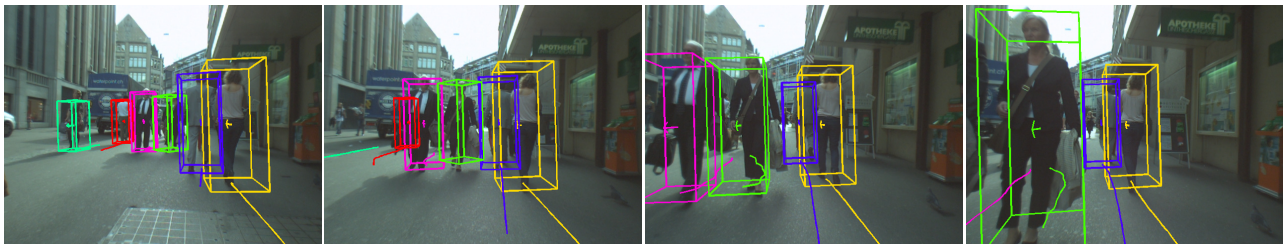
# A Mobile Vision System for Robust Multi-Person Tracking

Andreas Ess[1] Bastian Leibe[1] Konrad Schindler[1] Luc Van Gool[1,2]
[1]ETH Zurich, Switzerland [2]KU Leuven, Belgium
{aess,leibe,konrads}@vision.ee.ethz.ch vangool@esat.kuleuven.be



## Abstract

*We present a mobile vision system for multi-person tracking in busy environments. Specifically, the system integrates continuous visual odometry computation with tracking-by-detection in order to track pedestrians in spite of frequent occlusions and egomotion of the camera rig. To achieve reliable performance under real-world conditions, it has long been advocated to extract and combine as much visual information as possible. We propose a way to closely integrate the vision modules for visual odometry, pedestrian detection, depth estimation, and tracking. The integration naturally leads to several cognitive feedback loops between the modules. Among others, we propose a novel feedback connection from the object detector to visual odometry which utilizes the semantic knowledge of detection to stabilize localization. Feedback loops always carry the danger that erroneous feedback from one module is amplified and causes the entire system to become instable. We therefore incorporate automatic failure detection and recovery, allowing the system to continue when a module becomes unreliable. The approach is experimentally evaluated on several long and difficult video sequences from busy inner-city locations. Our results show that the proposed integration makes it possible to deliver stable tracking performance in scenes of previously infeasible complexity.*

## 1. Introduction

Computer vision has seen tremendous progress in recent years. Many individual disciplines have advanced to a state where algorithms are becoming applicable for real-world tasks. These successes have fostered the demand for actual vision systems. In particular, there is a strong need for mobile vision systems than can operate in unconstrained scenarios of daily human living. Building such systems has been a far-end goal of scene understanding since the 1970ies, but it is also a crucial requirement for many applications in the near future of mobile robotics and smart vehicles. So far, however, the sheer complexity of many real-world scenes has often stymied progress in this direction.

In this paper, we focus on an important building block for mobile vision applications, namely the capability to track multiple people in busy street scenes as seen from a mobile observer. This could be a mobile robot, an electric wheelchair, or a car passing through a crowded city center. As can be seen in the above figure, such a scenario puts extreme demands on the underlying vision algorithms. Many people are walking through the system's field of view, crossing and occluding each other, undergoing large scale changes, and occasionally even blocking almost the entire scene.

It has long been argued that scene analysis in such complex settings requires the combination of and careful interplay between several different vision modules. However, it is largely unclear how such a combination should be undertaken and which properties are critical for its success. In this paper, we propose a specific design how to integrate visual odometry, depth estimation, object detection, and tracking, and demonstrate its applicability in practice.

One important component of the proposed integration is the concept of cognitive feedback. The underlying idea is to derive higher-level semantic information from one vision module and feed it back to the other modules in order to improve performance there. Several instantiations of this concept have been successfully demonstrated in recent years, among them the feedback from recognition to segmentation [4, 17], from geometry estimation to object detection [13, 15], and the often-used feedback from tracking to de-

978-1-4244-2243-2/08/$25.00 ©2008 IEEE

tection (e.g. [16]). Here, we propose another such feedback path, namely to make visual odometry more robust through semantic information from object tracking.

However, the creation of feedback loops always carries the danger that measurement noise may be picked up and amplified to the point that the entire system becomes unstable (as in the case when a microphone is held too close to a connected loudspeaker). An important design question is therefore how to avoid such instabilities and guarantee robust performance. We specifically address this question by incorporating automatic failure detection and correction mechanisms into our system and show how they interact to stop error amplification. As our experiments will demonstrate, the resulting system achieves robust multi-object tracking performance on very challenging video data.

The paper is structured as follows. After discussing related work in the following section, Section 3 describes the different components of our system and their interactions in detail. Section 4 then introduces our novel cognitive feedback and shows how it improves system robustness. Finally, Section 5 presents experimental results.

## 2. Related Work

**Visual Odometry.** The majority of the work in visual odometry (VO) is based on local features and RANSAC-type hypothesize-and-test frameworks [21, 25]. Some other approaches include Hough-like methods [19] or recursive filtering [7, 8]. Most of these have however not been demonstrated on extended runs in realistic outdoor scenarios. The main problem with all these methods is the assumption that a dominant part of the scene changes only due to camera egomotion. As a result, these approaches are prone to failure in crowded scenes with many independently moving objects. While there has been work on multi-body Structure-from-Motion [18, 23], most systems are still constrained to short videos, and more importantly, assume sufficiently large, rigidly moving objects. In robotics, various approaches for SLAM in dynamic environments [3, 11] exist, related to the above, but mostly focusing on range data. In this paper, we propose to explicitly feed back information from object tracking to egomotion estimation, thereby introducing semantics.

**Object Detection.** Pedestrian detection has reached an impressive level [5, 17, 26, 27, 28]. By themselves, many available approaches already perform quite well on individual images. However, in order to achieve robustness to adverse imaging conditions, it is necessary to supply them with additional information. Examples for such supplements include motion cues [27, 6], stereo depth [10, 9], scene geometry [13, 15], or feedback from tracking [28, 16]. In this paper, we will integrate the latter three cues.

**Multi-body Tracking.** Many approaches are available for multi-object tracking from stationary cameras (*e.g.* [2, 14]). The task is however made considerably harder when the camera itself can move. In such cases, background subtraction is no longer a viable option, and tracking-by-detection approaches seem to be the most promising alternative [22, 10, 15, 28].

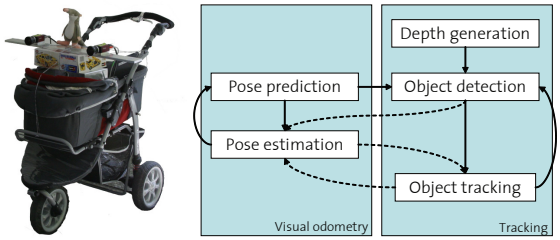



Figure 1. (Left) Mobile recording system equipped with camera pair. (Right) Components of our mobile vision system and their connections, executed for each frame of a video sequence. The dashed arrows indicate the novel cognitive loop examined in this paper.

For applications on moving vehicles, it has been shown that visual odometry can considerably help tracking, allowing it to operate in 3D world coordinates [15]. In this paper, we will complete this interaction to a loop by also feeding back information from tracking to help visual odometry. As we will show in Section 4, such a feedback is crucial for robust performance in crowded scenes. A similar idea for stabilizing visual odometry was suggested by [29]. However, their method is solely based on feature points (without occlusion handling). Thus, it does not incorporate semantics and is not suitable for articulated motions.

## 3. System

Our system is based on a mobile platform that is equipped with a pair of forward-looking cameras. Fig. 1 gives an overview of the proposed vision system. For each frame, the blocks are executed as indicated: first, a depth map is calculated, and the new frame's camera pose is predicted. Then objects are detected, taking advantage of trajectory knowledge and depth information. One of the novelties of this paper is to use this output for stabilizing visual odometry, which updates the pose estimate for the platform and the detections, before running the tracker on these updated detections. The whole system is held entirely causal, *i.e.* at any point in time, we only use information from the past and present frame pairs. The following subsections detail the three main components of the system, as well as how they are implemented in a robust manner.

### 3.1. Object Detection

The graphical model of Fig. 2 represents the core of our tracking-by-detection system. It builds upon ideas for single-frame scene analysis by [13, 9], but adapts their models with several improvements. It performs inference over

object detections $o_i$, supported by a ground plane $\boldsymbol{\pi}$ and local depth measurements $d_i$. Data is extracted per frame from the image $\mathcal{I}$ and the stereo depth map $\mathcal{D}$.

Briefly stated, the graphical model operates as follows. For each frame, a set of object hypotheses is provided by an object detector. Based on these, a stereo depth map, and prior information, the model structure is built up. Belief propagation is then used to find a geometrically consistent set of hypotheses, before a global optimization step resolves object-object occlusions.

Compared to earlier work, our model contains the following improvements: firstly, we detect occlusions in the depth maps (using a left-right check between the two views), which results in an occlusion map $\mathcal{O}$. This information is used to explicitly re-weight the prior of the depth flag $P(d_i)$ according to the level of occlusion in the depth map: in areas with missing depth estimates due to occlusion or otherwise insecure estimates, confidence in this cue is reduced. On the one hand, this is important in case of near objects that cause large occlusions. On the other hand, this also allows one to increase the confidence in the depth map in the remaining regions. Thus, not only the performance of the graphical model is improved, but object detections can be placed more accurately in world space using depth information instead of backprojecting bounding boxes.

Secondly, we remove the only explicit cycle between $\boldsymbol{\pi}$ and $d_i$ compared to [9], which simplifies the belief propagation algorithm used for inference and increases its stability. Using this new model, inference becomes:

$$P(\boldsymbol{\pi}_t, \boldsymbol{\pi}_{t-1}, o_i, d_i, \mathcal{E}) = P(\boldsymbol{\pi}_t|\boldsymbol{\pi}_{\mathcal{D}})P(\boldsymbol{\pi}_t|\boldsymbol{\pi}_{t-1})\prod_i Q_i \quad (1)$$

$$Q_i = P(o_i|\boldsymbol{\pi}_t, d_i)P(o_i|\mathcal{H}_{t0:t-1})P(\mathcal{I}|o_i)P(\mathcal{D}|d_i)P(\mathcal{O}|d_i),$$

where $\mathcal{E} = \{\mathcal{I}, \mathcal{D}, \mathcal{O}, \boldsymbol{\pi}_{\mathcal{D}}, \mathcal{H}_{t0:t-1}\}$ is the available evidence. An object's probability $P(o_i|\boldsymbol{\pi}_t, d_i)$ depends both on its geometric world features (distance, height) and its correspondence with the depth map (distance, uniform depth). The factor $P(o_i|\mathcal{H}_{t0:t-1})$ incorporates past trajectories $\mathcal{H}$ and $P(\mathcal{I}|o_i)$ the detector's probability. The time indices were omitted from $o_i$ and $d_i$ for the sake of brevity.

Finally, we introduce temporal dependencies, indicated by the dashed arrows in Fig. 2. For the ground plane, we propagate the previous state as a temporal prior $P(\boldsymbol{\pi}_t|\boldsymbol{\pi}_{t-1}) = (1-\alpha)P(\boldsymbol{\pi}_t) + \alpha P(\boldsymbol{\pi}_{t-1})$ that augments the per-frame information from the depth map, $P(\boldsymbol{\pi}_t|\boldsymbol{\pi}_{\mathcal{D}})$. For the detections, we add a spatial prior for object locations that are supported by candidate trajectories $\mathcal{H}_{t0:t-1}$ from tracking. As shown in Fig. 2, this dependency is non-Markovian due to the tracking framework explained in the following section. For details on training of the graphical model, we refer to [9].

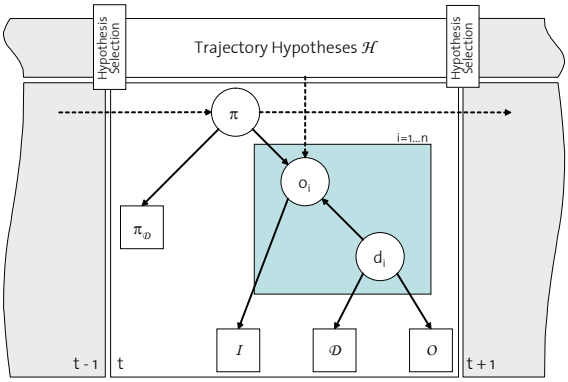



Figure 2. Graphical model for tracking-by-detection with additional depth information.

### 3.2. Tracking-by-Detection

Object detections are placed into a common world frame using camera positions estimated from visual odometry. The actual tracking system follows a multi-hypotheses approach, similar to the one described in [16]. We do not rely on background subtraction, but instead accumulate detections of the current and past frames in a space-time volume.

This volume is analyzed by growing trajectory hypotheses using independent Kalman filters. By starting this analysis from various points in time, an overcomplete set of trajectories is obtained, and pruned to a consistent estimate using model selection. Overlapping trajectory hypotheses are resolved with a global optimization step, in which trajectories compete for detections and space-time volume. For the mathematical details, we refer to [16]. The selected trajectories $\mathcal{H}$ are then used in the next frame to provide a spatial prior for the object detections.

The most important effects of this are automatic track initialization (usually, after about 5 detections), as well as the ability to recover temporarily lost tracks, thus enabling the system to track through occlusions. Obviously, such a tracking system critically depends on an accurate and smooth egomotion estimate.

### 3.3. Visual Odometry

To allow reasoning about object trajectories in the world-coordinate frame, the camera position is estimated using visual odometry. The employed system builds on previous work by [21]. See Fig. 3 for a flow diagram. In short, each incoming image is divided into a grid of $10 \times 10$ bins, and an approximately uniform number of points is detected in each bin using a Harris corner detector with locally adaptive thresholds. This encourages a feature distribution that allows stable localization. In the initial frame, stereo matching and triangulation provide a first estimate of the 3D structure. In subsequent frames, we use 3D-2D matching to get correspondences, followed by RANSAC with 3-point pose [20]. Bundle adjustment is run on a window of $n_b = 18$

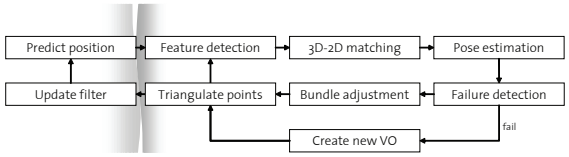



Figure 3. Flow diagram of the employed visual odometry system. The shaded regions indicate the insertion points for the feedback from object tracking.

past frames, smoothing the trajectory. Older frames are discarded, along with points that are only supported by these removed frames.

The key differences to previous systems are the use of 3D-2D matching to bridge temporally short occlusions of a feature point and to filter out independently moving objects at an early stage, as well as the use of a Kalman filter to predict the next camera position for object detection. This makes feature selection similar to the active search paradigm known from the SLAM literature [7]. Scene points are directly associated with a viewpoint-invariant SURF descriptor [1] that is adapted over time. In each frame, the 3D-2D correspondence search is then constrained by the predicted camera position. As mentioned above, only scene points without support in the past $n_b$ frames are discarded. This helps bridging temporally short occlusions (*e.g*. from a person passing through the image) by re-detecting 3D points that carry information from multiple viewpoints and are hence more stably localized. Attaching 2D descriptors to coarse 3D geometry is also a recent trend in object recognition [24].

In order to guarantee robust performance, we introduce an explicit failure detection mechanism. In case of failure, the Kalman filter estimate is used instead of the measurement; all scene points are cleared; and the VO system starts anew. This allows us to keep the tracker running without resetting it. While such a procedure may introduce a small drift, a locally smooth trajectory is more important for our application than accurate global localization. We believe that the latter is best done by integrating different sensor modalities, such as GPS and INS, see *e.g*. [30].

### 3.4. Failure Detection

For systems to be deployed in real-life scenarios, failure detection is an often overlooked, but critical component. In our case, ignoring odometry failures can lead to erratic tracking behavior, since tracking is performed in 3D world coordinates. As tracking is in turn used to constrain VO, those errors may be amplified further. Similarly, the feedback from object tracking as a spatial prior to detection can potentially lead to resonance effects if false detections are integrated into an increasing number of incorrect tracks. Finally, our system's reliance on a ground plane to constrain object detection may lead to incorrect or dropped detections if the ground plane is wrongly estimated. As our system relies on the close interplay between all components, each of these failure modes could in the worst case lead to system instability and must be addressed.

**Visual Odometry.** To detect visual odometry failures, we consider two measures: firstly the deviation of the calculated camera position from the smoothed filter estimate and secondly the covariance of the camera position. Thresholds can be set for both values according to the physical properties of the moving platform, *i.e*. its maximum speed and turn rate. Note that an evaluation of the covariance is only meaningful if based on rigid structures. Moving bodies with well distributed points could yield an equally small covariance, but for an incorrect position. When dynamic objects are disregarded, the covariance gives a reliable quality estimate for the feature distribution.

Note: while it would be possible to constrain pose sampling during RANSAC, this would not alleviate the problem of correspondences on moving objects. We therefore also use semantic information from object tracking, as will be explained in Section 4.

**Object Tracking.** Failure detection and correction is accomodated by the construction of our tracking approach. Instead of relying on a Markov assumption for propagating tracks over time, this approach builds upon a model selection framework to optimize tracks over a larger temporal window, similar to [16]. At each time instant, the tracking module explores a large number of concurrent track hypotheses in parallel and selects the most promising subset. This means that it can compensate for tracking errors and recover temporarily lost tracks.

**Object Detection and Ground Plane Estimation.** These two components are kept stable by the continuous use of additional information from stereo depth. Depth measurements are employed both to support the ground plane estimate and to verify object detections. Thus, false predictions from the tracking system are corrected. In addition, the temporal prior $P(\boldsymbol{\pi}_{t-1})$ smoothes noisy measurements.

## 4. Cognitive Feedback to Visual Odometry

Standard algorithms for visual odometry assume a predominantly static scene, treating moving objects just the same as incorrect correspondences. Most systems use robust hypothesize-and-test frameworks such as RANSAC or Least-Median-of-Squares for removing such outliers. Recently, some multi-body Structure-from-Motion systems have been demonstrated on realistic video scenes [18]. However, those remain constrained to rigidly moving bodies such as cars, and require a sufficient number of interest points for each model. We show that the use of basic scene understanding can effectively stabilize visual odometry by constraining localization efforts on regions that are likely to

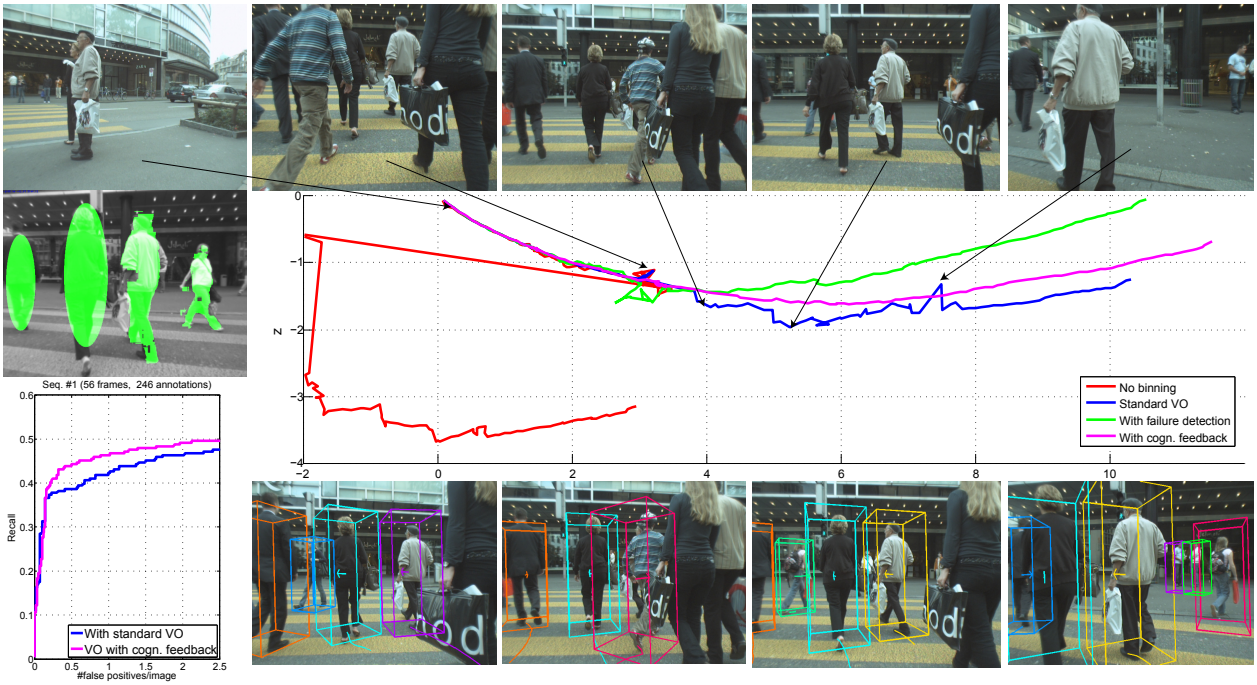



Figure 4. Trajectory estimation of our system with and without cognitive feedback. (Top) A few frames of a difficult sequence. (Middle) (Left) Example confidence map used by the cognitive feedback to adapt corner sampling. (Right) Trajectory estimates. (Bottom) (Left) Recall/false positives for single detections with standard VO and VO using feedback. (Right) Tracking results with cognitive feedback. As can be seen, the proposed feedback greatly stabilizes the egomotion estimation and leads to stable tracks. (Figure best viewed in color)

be part of the rigid scene.

In order to underline the importance of the proposed integration, consider the scene shown in Fig. 4, taken from one of our recordings. Here, our mobile platform arrives at a pedestrian crossing and waits for oncoming traffic to pass. Several other people are standing still in its field of view, allowing standard VO to lock onto features on their bodies. When the traffic light turns green, everybody starts to move at the same time, resulting in extreme clutter and blotting out most of the static background. Since most of the scene motion is consistent, VO fails catastrophically (as shown in the red curve). This is of course a worst-case scenario, but it is by no means an exotic case — on the contrary, situations like this will often occur in practical outdoor applications (we present another example in the results section).

Spatial binning for feature selection (as promoted in [21, 30]) improves the result in two respects: firstly, spatially better distributed features per se improve geometry estimation. Secondly, binning ensures that points are also sampled from less dominant background regions not covered by pedestrians. Still, the resulting path (shown in blue) contains several physically impossible jumps. Note here that a spike in the trajectory does not necessarily have to stem from that very frame. If many features on moving objects survive tracking (*e.g.* on a person's torso), RANSAC can easily be misled by those a few frames later. Failure detection using the Kalman filter and covariance analysis (in green) reduces spiking further, but is missing the semantic information that can prevent VO from attaching itself to moving bodies. Finally, the magenta line shows the result using our complete system, which succeeds in recovering a smooth trajectory. Detection performance improves as well (bottom row, left): when measuring recall over false positives per image (FPPI) on single detections, recall increases by $6\%$ at $0.5$ FPPI when using the cognitive feedback. A few example frames are displayed in the bottom right, confirming that the VO estimates also result in more plausible and correct tracks.

The intuition behind our proposed feedback procedure is to remove features on pedestrians using the output of the object tracker. For each tracked person, we mask out her/his projection in the image. If a detection is available for the person in the current frame, we use the confidence region returned by the object detector (in our case an ISM [17]). If this region contains too large holes or if the person is not detected, we substitute an axis-aligned ellipse at the person's predicted position. A few example masks are shown in Fig. 4(middle row, left).

Given this object mask for a frame, we now adapt the sampling of corners. In order to ensure a constant number of features, we adapt the number of corners to look for in bin $i$ as follows

$$N_i = \frac{N_{org}(1 - p_o^{(i)})}{1 - \sum_i p_o^{(i)} / n_{bins}}, \tag{2}$$

| Seq. | # Frames | Dist | VO Inliers | |
|---|---|---|---|---|
| | | | Standard | w/ Feedback |
| #1 | 220 | 12m | 30% | 40% |
| #2 | 1'208 | 120m | 27% | 33% |
| #3 | 999 | 110m | 39% | 41% |
| #4 | 950 | 82m | 40% | 45% |
| #5 | 840 | 43m | 12% | 32% |

Table 1. Overview of used test sequences (frames, approx. travelled distance), along with average percentage of VO inliers. The cognitive feedback consistently improves the inlier ratio, especially in highly dynamic scenes (#1,#5).

with $N_{org}$ the originally desired number of corners per bin, $p_o^{(i)}$ the percentage of occluded pixels in bin $i$, and $n_{bins}$ the number of bins (in our case $n_{bins} = 100$). Corners are only sampled from unmasked pixels. Even with imperfect segmentations, this approach improves localization by sampling the same number of feature points from regions where one is more likely to find rigid structure.

While this pedestrian crossing example represents a worst-case scenario for VO, the beneficial effect of the proposed cognitive feedback can also be seen in less extreme cases. For instance, for Seq. #2 (see Table 1), estimated walking speed *before* Kalman filtering only spikes 15 instead of 39 times (in 1'200 frames) above a practical upper limit of 3 meters/second when using cognitive feedback. This means that the fallback options are used less frequently, and in turn that dead reckoning and hence introduction of drift are reduced. By optimizing the sampling locations, the feedback generally improves the feature distribution and thus also the number of inliers. This can be seen in Table 1 for several test sequences (the other sequences will be introduced below) and also has practical consequences regarding speed: for RANSAC, the number of iterations $M = \log(1-p)/\log(1-(1-\eta)^3))$ is in our case only controlled by the percentage of expected outliers $\eta$ [12]. The desired probability $p = 0.99$ of an outlier-free solution and the number of points needed to produce a hypothesis is constant for our problem. In most of our examples, the increased number of inliers translates to about half the necessary samples.

## 5. Results

In order to evaluate our vision system, we applied it to another four sequences, showing strolls through busy pedestrian zones. In total, we thus used 4'200 frames. All sequences were acquired with similar mobile platforms and consist of two synchronized video streams recorded at 15fps.[1] The first such sequence ("Seq. #2") extends over 1'208 frames. We manually annotated all visible pedestrians in every fourth frame, resulting in 1'894 annotations. As additional baseline, we include a sequence ("Seq. #3") from [9] with 5'193 annotations in 999 frames. Finally, as a demonstration of the breaking point of our system, we show two other sequences with fast turns ("Seq. #4") and an extreme number of moving pedestrians ("Seq. #5"). For testing, all system parameters are kept the same throughout all sequences.

We quantitatively measure performance by comparing generated and annotated bounding boxes and plotting recall over false positives per image. Fig. 5 shows performance plots for Seqs. #2 and #3. Besides raw detector output ("Detector"), we consider two additional baselines: firstly, we emulate the system of [16] by an offline step of running VO, fitting ground planes through wheel contact points, and then running our tracker without depth-map information ("Tracker baseline"). Secondly, for Seq #3, we use the baseline from [9] ("GM baseline"). Even though our proposed system needs a few frames before initializing a track (losing recall) and even though it reports currently occluded hypotheses (increasing false positives), both baselines are clearly outperformed.

An interesting observation is the bad performance of the baseline tracker on Seq. #3. Here, the detector yields multiple hypotheses at different scales for many pedestrians. Due to the low camera placement, these cannot be disambiguated by the ground plane alone. Thus, misplaced detections generate wrong trajectories that in turn encourage bad detections, resulting in a very unstable system. Our system breaks this vicious circle by using depth information.

For Seqs.#2 and #5, Fig. 6 shows the driven trajectories overlayed on an aerial image using manually picked control points. For Seq. #5, we also show the trajectory obtained without cognitive feedback — as with the densely crowded Seq. #1, the VO system cannot cope with the complexity of the scene without semantic information.

We manually evaluated tracking performance in $450$ frames of Seq. #2 using similar criteria as described in [28] (Tab. 2). We consider the number of pedestrians, the number of trajectories (if a pedestrian is occluded for $> 10$ frames, we count a new trajectory), the number of mostly hit trajectories ($> 80\%$ covered), mostly missed trajectories ($< 20\%$ covered), the number of false alarms, and the number of ID switches (meaning the tracker drifts from one person to another). On average, $75\%$ of a trajectory are covered by the tracker. The missed trajectories belong mostly to pedestrians at smaller scales, and to two children that do not fit the size prior. Example tracking results for Seq. #2 are shown in Fig. 8. Our system's ability to track through occlusion is demonstrated in the top row: please note how the woman entering from the left has temporarily occluded almost every part of the image. Still, the tracker manages to pick up the trajectory of the woman on the right again (in red). Fig. 9 shows additional tracking results for Seqs.#3, #4, and #5. Again, our system manages to produce long

[1] Paper website: `http://vision.ee.ethz.ch/~aess/cvpr2008`

| # Persons | # Traj. | Mostly Hit | Mostly Missed | False Alarms | ID Switches |
|---|---|---|---|---|---|
| 30 | 45 | 36 | 6 | 4 | 2 |

Table 2. Quantitative tracking results for part of Seq. #2. (see text)

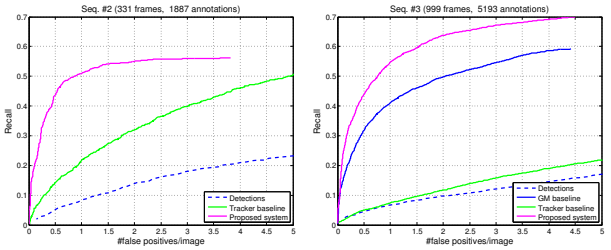



Figure 5. Single-frame detection performance on Seq. #2 and #3.

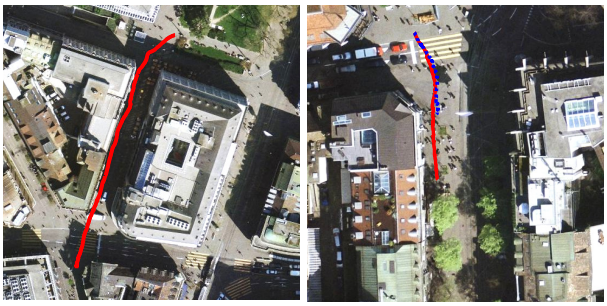


Figure 6. Travelled paths for Seqs. #2 and #5, overlaid on aerial images. (Right: result without feedback in blue, dashed)

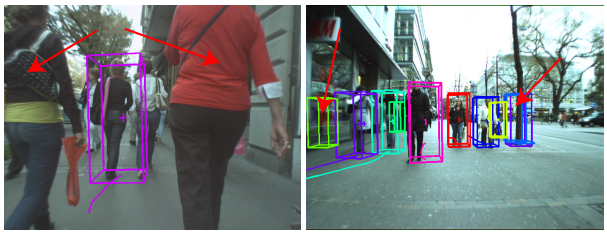


Figure 7. Typical false negatives (pedestrians at large scales) and false positives (reflections, trees) of our system.

and stable tracks in complex scenarios with a considerable degree of occlusion. In the top row, a pedestrian gets successfully tracked on his way around a few standing people and two pedestrians are detected at far distances. The middle row again demonstrates tracking through major occlusion. Finally, the bottom row shows an example scenario from Seq. #5 with many pedestrians blocking the camera's field-of-view. As mentioned above, scenes of this complexity are at the limit of what is currently possible with our system. The solution of such scenarios still needs further work. This will also address typical failures, as seen in Fig. 7.

## 6. Conclusion

In this paper, we have presented an integrated system for multi-person tracking from a mobile platform. The different modules (here, appearance-based object detection, depth estimation, tracking, and visual odometry) were integrated using a set of feedback channels. This proved to be a key factor in improving system performance. We showed that special care has to be taken to prevent system instabilities caused by erroneous feedback. Therefore, a set of failure prevention, detection, and recovery mechanisms was proposed. The resulting system can handle very challenging scenes. Still, there is some way to go before it becomes deployable in a real-world application. The individual components still need to be optimized further, both with respect to speed and performance. For instance, very close pedestrians, with only parts of their torso visible, are often missed by the current detector. A graceful degradation in form of image-based tracking might be a possibility to prevent system breakdown in such cases. Further combinations with other modules, such as world knowledge inferred *e.g.* from map services, provide other exciting feedback possibilities that we plan to investigate in the future.

**Acknowledgments.** This project has been funded in parts by Toyota Motor Corporation/Toyota Motor Europe and the EU projects DIRAC (IST-027787) and HERMES (IST-027110).

## References

[1] H. Bay, T. Tuytelaars, and L. van Gool. SURF: Speeded up robust features. In *ECCV*, 2006.

[2] J. Berclaz, F. Fleuret, and P. Fua. Robust people tracking with global trajectory optimization. In *CVPR*, 2006.

[3] C. Bibby and I. Reid. Simultaneous localisation and mapping in dynamic environments (SLAMIDE) with reversible data association. In *RSS*, 2007.

[4] E. Borenstein and S. Ullman. Class-specific, top-down segmentation. In *ECCV*, 2002.

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

[6] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In *ECCV'06*.

[7] A. Davison. Real-time simultaneous localization and mapping with a single camera. In *ICCV*, 2003.

[8] E. Eade and T. Drummond. Scalable monocular SLAM. In *CVPR*, 2006.

[9] A. Ess, B. Leibe, and L. van Gool. Depth and appearance for mobile scene analysis. In *ICCV*, 2007.

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

[11] D. Hähnel, R. Triebel, W. Burgard, and S. Thrun. Map building with mobile robots in dynamic environments. In *ICRA'03*.

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

[13] D. Hoiem, A. A. Efros, and M. Hebert. Putting objects in perspective. In *CVPR*, 2006.

[14] O. Lanz. Approximate bayesian multibody tracking. *PAMI*, 28(9):1436–1449, 2006.

[15] B. Leibe, N. Cornelis, K. Cornelis, and L. van Gool. Dynamic 3d scene analysis from a moving vehicle. In *CVPR'07*.

[16] B. Leibe, K. Schindler, and L.van Gool. Coupled detection and trajectory estimation for multi-object tracking. In *ICCV'07*.

[17] B. Leibe, E. Seemann, and B. Schiele. Pedestrian detection in crowded scenes. In *CVPR*, 2005.

[18] T. Li, V. Khallem, D. Singaraju, and R. Vidal. Projective factorization of multiple rigid-body motions. In *CVPR'07*.

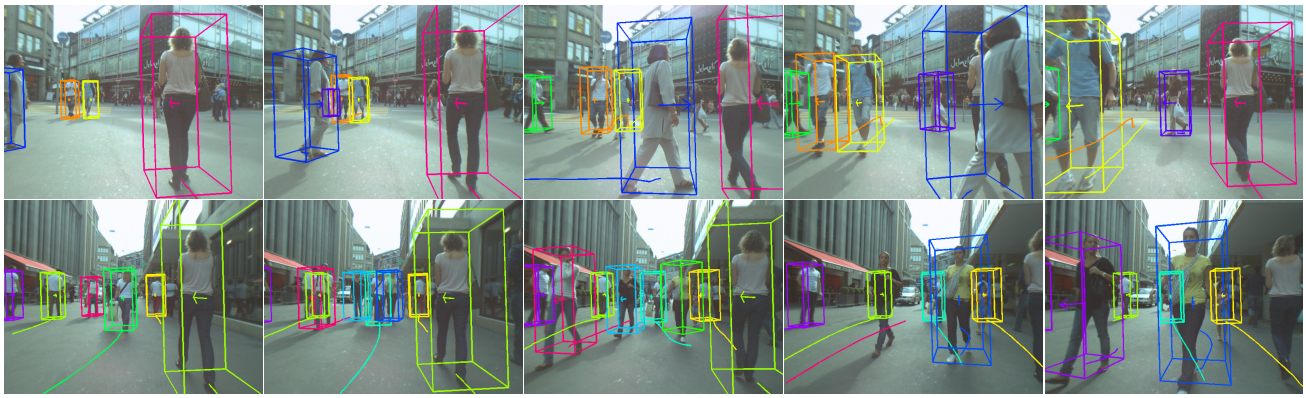


Figure 8. Two exemplary subsequences obtained using our mobile vision system in Seq.#2. Note the long trajectories and ability of the tracker to handle temporary occlusions in complex scenarios.

Figure 9. Selected tracking results for Seqs. #3 , #4, and #5.


[19] A. Makadia, C. Geyer, S. Sastry, and K. Daniilidis. Radon-based structure from motion without correspondences. In *CVPR*, 2005.

[20] D. Nistér. A minimal solution to the generalised 3-point pose problem. In *CVPR*, 2004.

[21] D. Nistér, O. Naroditsky, and J. R. Bergen. Visual odometry. In *CVPR*, 2004.

[22] K. Okuma, A. Taleghani, N. de Freitas, J. Little, and D. Lowe. A boosted particle filter: Multitarget detection and tracking. In *ECCV*, 2004.

[23] K. E. Ozden, K. Schindler, and L. van Gool. Simultaneous segmentation and 3d reconstruction of monocular image sequences. In *ICCV*, 2007.

[24] M. S. P. Yan, S. M. Khan. 3d model based object class detection in an arbitrary view. In *ICCV*, 2007.

[25] S. Se, D. Lowe, and J. Little. Global localization using distinctive visual features. In *IROS*, 2002.

[26] O. Tuzel, F. Porikli, and P. Meer. Human detection via classification on Riemannian manifolds. In *CVPR*, 2007.

[27] P. Viola, M. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. In *ICCV*, 2003.

[28] B. Wu and R. Nevatia. Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet part detectors. *IJCV*, 75(2):247–266, 2007.

[29] K. Yamaguchi, T. Kato, and Y. Ninomiya. Vehicle ego-motion estimation and moving object detection using a monocular camera. In *ICPR*, 2006.

[30] Z. Zhu, T. Oskiper, O. Naroditsky, S. Samarasekera, H. Sawhney, and R. Kumar. An improved stereo-based visual odometry system. In *Perf. Metr. for Intell. Sys. (PerMIS)*'06.