increased precision. The refined detections using CCF also improve the recall value significantly from 3% to 25% for cam3. More positive detections are obtained due to the location specific threshold adaptation. The use of multiple camera views is able to resolve partial occlusions. The best performance is therefore obtained after multi-camera fusion, i.e. the F-measure of 54 for multi-view compared to 35 and 41 in the individual views. But the multi-view fusion's improvement in recall is at the cost of a decline in precision, i.e. from 59% and 92% for single views to 80% for multi-view. This is a consequence of accumulated errors, such as in cases where a single false detection in each of the two camera views is finally fused to two false positive person detection results.

Figure 7 demonstrates examples of improved person detection results for cam1 (top row) and cam3 (bottom row) after application of the proposed steps. More person detection results are available as supplementary video material [2].

[2] http://www.youtube.com/watch?v=7mgiTNKOgaY



(a) HOG person detections (b) CCF (c) multi-view

Figure 7. Example detections for cam1 (top row) and cam3 (bottom row) after applying (a) the HOG-based person detector, (b) cascaded confidence filter and (c) multiple camera fusion. False detections in (a) are indicated in red, missed detections by a dashed green box.

4. Conclusions

This paper presented a series of algorithms for improving the robustness of person detection technology in challenging industrial environments. The most important boost in precision is due to CCF imposing a ground plane constraint which was automatically estimated in this paper. The fact that only person detection is needed to establish a homography between two overlapping views is a nice feature because it utilizes the output of existing state-of-the-art human detection technology and is automated, unlike manual checkerboard calibration. This is important for an application in industrial environments and scenes which lack in interest points located on the common ground plane. For multi-camera fusion the use of person detection output is rather challenging. Specially if multiple people are located close together, imprecise detections where bounding boxes are not exactly centered around the persons, lead to wrong cross-camera correspondence and false fusion results. A relatively small displacement of a few pixels in one camera view may result in a considerable error when being projected to another camera view. Our experiments showed that due to this reason the accuracy was slightly decreased when multiple overlapping camera views are used, but the overall performance was generally improved, since occlusions can be better resolved.

References

[1] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. CVPR*, 2005. 2, 3

[2] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *Proc. CVPR*, 2008. 1

[3] M. Fischler and R. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Comm. of ACM*, 1981. 2

[4] A. Frangi, W. Niessen, K. Vincken, and M. Viergever. Multiscale vessel enhancement filtering. 1998. 3

[5] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, second edition, 2004. 3

[6] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004. 3

[7] L. Middleton and J. R. Snowdon. Histogram of confidences for person detection. In *17th IEEE International Conference on Image Processing*, 2010. 1

[8] R. Mörzinger and M. Thaler. Improving person detection in videos by automatic scene adaptation. In *Proc. Int. Conf. on Comp. Vision Theory and App.*, 2010. 2

[9] J. Orwell, P. Remagnino, and G. Jones. Multi-camera colour tracking. In *IEEE Workshop on Visual Surveillance*, 1999. 3

[10] S. Stalder, H. Grabner, and L. V. Gool. Cascaded confidence filtering for improved tracking-by-detection. In *Proc. ECCV*, 2010. 2, 3

[11] C. Stauffer and W. Grimson. Adaptive background mixture models for real-time tracking. In *Proc. CVPR*, 1999. 1

[12] M. Thaler and R. Mörzinger. Automatic inter-image homography estimation from person detections. In *Proc. AVSS*, 2010. 4