DIRAC
Detection and Identification of Rare Audiovisual Cues

Insperata accident magis saepe quam quae speres.
(Things you do not expect happen more often than
things you do expect) Plautus (ca 200(B.C.)

Project no: 027787

DIRAC

Detection and Identification of Rare Audio-visual Cues

Integrated Project
IST - Priority 2

DELIVERABLE NO: D3.5

DYNAMIC 3D SCENE ANALYSIS USING COGNITIVE LOOPS.

Recognition of Both Static and Dynamic Objects Benefitting from and Feeding back to 3D Reconstruction.

Date of deliverable: 31.06.2007
Actual submission date:

Start date of project: 01.01.2006 Duration: 60 months

Organization name of lead contractor for this deliverable: **ETH Zurich**

Revision [draft, 1, 2, …]

Project co-funded by the European Commission within the Sixth Framework Program (2002-2006)		
Dissemination Level		
PU	Public	
PP	Restricted to other program participants (including the Commission Services)	X
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

D3.5 Dynamic 3D Scene Analysis using Cognitive Loops

Recognition of both Static and Dynamic Objects Benefiting from and Feeding back to 3D Reconstruction

Eidgenoessische Technische Hochschule Zuerich (ETHZ)
Katholieke Universiteit Leuven (KUL)
Czech Technical University Prague (CTU)
Hebrew University Jerusalem (HUJI)

Abstract:

This deliverable presents methods for dynamic visual scene analysis in very challenging outdoor scenarios. Starting from a mobile sensing setup very similar to the AWEAR application scenario, we combine many different vision components and interface them to collaborate in cognitive feedback loops. Specifically, we integrate Structure-from-Motion (SfM), dense reconstruction, appearance-based object detection, trajectory estimation, and multi-object tracking into a combined system and show how those components can benefit from each other's continuous support. We apply the resulting system to challenging video sequences of strolls through busy pedestrian zones and demonstrate that the proposed integration makes it possible to deliver stable scene analysis in such difficult settings. The presented results thus constitute a major step towards the envisioned DIRAC application scenario of a cognitive walking aid.

Table of Contents

1. Introduction

DIRAC's envisioned application scenario is the development of an audio-visual device that would be either stationary surveying a room or mounted on a walking aid as a portable assistant for elderly care. It would be able to learn daily routines of its user, help him/her navigate around the environment, identify unfamiliar and/or potentially dangerous situations, and issue alarms when such a situation is encountered. In the context of this application scenario, it is important to visually analyze the environment by processing camera input.

The DIRAC goal of detecting rare audio-visual events in such unconstrained scenarios necessitates an advanced level of scene understanding. In order to detect what is rare, one first needs to learn what is normal. In order to build up expectations about how other people will act, one first needs to learn their typical behaviors. This of course requires to detect them in the first place and to maintain an association of their identities over a certain time span. Context also plays a vital role for scene understanding. In order to interpret other people's observed actions, it is necessary to know about the type of environment they act in, their relative positions in this environment, as well as about other objects and outside influences that act upon them. As we live in a 3D world, it is only natural that true scene understanding should incorporate a notion of this 3D space and its associated constraints. Building a system with such capabilities has been a far-end goal of scene understanding since the 1970ies, but so far the sheer complexity of many real-world scenes has often stymied progress in this direction.

However, this does not mean that such an endeavor would be hopeless. Computer vision has made significant progress in recent years, and many of its individual disciplines have advanced to a state where algorithms are becoming applicable to real-world tasks. In this deliverable, we therefore explore how different visual components can be used for dynamic visual scene analysis in complex outdoor scenes and how they can be integrated in order to support each other. Specifically, we combine Structure-from-Motion (SfM), dense reconstruction, appearance-based object detection, trajectory estimation, and tracking into a combined system. We build upon the system for bottom-up 3D reconstruction described in D3-1 and extend it with the other components. As our results will show, the proposed integration makes it possible to deliver stable scene analysis performance in scenes of previously infeasible complexity.

A central component in the proposed integration is the concept of cognitive feedbacks. 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. In the work described in this deliverable, we incorporate several such feedback paths to interface the various components: from Structure-from-Motion (SfM) to recognition, from recognition to depth estimation, from tracking to recognition, and from both recognition and tracking back to SfM. As we will show, those cognitive loops are an important ingredient to delivering robust performance in very challenging real-world scenarios.

At the current stage, the work presented in this deliverable is still based on perspective cameras. This is largely a consequence of our need to capture data at sufficiently high frame rates to enable tracking, which was not yet possible with the first prototype omni-directional setup developed in WP1. Based on the results obtained and the experiences made in Y2, this work will be extended to an omni-directional scenario for use with the next AWEAR prototype, which will allow data capture at higher frame rates. Results for this omni-directional scenario will then be presented in the M36 deliverable D3-7.

Figure 1: The three prototype setups used to test out the different stages of our mobile vision system. (left) The car setup allows the simplest scenario with relatively stable platform motion, wide baseline, and high camera placement for optimal scene coverage. This is made more difficult when switching to the child stroller setups (middle, right). Here, camera motion is far more noisy; the setups allow only for a smaller baseline; and the lower camera placement makes it harder to observe a sufficient portion of the scene. The third prototype (right) is already close in size to the envisioned AWEAR application scenario of an intelligent walking aid.

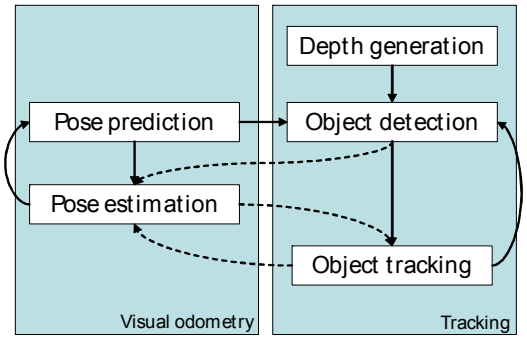


Figure 2. Overview of the functional relationship between the different components of our system.

2. System Overview

In the course of the work leading to this deliverable, we created a series of prototype data acquisition setups of steadily increasing difficulty, as shown in Figure 1. The first setup was built on a car, which provided stable camera motion, as well as a relatively wide baseline of 1.5m. First results from this setup were already presented in D3-1. In order to obtain data that is more similar to the intelligent walking aid application envisioned in the AWEAR scenario, we scaled down this setup to two prototypes using a child stroller as mobile base (Figure 1 middle and right). This considerably smaller setup results in far more difficult sensing conditions. Its size constraints only allow for a camera baseline of about 40cm, which strongly limits the achievable accuracy for stereo depth estimation. In addition, the lower camera placement implies that pedestrians will often block a much larger part of the vehicle's field-of-view, making it harder to find static scene features for SfM. Finally, the smaller wheel size and uneven ground result in numerous bumps in the effective camera trajectory, which disturbs SfM estimation. Altogether, those effects pose significant challenges for all stages of our approach to achieve robust system performance.

We address those challenges by integrating and closely coupling several different vision modalities: SfM, dense stereo estimation, appearance-based object detection, and detection-based tracking. Figure 2 shows an overview how those components interact in our system. For each frame pair of the input video streams, we first estimate the camera location and scene geometry. This part is based on the bottom-up reconstruction framework developed in D3-1 and will be briefly summarized in Section 3. In parallel, we perform appearance-based object detection on both input video streams in order to detect other traffic participants (pedestrians, cars, bicyclists, etc.) in the camera vehicle's field of view. An automatically estimated ground plane from the reconstruction pathway is used in order to constrain object detection to

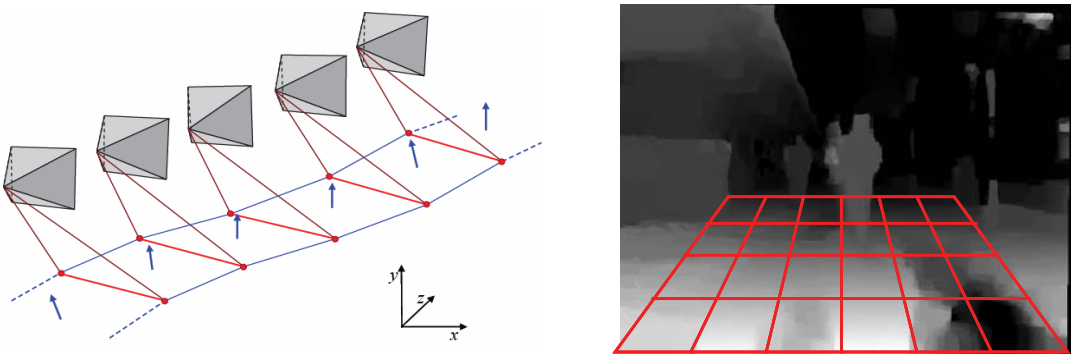
promising image locations, which considerably improves recognition performance. In addition, recognition is supported by dense stereo depth measurements, which are used to verify detections and localize them more accurately. This is described in Section 4. Using the estimated camera location from SfM, detection bounding boxes are converted into world coordinates, which are integrated over time in order to estimate physically plausible object trajectories, as explained in Section 5. In Section 6, we show how those last two steps, detection and trajectory estimation, can be coupled into a combined optimization problem in order to further improve robustness. Finally, we close the feedback cycle by supplying SfM with information from detection and tracking in order to obtain more accurate localization estimates (Section 7). The final integrated system is then presented in Section 8. Experimental results and references to more detailed explanations in the appended research papers will be given throughout this report.

3. Online Scene Geometry Estimation

Our approach makes the following two uses of automatically estimated scene geometry information. First, it employs the knowledge about the scene's ground plane in order to restrict possible object locations during detection. Second, a camera calibration obtained from SfM allows us to integrate individual detections over time in a world coordinate frame and group them into trajectories over a spacetime window. As we will show in Section 5, this makes it possible to perform 3D object tracking from a moving vehicle.

We investigated two different methods for estimating the current ground plane from visual data. In a joint ETH-KUL work [1, 4, both appended to this deliverable], we used the camera trajectory obtained from SfM in order to infer the ground plane location. Figure 3(a) visualizes this procedure. Taking as input the estimated camera locations from SfM, our approach assumes a rigid vehicle geometry to reconstruct the wheel contact points on the road surface at the time the images were taken. By connecting those wheel contact points into trapezoidal patches, we can obtain local normal measurements, which are further smoothed over a larger spatial window. Empirically, averaging the normals over a length of 3m (or roughly the wheel-base of the vehicle) turned out to be optimal for a variety of cases [1]. The obtained ground plane is then extrapolated to the distance in order to provide an estimate for the current frame.

This approach performs well in practice, but it requires a small temporal look-ahead. The reason is that SfM estimates for forward-looking cameras are typically quite noisy because of numerical instabilities. They therefore need to be corrected by bundle adjustment over a certain (small) temporal window. This means that the SfM results will only be stable after the current window has been processed. Depending on data capture frame rate, the necessary minimum window size for bundle adjustment is typically between 3-6 frames, resulting in on average half that many frames before the results become available. In addition, abrupt changes of the ground plane (such as when driving over a speed bump) may only be reflected in the ground plane estimate after such a location has been passed by the vehicle's front wheel. Taken together, those two effects result in an effective delay before the SfM measurements can be safely used, which is obviously undesirable for online applications.



(a) (b)

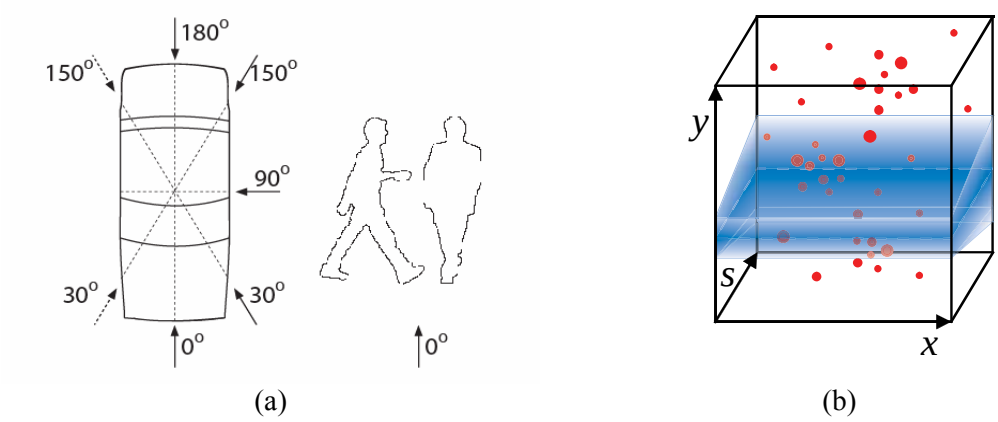
Figure 3. The two approaches we employ for obtaining ground plane information in this deliverable. (a) Using the camera locations estimated by Structure-from-Motion for past frames, we reconstruct points on the road surface our vehicle traveled over and extrapolate the resulting ground surface to future frames. (b) We employ depth measurements from dense stereo in order to verify the consistency of ground plane hypotheses using a robust least-median-of-squares estimator.

For this reason, we have also investigated a different method for estimating the ground plane described in [3, appended to this deliverable] and depicted in Figure 3(b). This approach computes a dense stereo depth map for every frame pair and then verifies how consistent a given ground plane hypothesis is with the observed depth measurements using a robust least-median-of-squares estimator. This method works very well if there is enough texture on the ground surface to yield depth measurements. In addition, it is entirely causal, meaning it does not require a temporal look-ahead, but only relies on information from the current frame. However, it becomes less reliable in crowded scenes with many pedestrians at close ranges, where less of the ground is visible. In the following section, we therefore couple object detection and ground plane estimation in a common graphical model that allows both search problems to be solved together.

4. Object Detection

The object recognition system is based on a battery of single-view, single-category object detectors. In this work, we use the ISM detector from [7]. This approach lets local features, extracted around interest regions, vote for the object center in a 3-dimensional Hough space, followed by a top-down segmentation and verification step. For our application, we use the robust multi-cue extension from [8, D1-2], which integrates local Shape Context descriptors [11] computed at *Harris-Laplace*, *Hessian-Laplace*, and *Difference-of-Gaussian* interest regions [10, 11]. Those features have been selected based on an extensive evaluation performed as part of deliverable D1-2 in WP1. For a more detailed description, we refer to this deliverable and the corresponding publication [8].

In order to capture the varying object appearance from multiple viewpoints, we use 5 single-view detectors for the different car orientations and one additional detector for pedestrians (see Figure 4(a)). We do not differentiate between pedestrians and bicyclists here, as they are often indistinguishable from a distance and our detector responds well to both categories. In the rest of this section, we present two different approaches for integrating the resulting object detections with the estimated scene geometry.



(a) (b)

Figure 4. (a) Training viewpoints used for cars and pedestrians in our experiments. (b) The estimated ground plane significantly reduces the search space for object detections to a corridor in the (*x,y,scale*) volume.

4.1 Integration of Ground Plane Constraints

As already described in D3-1, the estimated scene geometry can be used to significantly constrain the search space for object detection to a corridor in the (*x,y,scale*) volume. This effect is shown in Figure 4(b). In the first approach presented here, we employ this strategy in order to achieve significant detector speed-ups and to filter out many false positives.

Given a ground plane estimate from SfM, we limit object detection to the above-mentioned (*x,y,scale*) corridor. By projecting a ray through the base point of each detection bounding box and intersecting it with the ground plane, we can localize each detection hypothesis in 3D and estimate its real-world size. By comparing this size estimate to a learned distribution of real object sizes, we can then express the likelihood for a real-world object being present in this location given the observed image evidence. Details how this is formally done can be found in the appended paper [1]. The resulting verification procedure considerably improves object detection performance to a level where the obtained detections can be used for real-world applications [1, 2, 4].

4.2 Simultaneous Object Detection and Ground Plane Estimation

However, a potential problem with the approach described in the previous section is that it relies on a hard decision about the ground plane location at the very beginning of the sensing process. If this decision is wrong, potential object hypotheses will not be sampled from certain image regions, which may result in missing detections. In this section, we therefore present a method to avoid this hard decision. Using input from pedestrian detection and dense stereo, we want to jointly estimate both scene geometry and object locations. This is achieved by integrating the different cues in a graphical model, which allows inference in all directions.

Figure 5 shows the graphical model we use for inference over object hypotheses. It is based on three different kinds of inputs: ground plane measurements $\pi_{\mathcal{D}}$ based on the dense stereo depth map described in Section 3; object hypotheses $o_1,...,o_n$ from an appearance-based object detector applied to the input image I; and a depth verification $d_1,...,d_n$ flag that checks if the hypothesized object is consistent with the observed depth distribution $\mathcal{D}$ in the same image location. The model is parameterized over the hidden variable π defining the ground plane parameters. Please refer to the appended paper [3] for details of the implementation.

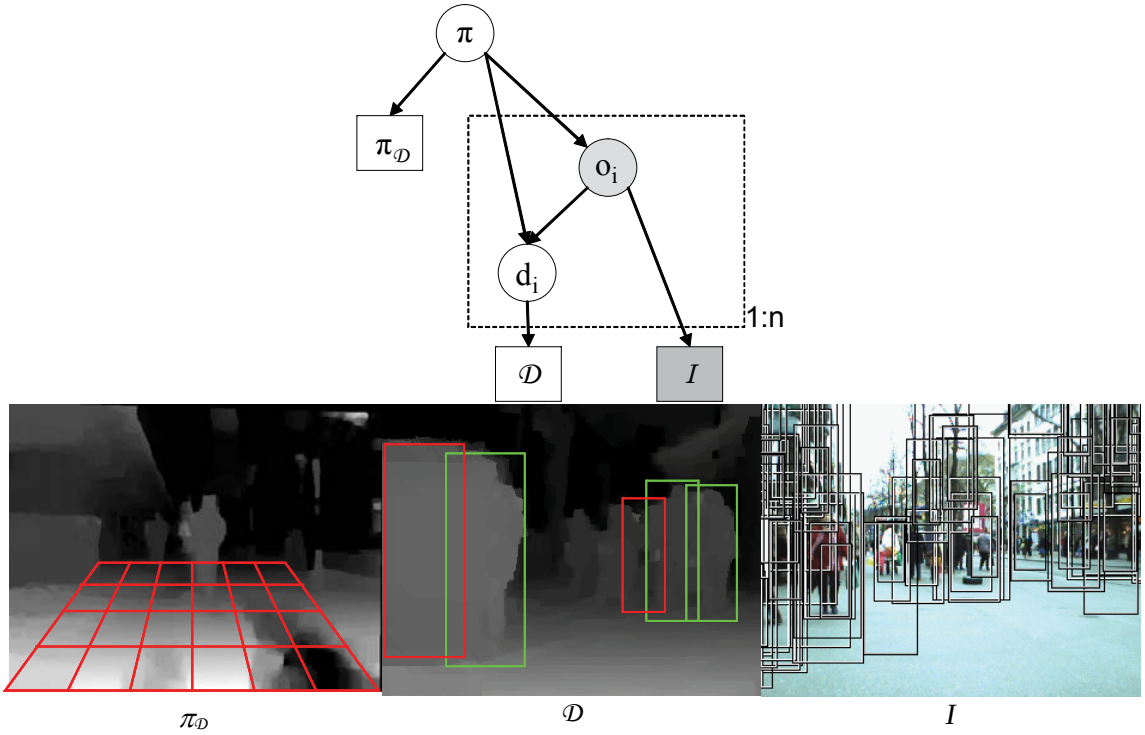


Figure 5. (top) The Graphical Model used for simultaneous object detection and ground plane estimation. (bottom) The three input cues for this model: (left) ground plane measurements based on the dense stereo depth map; (middle) depth verification for each object hypothesis; (right) output of the appearance-based object detector on image *I*.

As stated above, the interesting property of this graphical model is that it can perform inference in both directions. Thus, if we observe a largely empty scene, the ground plane measurements can constrain object detection to promising object locations. If, on the other hand, we encounter a crowded scene with many people appearing at close ranges and only a small fraction of the ground surface being visible, successful object detections can in turn constrain the ground plane location. Thus, our approach implements another cognitive loop in which detection and geometry estimation closely collaborate towards a common goal.

Figure 6 presents example detection results using this approach on several challenging video sequences from busy pedestrian zones in Zurich. The test data was acquired using the child stroller setup shown in Figure 1(b), which is already quite close to the target scenario of the final AWEAR demonstrator, just at this point still with perspective instead of omni-directional cameras. As the results in Figure 6 show, our approach achieves very good detection results in such difficult scenes and successfully detects many of the pedestrians visible there with only few false positives. A detailed experimental evaluation of the different system components and a comparison to other state-of-the-art detectors can be found in [3, appended], confirming this result also quantitatively.

Figure 6. Example pedestrian detection results of our approach from Section 4.1 on three different test sequences. These results confirm that our simultaneous object detection and ground plane estimation method achieves very good detection performance with only few false positives (shown in red).

5. Spacetime Trajectory Estimation

The results from the previous section confirmed that our object detection framework reaches a suitable performance level for real-world applications. In this section, we now introduce an algorithm to integrate the detections over time and group them into object trajectories. The framework from the previous deliverable D3-1 could only perform such an integration for static scene objects, such as parked cars. In this section, we extend the approach described there to also handle dynamic objects and track their motion over time.

The key idea behind our proposed tracking estimation procedure is that we accumulate object detections in a spacetime observation volume, which we subsequently analyze in order to find physically plausible object trajectories. Each detection is entered into this spacetime volume with its ground plane location and time stamp. If we now consider a static scene object, we would expect to see the corresponding detections to form a vertical trajectory. For a moving object, the trajectory will be tilted according to the motion speed. The basic idea of our approach is now to collect a large set of candidate trajectories and then apply model selection in order to choose the optimal subset that best explains the observed data.

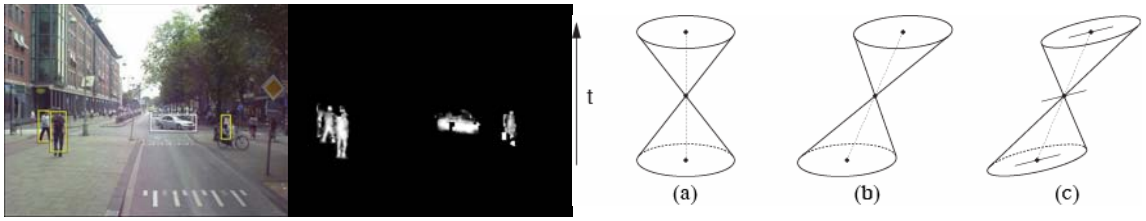


Figure 7. (left) Detections and corresponding segmentations used to learn the object-specific color model. (right) Visualization of example event cones for (a) a static object with unknown orientation; (b) a holonomically moving object; (c) a non-holonomically moving object.

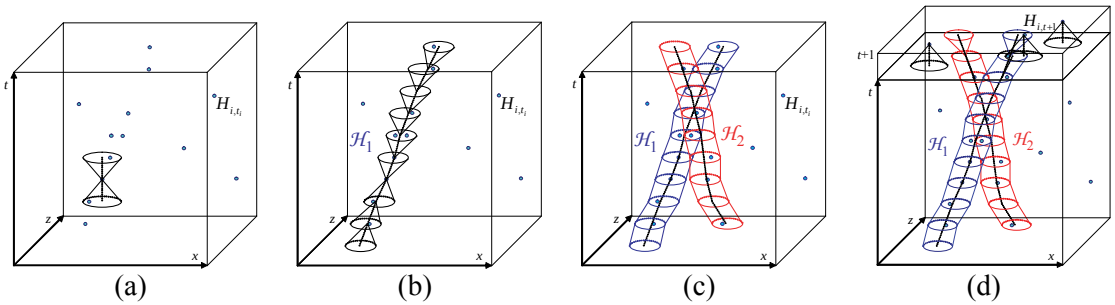


Figure 8. Visualization of the trajectory growing procedure. (a) Starting from an observation, we collect all detections that fall inside its event cone in the adjoining time steps and evaluate them under the trajectory model. (b) We adapt the trajectory based on inlier points and iterate this process both forward and backward in time. (c) This results in a set of candidate trajectories, which are passed to the hypothesis selection stage. (d) For efficiency reasons, trajectories are not built up from scratch at each time step, but are grown incrementally.

Each trajectory is defined via an object-specific appearance model and a dynamic model. The appearance model is represented as an 8x8x8 color histogram computed over the top-down segmentation returned by the object detector (see Figure 7(left)). The dynamic model is an Extended Kalman Filter (EKF) that specifies the event cone of an object, i.e. the spacetime volume that it can physically reach from its current position given its maximal velocity and turn rate. Here, we assume different motion models for pedestrians and cars. For pedestrians, we assume holonomic motion on the ground plane, meaning that they can move without external constraints. For cars, we use the knowledge that they can only move in the direction of their main axis and only turn while moving by adopting a non-holonomic motion model. Example event cones for different cases are shown in Figure 7(right).

We thus search for plausible trajectories through the spacetime observation volume by linking up event cones, as shown in Figure 8. Starting from an observation $H_{i,t}$, we follow its event cone up and down the timeline and collect all observations that fall inside this volume in the adjoining time steps. Since we do not know the starting velocity $v_{i,t}$ yet, we begin with the case in Figure 7(a). In all subsequent time steps, however, we can reestimate the object state from the new evidence and adapt the growing trajectory according to the EKF equations. Although any single trajectory hypothesis is thus estimated by a bidirectional EKF, our approach goes beyond Kalman Filters in a very important respect: we are not restricted to tracking a single hypothesis. Instead, we start independent trajectory searches from all available observations (at all time steps) and collect the corresponding hypotheses. The final scene interpretation is then obtained by a global optimization criterion which selects the combination of trajectory hypotheses that best explains the observed data under the constraints that each observation may at most belong to a single object and no two objects may occupy the same physical space at the same time. This makes it possible to enforce physical exclusion constraints such that a pedestrian may not walk through a car and vice versa.

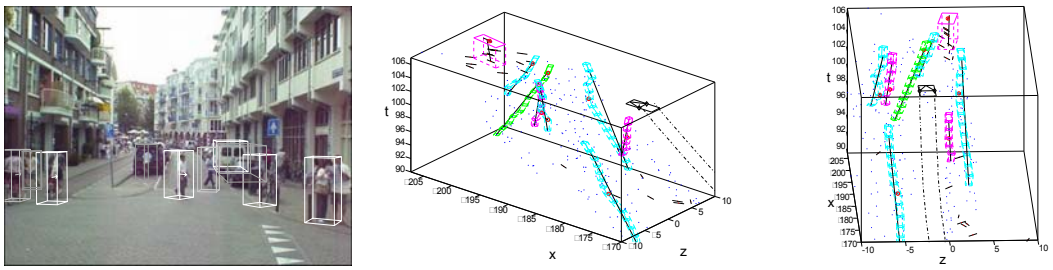


Figure 9. (left) Online 3D localization and trajectory estimation results of our system obtained from inside a moving vehicle. (right) Visualization of the corresponding spacetime trajectory estimates for this scene.

Our basic mathematical tool for this step is a model selection framework as introduced in [9] and adapted in [7]. This framework is based on a formulation based on the Minimum Description Length (MDL) principle. Briefly stated, it assigns each trajectory hypothesis a merit term, based on how well this hypothesis explains the observed data, and a base cost that penalizes more complex hypotheses. As each data point can only be assigned to a single model, overlapping hypothetical models compete for data points. This competition translates to interaction costs, which apply only if both hypotheses are selected and which are then subtracted from the score of the hypothesis combination. Leonardis et al. [9] have shown that if only pairwise interactions are considered, then the optimal set of models can be found by solving a Quadratic Boolean Problem (QBP). In this work, we use the multibranch gradient ascent method from [13] to solve the optimization problem.

This model selection procedure is performed after each frame (or frame pair) of the input sequence. As its result, we obtain a set of selected object trajectories that correspond to the best explanation of the current world state *given all evidence available up to now*. Each selected hypothesis comes with its own history, i.e. it reaches back into the past to observations that it could explain there. We can thus follow a trajectory back in time to determine where a pedestrian came from when he first stepped into view, even though no trajectory was selected for him back then. Figure 9 visualizes the estimated spacetime trajectories for such a case.

Another important property of our method is that when several trajectory hypotheses compete for the same data, a different hypothesis may be selected after each frame. We are thus no longer bound by a Markovian assumption, but effectively obtain a non-Markovian multi-object tracking framework that can compensate for previous errors and recover temporarily lost tracks. In contrast to previous multi-object tracking approaches such as Multi-Hypothesis Tracking (MHT) [12] and Joint Probabilistic Data Association Filters (JPDAFs) [6], which scale exponentially with the number of considered time steps, our approach can keep a significantly longer history, since only relatively few trajectory hypotheses need to be stored. This work was presented in a joint ETH-KUL paper [1] and was awarded the CVPR'07 Best Paper Award out of more than 1250 submissions and 352 accepted papers.

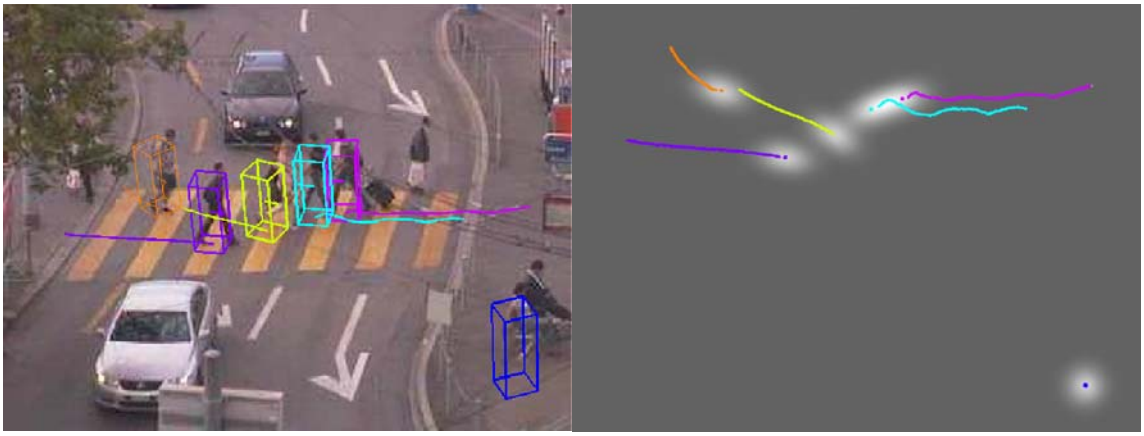


Figure 10. Influence of past trajectories on object detection in the coupled optimization problem. (left) a frame from one of our test sequences and detected pedestrians. (right) Top view of the detection prior for the next frame showing previous trajectories, predicted positions, and detection prior (brighter color means higher probability).

6. Coupled Detection and Tracking

In the two previous sections, we developed methods for detecting objects and for grouping those detections into spacetime trajectories. However, the two tasks are closely coupled: the merit of a putative trajectory depends on the number and strength of the underlying detections, while the merit of a putative detection depends on the current object trajectories, which impose a prior on object locations. These dependencies lead to further interactions between detections and trajectories. In the spirit of the work in DIRAC, we therefore want to close another cognitive loop by coupling the two processes and thus allow feedback from tracking to detection.

However, we have to keep in mind that the relationship between detections and trajectories is not entirely symmetric: trajectories ultimately rely on detections to be propagated, but new detections can occur without a trajectory to assign them to (e.g. when a new object enters the scene). We therefore need to enable detections to survive without contributing to an actual trajectory. In [2 ,4], we therefore developed a novel mechanism that couples detection and trajectory estimation in a jointly optimized combined QBP. We accomodate the above-mentioned asymmetry by introducing a list of virtual trajectories, one for each detection in the current image, which can explain detections whose score outweighs the base cost, but which are not claimed by any selected trajectory.

Thus, coupling has the following beneficial effects. First, it supports novel object detections that are consistent with existing trajectories. Existing trajectories effectively impose a prior on certain object locations, which raises the chance of generating novel detections there above the uniform background level (see Figure 10). Second, the evidence from novel detections aids trajectories with which those detections are consistent by allowing them to account the new information as support. As our experiments from [2, 4] show, the resulting feedback from tracking to detection improves total system performance and yields more stable tracks.

6.1 Identity Management

The hypothesis selection framework helps to ensure that all available information is used at each time step. However, it delivers an independent explanation at each time step and hence does not by itself keep track of object identities. Frame-to-frame propagation of tracked object identities is a crucial capability of tracking (as opposed to frame-by-frame detection).

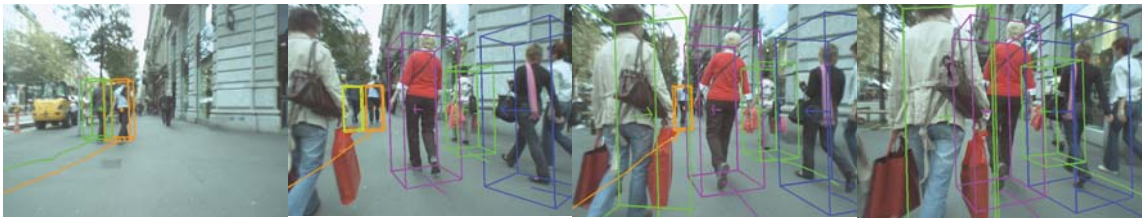


Figure 11. Example tracking results in very crowded scenes that are only made possible by feeding back information from object detection and tracking to Structure-from-Motion. Details of this feedback are described in deliverable D1-4.

Propagating identity is trivial in the case where a trajectory has been generated by extending one from the previous frame, where the hypothesis ID is simply passed on, as in a recursive tracker. However, one of the core strengths of the presented approach is that it does not rely on stepwise trajectory extension alone. If at any time a newly generated hypothesis provides a better explanation for the observed evidence than an extended one, it will replace the older version. However, in this situation the new trajectory should inherit the old identity, in order to avoid an identity switch.

In [2, 4], we have proposed a simple identity management strategy based on the associated data points: the identities of all selected trajectories are written into a buffer, together with the corresponding set of explained detections. This set is continuously updated as the trajectories grow. Each time a new trajectory is selected for the first time, this trajectory is compared to the buffer, and if its set of explained detections is similar to a buffer entry, it is identified as the new representative of that ID, replacing the older entry. If it does not match any known trajectory, it is added to the buffer with a new ID. This strategy works well in practice and yields stable person identities over long test sequences, as verified by quantitative experiments in the appended papers [2, 4].

7. Feedback to 3D Localization

The previous sections have shown that Structure-from-Motion can considerably help tracking in the envisioned AWEAR scenario by allowing our system to operate in 3D world coordinates. In this section, we will complete this interaction to a loop by also feeding back information from tracking to help visual odometry. As we will show, such a feedback is crucial for robust performance in crowded scenes, such as the one depicted in Figure 11. Here, 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. Such a scenario is very problematic for standard SfM algorithms, which assume a predominantly static scene and treat 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. We show that the use of basic scene understanding can effectively stabilize visual odometry by constraining localization efforts on regions that are likely to be part of the rigid scene.

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 in [5] demonstrate, the resulting system achieves robust multi-object tracking performance on very challenging video data. In this section, we will outline the basic ideas behind the proposed integration. For details of its implementation, we refer to deliverable D1-4 and [5].

7.1 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.

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, though for an incorrect position. With estimation based only on rigid structures, the covariance gives a reliable quality estimate for the feature distribution.

In case of a detected odometry failure, the filter estimate is used instead of the measurement; all scene points are cleared; and the Structure-from-Motion system starts anew. This allows us to keep the object 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, which can also be obtained through other cues such as GPS.

7.2 Feedback from Scene Understanding to SfM

The intuition behind our proposed procedure is to remove features on pedestrians using the output of the object tracker. For each tracked person, we mask out its 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. 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. Given this object mask for a frame, we then adapt the sampling of corners in order to ensure that a constant number of features is sampled from each unmasked image region. 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 useful structure. Together with the automatic failure detection, this results in considerably improved robustness of the SfM subsystem, as our experiments in [5] demonstrate.

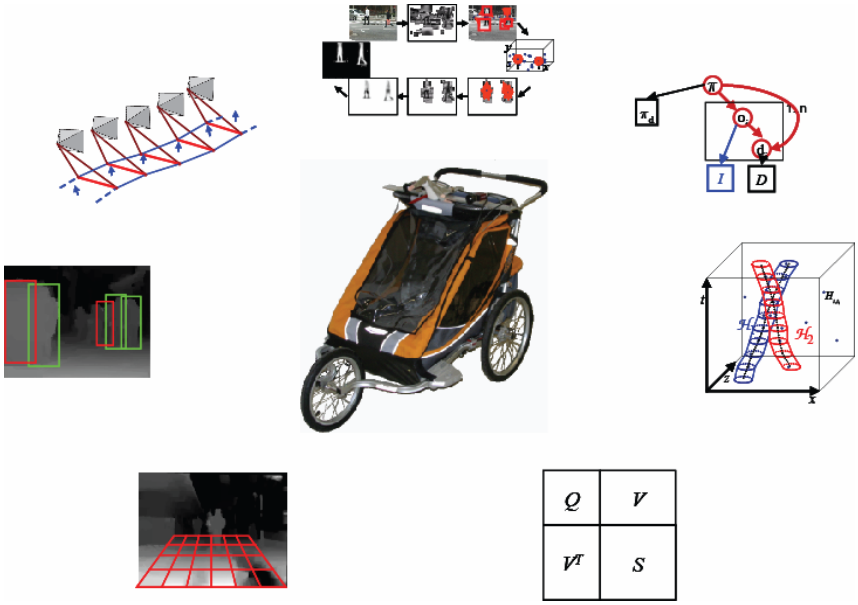


Figure 12. Putting it all together. In this work, we integrate appearance-based object detection, Structure-from-Motion, dense stereo verification, ground plane estimation, trajectory estimation, and multi-object tracking.

8. Putting it all together...

Figure 12 depicts the vision we pursue in DIRAC to put all of the above components together and closely interface them in order to tackle tasks that none of the individual components could handle on its own. In the course of this deliverable, we have developed a series of building blocks for such an endeavor – now we can connect them and explore the potential of cognitive feedback loops for visual scene understanding in the AWEAR scenario.

Figure 13 shows the graphical model that represents the core of our integrated system. It builds upon the ideas for single-frame scene analysis developed in Section 4.2, but extends them with temporal information from the tracking-by-detection approach described in Section5. Both object detection and trajectory estimation are tightly coupled in a combined optimization problem, as shown in Section 6.

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, an additional stereo depth map, and prior information, the model structure is built up. Belief propagation is then used to find a geometrically consistent set of object hypotheses. At the same time, the spacetime volume of previous detections is analyzed to find a set of trajectory hypotheses. Both object and trajectory hypotheses are then considered together in a coupled global optimization step, using the approach from Section 6. The effect of this coupling is to add a spatial prior for object locations that are supported by candidate trajectories from tracking. As shown in Figure 13, this dependency is non-Markovian due to the employed tracking framework.

The output from object tracking is in turn used for stabilizing SfM, 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 current and previous frame pairs.

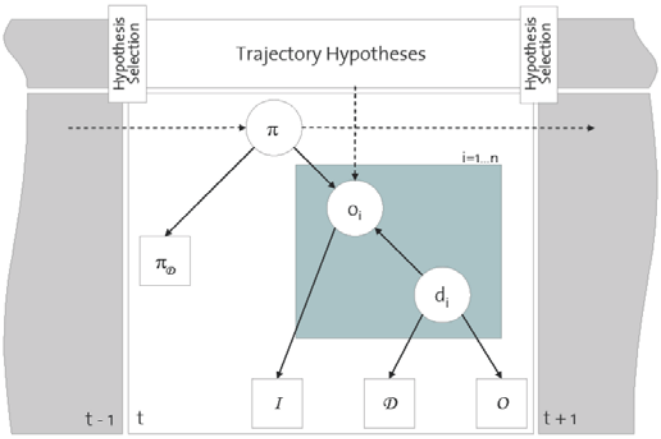


Figure 13. The Graphical Model that integrates appearance-based detection with stereo depth measurements and probabilistic multi-object tracking at the core of our approach.

Figure 14. Example tracking results of our combined system on several challenging test sequences.

We have implemented this integrated system in [5], where we also present a performance evaluation on several very challenging test sequences showing strolls with the AWEAR prototypes from Figure 1(middle, right) through busy pedestrian zones. Altogether, our test set consists of 5 sequences with a total of 4,217 frames spanning 367m travel distance and containing several hundred pedestrians in the vehicle's field of view.

Figure 14 shows example tracking results for several of those test sequences. Our system's ability to track through occlusion is demonstrated in the top row: 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). In the third row, a pedestrian gets successfully tracked on his way around a few standing people, and two

pedestrians are detected at far distances. The final row again demonstrates tracking through major occlusion. Altogether, those results show that our system manages to produce long and stable tracks in complex scenarios.

9. Conclusion

The aim of this deliverable was to provide basic building blocks for dynamic visual scene analysis in the AWEAR application scenario, as well as to explore ways to connect them in cognitive feedback loops. This goal has been achieved by the research presented above. We proposed an architecture to integrate the different vision components (Structure-from-Motion, dense stereo reconstruction, appearance-based object detection, trajectory estimation, and multi-object tracking) and closely coupled them by building up cognitive feedback loops. 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 and thus constitutes a valuable building block for the AWEAR application scenario.

As already stated above, the work presented in this deliverable is still based on perspective cameras. The main reason for this was that the current AWEAR prototype can only capture omni-directional images at 3 frames per second, which is not sufficient to guarantee stable tracking in difficult settings. However, this issue will be resolved by using better cameras in the final AWEAR platform, and the next steps will therefore be concerned with transferring the developed vision capabilities to such a setup. Using omni-directional cameras will bring considerable advantages for the envisioned outdoor application scenario. With their wider field-of-view, those sensors are better suited for SfM and self-localization, since they can capture a far larger portion of the surrounding rigid scene. Consequently, moving objects will not present as much of an obstacle as in the current setup with perspective cameras. Still, the targeted walking aid application implies that cameras will be mounted in a similar location at hip height on the mobile platform, so that people at close range may still block a large portion of the cameras' view. This will still make it necessary to incorporate the proposed cognitive feedback from detection and tracking to geometry estimation.

10. References

[**1**] B. Leibe, N. Cornelis, K. Cornelis, L. Van Gool. "Dynamic 3D Scene Analysis from a Moving Vehicle", in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'07)*, Minneapolis, USA, June 2007. **(appended to this document)**

[**2**] B. Leibe, K. Schindler, L. Van Gool. "Coupled Detection and Trajectory Estimation for Multi-Object Tracking", in *International Conference on Computer Vision (ICCV'07)*, Rio de Janeiro, Brasil, Oct. 2007. **(appended to this document)**

[**3**] A. Ess, B. Leibe, K. Schindler, L. Van Gool. "Depth and Appearance for Mobile Scene Analysis", in *International Conference on Computer Vision (ICCV'07)*, Rio de Janeiro, Brasil, Oct. 2007. **(appended to this document)**

[**4**] B. Leibe, K. Schindler, N. Cornelis, L. Van Gool. "Coupled Detection and Tracking from Static and Moving Cameras", submitted to *IEEE Transactions on Pattern Analysis and Machine Intelligenc*, Dec. 2007. **(appended to this document)**

[**5**] A. Ess, B. Leibe, K. Schindler, L. Van Gool. "A Mobile Vision System for Robust Multi-Person Tracking", submitted to *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'08)*, Anchorage, USA, Oct. 2008. **(appended to deliverable D1.4)**

[6] T. Fortmann, Y. Bar Shalom, M. Scheffe, "Sonar Tracking of Multiple Targets using Joint Probabilistic Data Association", in IEEE Journal of Oceanic Engineering, Vol. 8(3), pp. 173-184, 1983.

[7] B. Leibe, A. Leonardis, B. Schiele, "Robust Object Detection with Interleaved Categorization and Segmentation", in *International Journal of Computer Vision*, Vol. 77, No. 1-3, May 2008.

[8] B. Leibe, K. Mikolajczyk, B. Schiele, "Segmentation-Based Multi-Cue Integration", in *British Machine Vision Conference (BMVC'06)*, Edinburgh, UK, Sept. 2006.

[9] A. Leonardis, A. Gupta, R. Bajcsy. "Segmentation of Range Images as the Search for Geometric Parametric Models." in *International Journal of Computer Vision*, Vol. 14, pp. 253–277, 1995.

[10] D. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints", in *International Journal of Computer Vision*, Vol. 60(2), pp. 91-110, 2004.

[11] K. Mikolajczyk, C. Schmid, "A Performance Evaluation of Local Descriptors", in *IEEE Transactions of Pattern Analysis and Machine Intelligence*, Vol. 27(10), 2005.

[12] D. Reid, "An Algorithm for Tracking Multiple Targets", in *IEEE Transactions on Automatic Control*, Vol. 24(6), pp. 843-854, 1979.

[13] K. Schindler, J. U, and H. Wang. „Perspective n-View Multibody Structure-and-Motion through Model Selection." In *European Conference on Computer Vision (ECCV'06)*, pp. 606–619, 2006.

11. Annex

See appended publications, next pages