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: D1.4
Omnidirectional image acquisition and processing using high-level information

Date of deliverable:
Actual submission date:

Start date of project: 01.01.2006 Duration: 60 months

Organization name of lead contractor for this deliverable: ex: **CTU, ETHZ**

Revision [V0]

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)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	X
CO	Confidential, only for members of the consortium (including the Commission Services)	

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.)

D1.4 OMNIDIRECTIONAL IMAGE ACQUISITION AND PROCESSING USING HIGH-LEVEL INFORMATION

Tomas Pajdla., Bastian Leibe, Hynek Bakstein, Akihiko Torii

Czech Technical University Prague (CTU)
Eidgenoessische Technische Hochschule Zuerich (ETHZ)

Abstract:

In this deliverable we describe image processing and scene analysis modules for the entire processing pipeline of omnidirectional image analysis: starting by a sequence of omnidirectional images, camera calibration, camera motion estimation, view stabilization and rectification, camera tracking and structure from motion, including camera tracking failure, detection appearance-based object detection and recognition, and the feedback from the object recognition to the camera tracking and structure from motion. We describe the individual modules developed as well a pipeline starting from omnidirectional image acquisition to pedestrian detection.

Table of Content

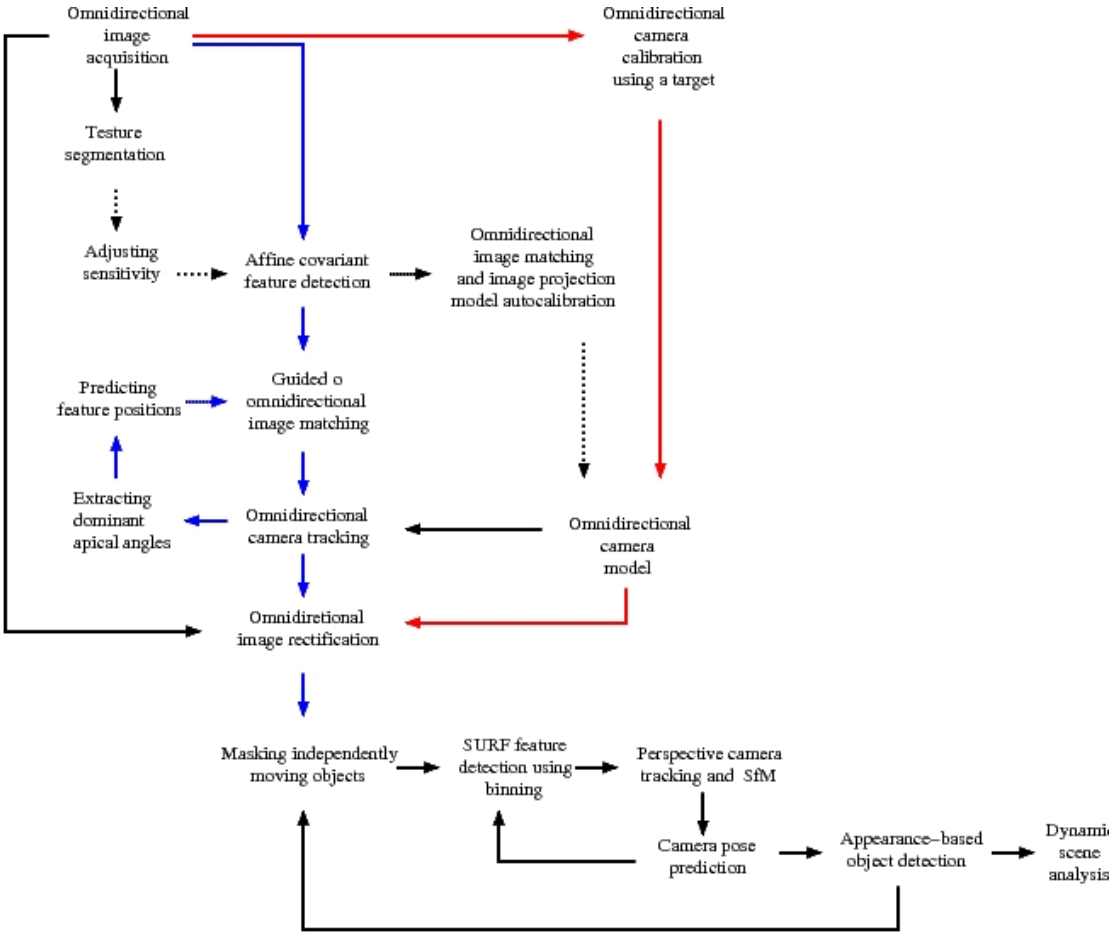

Figure 1 Omnidirectional image processing and dynamic scene analysis processing blocks. Full arrows represent fully implemented processing path. Dashed arrows represent experimental processing path.

1. Introduction

In this deliverable we describe image processing and scene analysis modules for the processing pipeline containing a loop starting by a sequence of omnidirectional images, camera calibration, camera motion estimation, view stabilization and rectification, camera tracking and structure from motion, including camera tracking failure, detection appearance-based object detection and recognition, and the feedback from the object recognition to the camera tracking and structure from motion.

Figure 1 shows omnidirectional image processing blocks and their connection to the object detection and dynamic scene analysis pipeline. The full arrows represent already implemented connections, while dashed arrows stand for the interfaces that already were fixed and are in the phase of implementation and testing.

The processing pipeline currently consists of an omnidirectional image acquisition and processing that produces overlapping perspective cut-outs covering the omnidirectional field of view followed by object detection and dynamic scene analysis on the perspective images The pipeline contains several places where a higher-level information is used to drive the extraction of lower level information. First, sensitivity of omnidirectional feature detectors can be adapted depending on the type of texture to guarantee uniform coverage of the field of view by points for matching and to suppress features in areas where chance of successful match is low. Secondly, estimated camera motion is used to guide matching in order to extend features into areas where they have not been detected and to select features corresponding to a partial reconstruction of the scene. Finally, features detected on independently moving objects, like pedestrians and cars, are masked to promote the features detected on the background with respect to which the camera motion is estimated. The structure of the pipeline, as well as the individual blocks, are being invented and designed based on the experience gained from experiments carried out in WP3 and WP6. It is already clear that the next step in constructing the pipeline will be to integrate the omnidirectional imaging into the dynamic scene analysis.

Next, we shall concentrate on describing the blocks of the above pipeline.

2. Omnidirectional image acquisition

AWEAR-β mobile AWEAR-β static

Figure 2. Omnidirectional image acquisition of the AWEAR-β platform.

Current omnidirectional image acquisition is part of the preliminary AWEAR platform. It consists of a stereopair of Kyocera FineCam M410R perspective acquiring 4 MPix images at the rate of 3 frames per second, equipped with the omnidirectional Nikon FC-E9 lenses offering a 183° circular view field 1]. Audio acquisition appears in two forms. In its mobile form, the audio recording is provided by a single stereo microphone of a Canon XM-1 DV camera. In its static form, the audio acquisition can be accomplished either by using up to 4 USB microphones or by two triplets of OL anthropomorphic hearing aid microphones.

3. Texture segmentation

Figure 3. Multilabel texture segmentation is able to distinguish featureless areas (sky and sand on the left) from highly textured areas (palm tree on the left), or bushes and other natural textures from guiding and other artificial textures (right).

To make robust and fast image matching, it is useful to assign labels to detected image regions that distinguish between regions with high and with low chance of finding a match.

We have previously observed that there exist several typical types of textures in image sequences acquired in the outdoor city environment, such as, geometrical shapes generated by buildings, small, complex and greenish shapes generated by leaves of trees and bushes, and large, uniform and bluish patches on the sky, see Figure **3**. Some of the characteristics are local, e.g. colour, but others, e.g. being surrounded by similar shapes, is global and can be extracted only by using image or other context.

In principal, it is difficult to know what are the useful characteristics beforehand and therefore the types of the regions, represented by unique labels, as well as the assignment between the labels the ability of finding a match needs to be learned from examples of successfully and unsuccessfully matched images.

It has been demonstrated [3] that properties recovered by general image segmentation can help in interpreting the types of image regions. In [3], for instance, such information could be used to distinguish between ground plane and facade patches. Their segmentation was built on a leave-one-out-against-the-others two-label segmentation and by generating a number of alternative segmentations by changing the parameters of the segmentation algorithm.

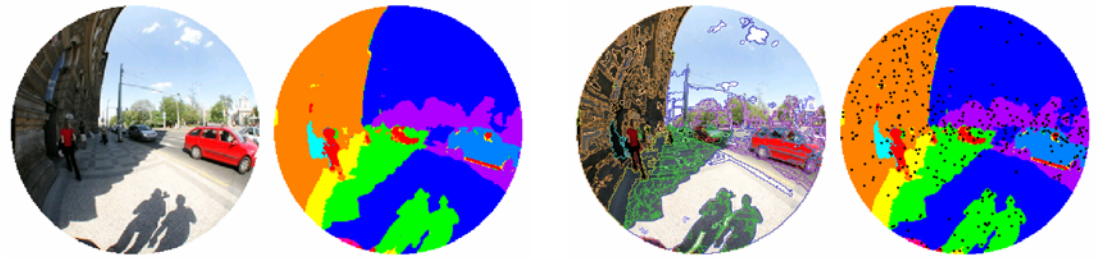

Figure 4. Original omnidirectional image, its 8-label segmentation, MSER regions detected inside the segmented regions, and the corresponding feature points. The segmentation provides labels to the feature points allowing to sample the feature points according to the type of the texture by the PROSAC ordered sampling algorithm [2]

We use a similar approach to generating a texture type descriptor for detected image regions which then enter the matching. The regions are influenced by the texture segmentation in two ways. Firstly, the detected regions are confined to be contained in a single texture region. Secondly, the label of the texture region its associated texture descriptor becomes part of the region feature descriptor used for matching. Thus, regions detected on one type of texture, e.g. on the sky, in one image may be matched only against regions from the same type of texture in the other image, Figure 4.

We have developed [4] a technique for learning types of textures from images and using them to segment the images efficiently with new Markov Random Field (MRF) solvers based on the linear programming and a relaxation of the original integer programming problem. We have formulated the multi-label segmentation into regions coherent in texture and colour as a MAX-SUM problem for which efficient linear programming based solvers have recently appeared. By handling more than two labels, we can represent more textures and used them in a natural way to segment the image. We build the MRF on superpixels to speed up the segmentation while preserving colour and texture. We used a new quality functions for setting the MRF, exploiting priors from small representative image seeds, provided either manually or automatically. In experiments with the Berkeley segmentation database, we have observed that our automatic segmentation method outperformed previous techniques in terms of the Global Consistency Error and consistently separated natural textures from the artificial ones.Texture segmentation based on learned texture prototypes typically divides image into regions of similar properties. The types of textures are learned from the segmented image and thus it adapts to observed scenes. If necessary, this approach can be also used to segment sequences of images by learning typical texture prototypes from consecutive subsequences on the fly.

One of the basic problems with future detectors is that it is necessary to set a meaningful sensitivity threshold for suppressing too weak detections at early stage of the processing to avoid spending too much of computational resources on trying impossible combinations in the later phase of matching and recognition. It is often impossible to set one threshold for complete image as, e.g., shown in Figure 4. The problem is that the level of detail changes dramatically when changing from the sidewalk to buildings and then to the sky. We can notice that the segmentation to large extent splits the image into regions of homogeneous

textures, i.e. to sky, bushes, faces of buildings, and several other regions. Not all regions necessary correspond to "best human segmentation, but they always have similar textural quality. We observed that when the removal of the weak features is done inside each of the segmented regions, based on the statistics of feature responses there, we retain more features in areas that are not so populated by features and suppress features in those areas where they are in abundance. Such approach reduces computational requirements and leads to a more uniform covering of the image, which is particularly important for good-quality structure from motion.

This approach implicitly relies on the fact that the segmentation will correspond in the images to be matched. This can, of course, hold only approximately and it may happen that segmentations considerably differ due to the viewpoint change and induced occlusions. However, if the sensitivity is connected to the feature class, rather than to the region itself, then features in (at least large) regions are detected with similar thresholds and will get associated similar texture descriptors. Since this is not absolutely reliable, we use the ordered sampling [2] to sample matches with more similar descriptors earlier but allow for sampling even less likely mathes later. It is still an open question how to combine texture descriptors with feature descriptors in a good way.

4. Affine covariant feature detection

Figure 5. Images with large camera motions can be tracked with affine covariant features, such as SIFT, MSER, and APTS.

Large camera motions, such as shown in **Figure 5**, call for affinely covariant feature detectors [1, 5]. Role of these features is to large extent complementary to using fast SURF features developed in WP3 and used in the dynamic scene analysis of the dynamic AWEAR platform. SURF features are highly successful for small camera motions and facilitate real time tracking. On the other hand, when camera makes a very fast turn, which often happens with hand held or body mounted cameras, they often can't be used since their descriptors are not sufficiently specific and discriminative. Affinely covariant features bear sufficiently complex descriptors to be matched under large image changes. On the other hand, they require more effort to be computated and currently can't be used for real time processing. In current pipeline, Figure 1, the problem of larger motions is partly handled by detecting failures of the

SfM module and using previously computed 3D structure to catch up in later frames. This has, however, limitations since large motions often dramatically change the positions of the matches and the tracking paradigm has to be replaced by the recognition paradigm to catch up. We foresee a possibility to use the affinely covariant regions in helping to recover after the tracking has failed. Since this happen relatively infrequently, it is still feasible for real-time tracking.

5. Guided omnidirectional image matching

Original matches Matches extended by guided matching

Figure 6. Guided matching can improve considerable improve the coverage by matches in difficult situations when many image features have very similar descriptors.

Another important problem in finding image matches arises when too many features share similar descriptors. This obviously happens when descriptors are not sufficiently discriminative but also when there are repetitive structures, as often happens in urban environment or many random structures, as often happens on bushes and trees.

Figure 6 shows an example of an omnidirectional image pair with large camera rotation and many features on bushes. These features have very similar descriptors and therefore the chance of constructing correct tentative matches based on the descriptor similarity is relatively low, 1,5 % in this case [2]. The situation can be to large extent remedied by using a guided matching. The number of matches almost always increases and, more importantly it often much better covers the field of view for difficult image pairs, Figure 6.

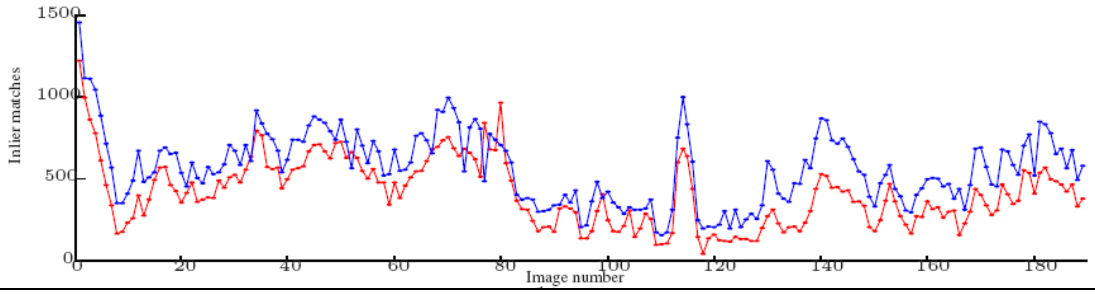

Figure 7. The number of matches before (red) and (blue) after guided matching on the “walk 2” sequence.

Our guided matching algorithm gradually adds more and more matches by using the knowledge of actual estimate of the epipolar geometry to constraint the regions in which the tentative matches are constructed. The probability of finding many very similar features in a small part of image is lower which reduces the problem with repetitive structures. We extend the matches by restricting the position of tentative matches to generate areas with smooth apical angles [6]. See the algorithm in Appendix 1. Our guided matching tries to find the matches to prefer smooth surfaces.

6. Omnidirectional camera tracking and autocalibration

Original 120° FOV image Image rectified using the autocalibrated model.

Figure 8. Internal as well as external camera parameters are estimated during the autocalibration process. The image on the right demonstrates that the estimated parameters faithfully model the camera by straightening an image cutout.

Figure 1 shows two ways towards the camera calibration. The first way uses a calibration target with known dimensions to estimate the internal camera parameters. This has been described in the deliverable [7]. Using calibration targets is a valid approach for industrial applications and helps to substitute results at early phases of the projects for the development of the following modules, but it is not a satisfactory approach for any routine, not speaking about massive, use of the technology. We are therefore developing a second route that will be fully automatic and will autrocalibrate omnidirectional cameras from image matches only. Core capabilities of such autocalibration technology are

1. image matching, and
2. robust camera motion and model fitting

that have to work altogether in a RANSAC framework [2]. The key component of this approach is the computation of the camera model from a smallest possible sample of data points, known as minimal problem solving.

We have formulated and solved three new minimal problems for cameras with radial distortion. The solution is based on the Groebner based method for solving systems of polynomial equations. The first minimal problem, the simultaneous estimation of fundamental matrix and a common radial distortion parameter for two uncalibrated cameras from eight image point correspondences, has been implemented in floating point arithmetic [8]. The two remaining minimal problems (i) simultaneous estimation of essential matrix and a common radial distortion parameter for two partially calibrated cameras from six image point correspondences and (ii) simultaneous estimation of fundamental matrix and different radial distortion parameters for two uncalibrated cameras from nine image point correspondences, are numerically more challenging and therefore have been first implemented in exact rational arithmetic only [9]. Next, we developed efficient and robust floating-point solvers [10].

The three minimal problems can be used in a calibration process of standard or omnidirectional cameras and integrated into the 3D reconstruction framework. They are based on the one-parameter division distortion model given by the formula

$$p_u \sim \frac{p_d}{(1 + \lambda r_d^2)},$$

where λ is the unknown distortion parameter, $p_u = (x_u, y_u, 1)$, resp. $p_d = (x_d, y_d, 1)$, are the corresponding undistorted, resp. distorted, image points, and r_d is the radius of p_d w.r.t. the distortion center. Using this model it is possible to handle up to 66% cutouts from omnidirectional images and using the algorithm for the last minimal problem also images with relatively different radial distortion parameters.

Having a camera model, the omnidirectional structure from motion developed in WP3 and WP6 [7, 11]

7. Omnidirectional image rectification

Figure 9. Omnidirectional rectification into overlapping as perspective cutouts (top) and into a cylindrical panoramic image (bottom). The cylindrical panoramic image allows representing a complete view of persons standing very close to the camera in images of acceptable size.

Rectification of omnidirectional images provides the interface to ETHZ object recognition and scene interpretation. To cover the complete field of view, or at least its interesting part, we generate a set of overlapping perspective images, on which the standard KUL structure from motion and ETHZ object detection and dynamic scene rectification can be performed. The rectification uses camera a calibration model calibrated by the procedures described above.

Figure 9 shows two types of omnidirectional image rectifications used for pedestrian detection. The conventional approach, which is currently acceptable by the ETHZ object detection, consists of a series of overlapping perspective views. The alternative approach, we are working on, is a single panoramic rectification. There are several reasons why the panoramic rectification should be a better interface between the lower omnidirectional image processing and the higher dynamic scene analysis. First, the panoramic rectification can more efficiently represent larger view angles since no parts of the view angle need to be covered twice. Secondly, the panoramic rectification allows to represent much larger view angle by nonlinearly rescaling the vertical axis as shown in Figure 9 (bottom). The half-spherical view-field is mapped onto the image plane by a projection constructed by a conveniently chosen projection center to balance the size of the person. This is a satisfactory approach for persons in larger distance from the camera. For persons standing closer to the camera, the proportions of the upper and lower body-part becomes unnatural and is not properly modeled by the human model learned on standard perspective images.

It is an open question how to design a convenient projection function to be able to use existing databases of perspective images for training detectors functioning also on panoramic images.

The second technical problem that has to be addressed to be able to use panoramic images as the interface for ETHZ structure from motion is how to use ground plane constraint in these images. It is rather a technical problem of the implementation but its solution might need to touch many modules in the existing systems. It is not yet clear how this is difficult. The ultimate approach is to use the omnidirectional camera tracking developed in WP3 and WP6 by CT|U and KUL to replace the existing perspective structure from motion. This, again, is not as simple task as it may seem since the KUL 3D reconstruction pipeline is optimized for perspective images. This issue will have to be decided based on next series of experiments with the existing pipeline.

8. Feedback from object tracking and dynamic scene analysis to feature detection

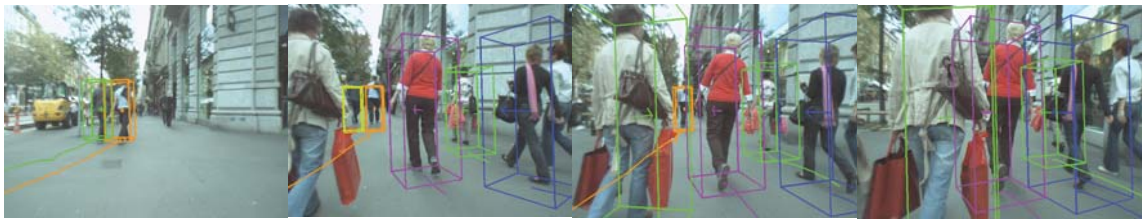

Figure 10. Example tracking results in very crowded scenes that are only made possible by feeding back information from object detection and tracking to Structure-from-Motion.

As already the first-year deliverable D3.1 has shown, the SfM results can considerably help recognition by supplying important information about the underlying scene geometry. In this section, we will complete this interaction to a loop by also feeding back information from scene understanding to help SfM. As we will show, such a feedback is crucial for robust performance in crowded scenes, such as the one depicted in Figure 10. 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 SfM 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 [12] 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 the appended paper [12].

In contrast to the previous sections, the work presented in this section 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 with SURF features, 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 and potentially use of the features suitable for omnidirecitonal images and allowing for faster camera rotations. Results for this omni-directional scenario will then be presented in the M36 deliverable D3-7.

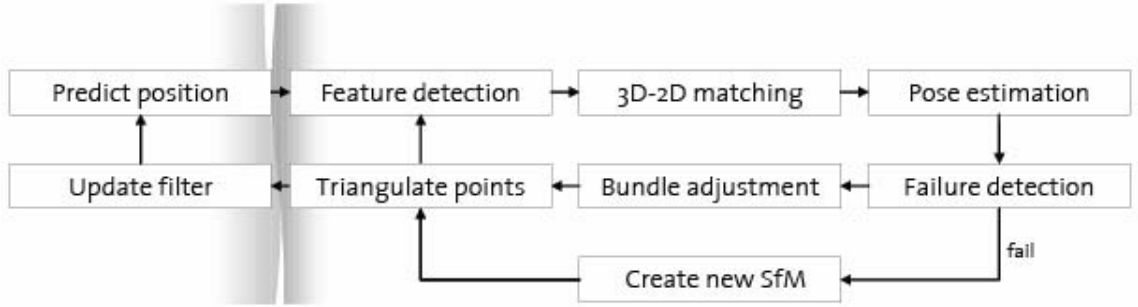

Figure 11. Flow diagram of the employed Structure-from-Motion system. The shaded regions indicate the insertion points for the semantic feedback from object tracking.

1.1 Spatial Binning for SfM

Figure 11 shows a flow diagram of our basic SfM system. The employed system builds on previous work by [13, 14], as well as on methods developed in deliverable D1.1. In order to achieve the necessary robustness, we use a spatial binning scheme, which encourages a feature distribution that allows stable localization. With this scheme, each incoming image is divided into a grid of 10×10 bins, and an approximately uniform number of points is detected in each bin using a Harris corner detector with locally adaptive thresholds. 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 [15]. Bundle adjustment is run on a window of n_b = 18 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 [16]. Scene points are directly associated with a viewpoint-invariant feature descriptor 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-finding 3D points that carry information from multiple viewpoints and are hence more stably localized.

1.2 Recognition-based Dynamic Scene Analysis

The goal of this section is to improve robustness of the SfM system by feeding back semantic information from object recognition and tracking. For this, we use the dynamic scene analysis framework presented in deliverable D3.5. Here, we just give a short outline of the overall approach and refer to D3.5 for details.

For each frame pair of the input video streams, we first estimate the camera location and scene geometry using the above SfM system. 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. Using the estimated camera location from SfM, detection bounding boxes are converted into 3D world coordinates, which are then integrated over time in order to estimate physically plausible object trajectories. A final global optimization criterion takes object-object interactions into account in order to select a subset of trajectories that together best explain the observed data under the physical constraint that no two objects may occupy the same space at the same time. As the results from D3.5 show, the resulting system achieves robust detection and tracking performance in very challenging real-world scenes.

1.3 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 people tracking behavior, since tracking is performed in 3D world coordinates. As tracking is in turn used to constrain SfM, 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 estimates 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.

1.4 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 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 (see Figure 12(left)). 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 the following section will demonstrate.

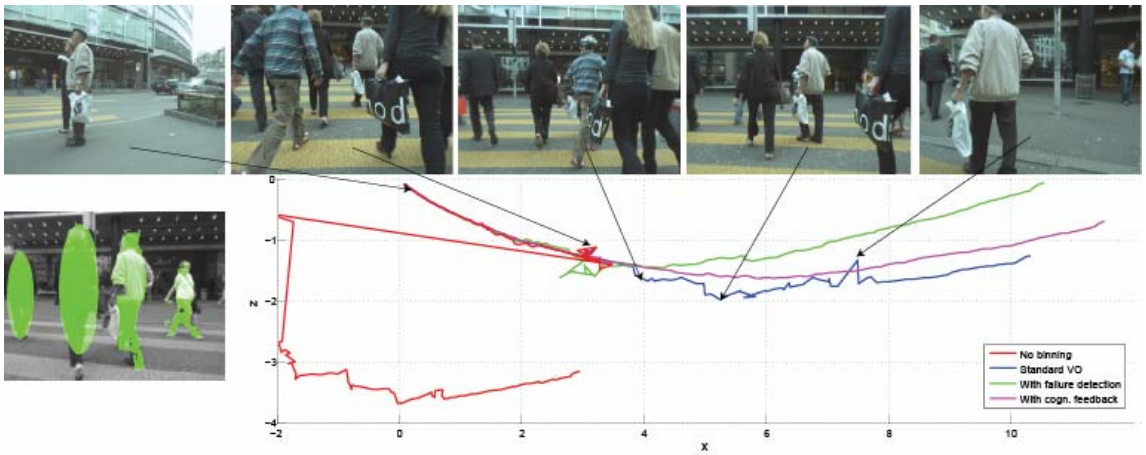

Figure 12. Trajectory estimation results of our system with and without cognitive feedback. (top) A few frames of a difficult sequence, where the camera vehicle arrives at a crowded pedestrian crossing. (left) Example top-down mask used by the cognitive feedback to adapt feature sampling. (bottom) Resulting trajectory estimates with and without cognitive feedback (the plot shows a bird's eye view on the ground plane). As can be seen from this plot, our proposed method results in a much cleaner trajectory and avoids catastrophic failures caused by excessive scene motion.

9. Experimental Results

In order to underline the importance of the proposed integration, consider the scene shown in Figure 12, 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 SfM 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, SfM 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.

Spatial binning for feature selection (as promoted in [4, 14]) 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 SfM from attaching itself to moving bodies. Finally, the magenta line shows the result using our complete system, which succeeds in recovering a smooth trajectory.

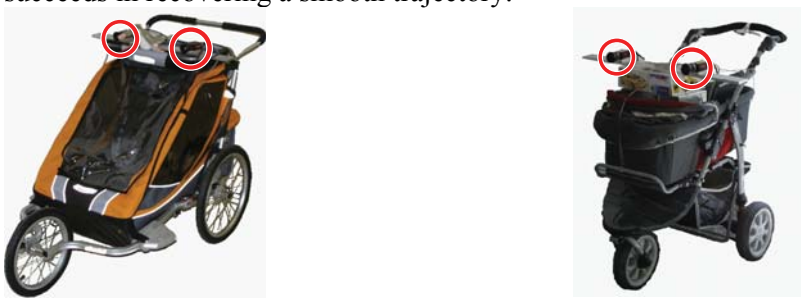

Figure 13. The two prototype setups used to test out the cognitive feedback from object tracking to SfM. The employed child stroller platform is already close in size to the envisioned AWEAR application scenario of an intelligent walking aid. Such a setup brings several important difficulties with it. Size constraints allow only for a relatively small baseline. In addition, the low camera placement makes it hard to observe a sufficient portion of the scene. Finally, camera motion is not always smooth due to uneven ground surfaces and the relatively small wheel diameters.

In [12], we present a detailed performance evaluation on several very challenging test sequences showing strolls with the AWEAR prototypes from Figure 13 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. Again, we refer to the paper [12] for details of the evaluation.

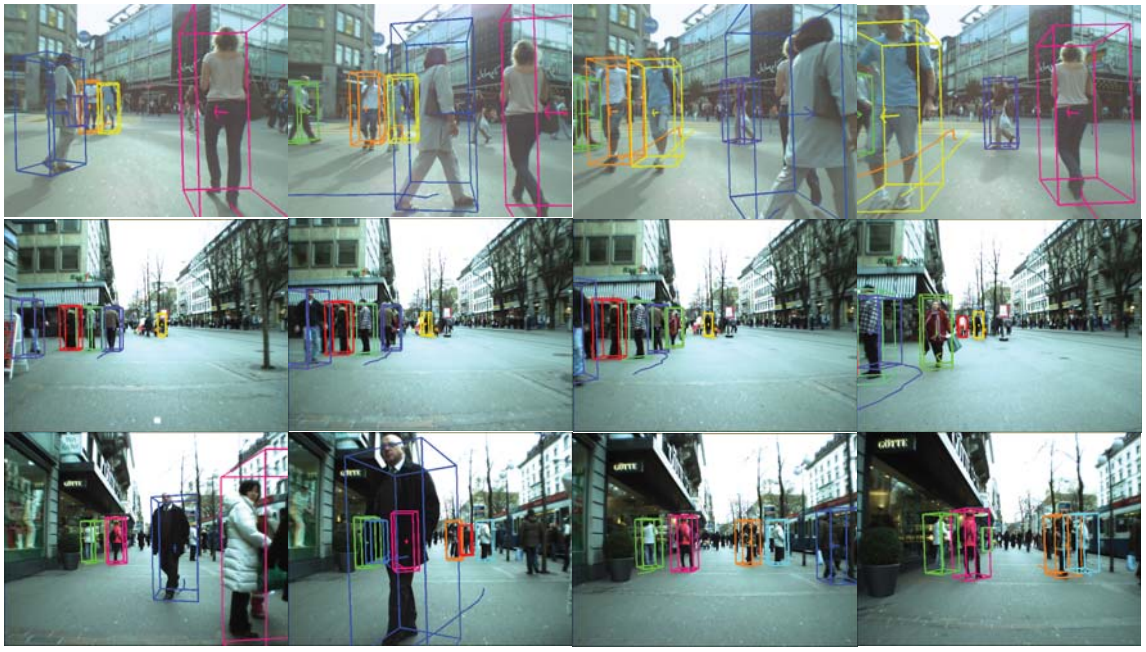

Figure 14. Example tracking results of our combined system on several challenging test sequences.

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 second 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.

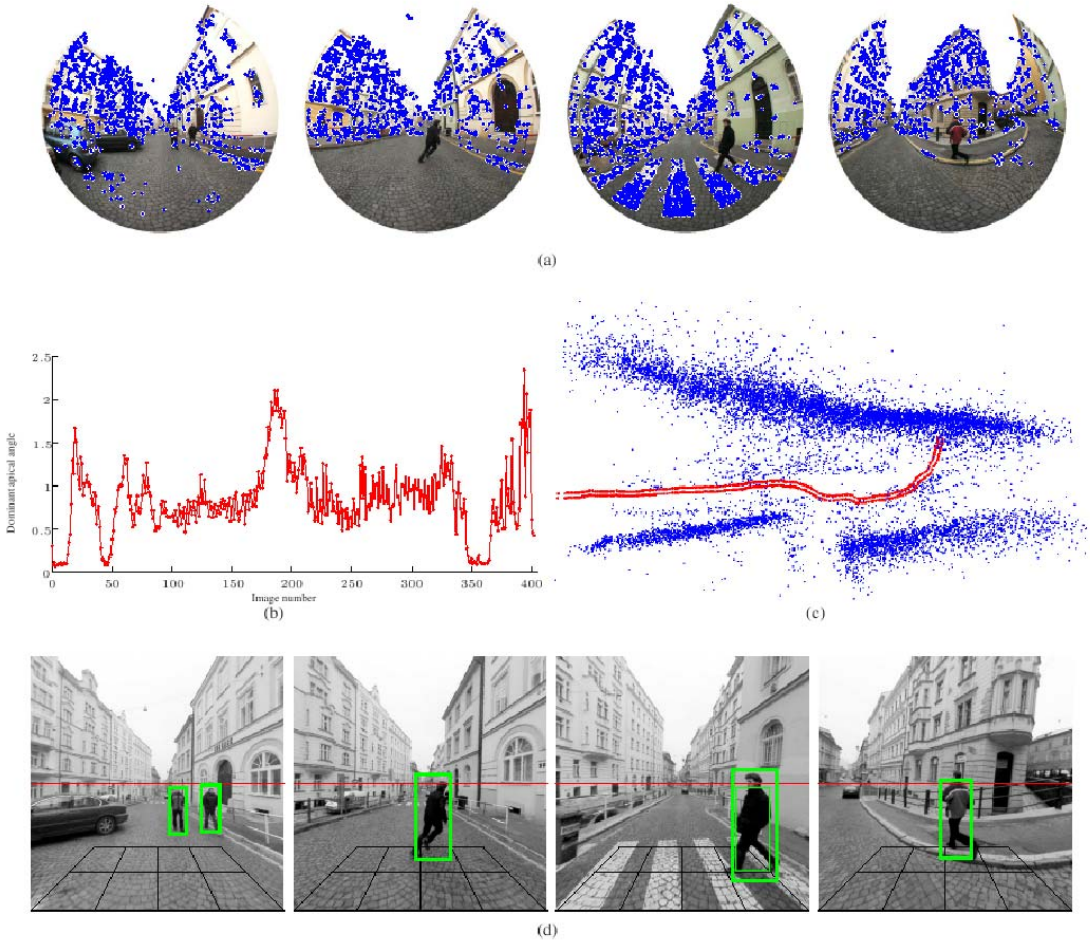

Figure 15. ETHZ pedestrian detection in omnidirectional images. (a) Omnidirectional images with matches obtained by looking for the largest support of an omnidirectional camera model by the ordered sampling of tentative matches formed from affine covariant image features. (b) Estimated relative length of the camera translation and (c) the corresponding camera path. (d) Rectified perspective cutouts with the camera position obtained by the KUL SfM upgraded for omnidirectional images (WP3, WP6), the ground plane (black grid), the horizon (red line) and detected pedestrians (green rectangles).

Figure 15 shows the result of the first integration of the omnidirectional image processing with higher-level image extraction as described above. In this example, the ETHZ dynamic scene analysis has not been used in its full capacity of dynamic scene analysis but only with pedestrian detection. The use of other features would, however, did not change current level of integration.

We were able to go from omnidirectional image acquisition till pedestrian detection as shown in Figure 1. The calibration for this experiment has been obtained through the red path, i.e. using the calibration pattern, but in the meantime the blue path become ready and will be tested in the next experiments.

Standard structure from motion, such as the KUL implementation, relies on a good initialization of the structure from camera motion with sufficient (dominant) translation. If the beginning of the acquired sequence does not contain sufficiently large translation, the initial structure is uncertain and the structure from motion may easily fail to start. This problem becomes particularly serious when working with hand-held cameras. Human walking and motion is much more irregular than motion of a car. Humans, unlike cars, can easily rotate on a spot without translating. That brings a problem how to detect small camera translations.

We have proposed a measure of the amount of the camera translation by the dominant apical angle, the angle under which the camera centers are seen from the perspective of the reconstructed scene points. Simulated experiments show that the dominant apical angle is a linear function of the length of the true camera translation. By skipping image pairs with too small motion, we could reliably initialize the omnidirectional upgrade of the KUL structure from motion and compute accurate camera trajectory in order to rectify images and to use the ground plane constraint in pedestrian recognition.

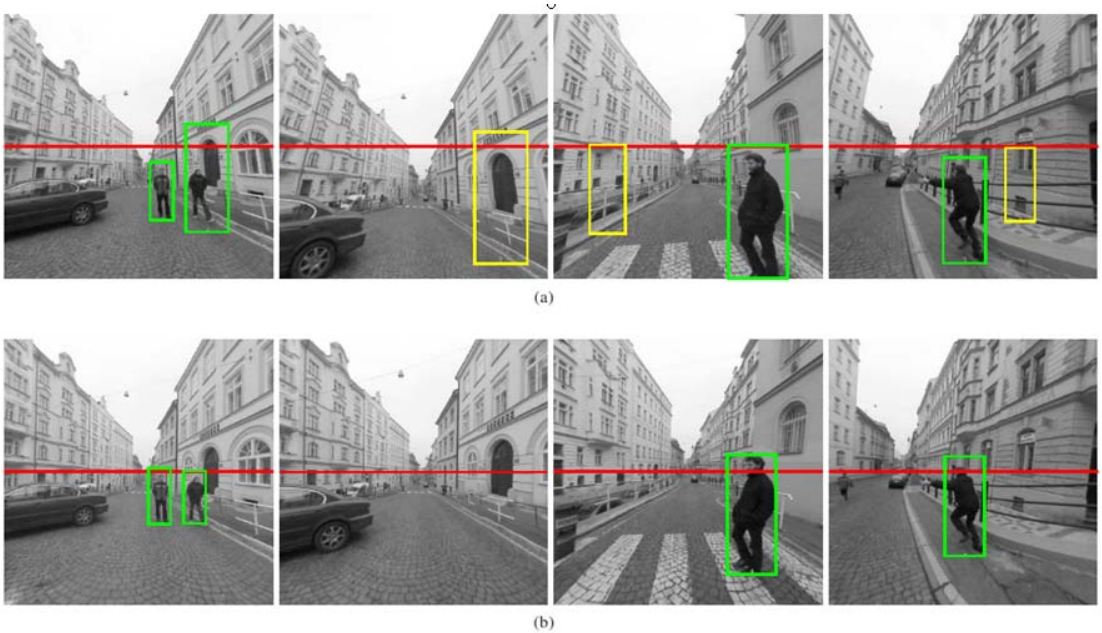

Figure 16. Perspective cutout rectification. (a) A sequence of perspective cutouts with pedestrian recognition without using the rectification to correctly track the ground plane. (b) the same result with the ground plane correctly tracked using the omnidirectional camera tracking based on KUL SfM. The number of false positives considerably decreases when the correct ground plane constraint is used.

See [6] for more details on using dominant apical angles to initialize KUL structure from motion.

Figure 16 shows an example of the improvement of the pedestrian detection after the ground plane constraint has been correctly tracked through the sequence.

10.Conclusion

We have succeeded to construct a pipeline starting from omnidirectional image acquisition to pedestrian detection.

The next work will concentrate on further integration of the omnidirectional image processing, feature extraction, camera tracking, and on closing feedback from dynamic scene analysis to low-level feature extraction based on the association between feature descriptors, including the texture descriptor, and the success of image matching.

11. References

[1] H. Bakstein, Z. Kukelova, A. Torii, T. Pajdla. Omnidirectional sensors and features for tracking and 3D reconstruction.DIRAC Deliverable D1-1, 2007
[2] A. Torii and T. Pajdla. Omnidirectional Camera Motion Estimation. International Conference on Computer Vision Theory and Applications (VISAPP'08), Funchal, Madeira, Portugal, January 2008.
[3] D. Hoiem, A.A. Efros, and M. Hebert, "Geometric Context from a Single Image", ICCV 2005.

[4] B.Micusik and T.Pajdla. Multi-label Image Segmentation via Max-Sum Solver. CVPR 2007.
[5] Feature detectors for body parts, tracking, and fast object detection. DIRAC Deliverable D1-2, 2007.
[6] A. Torii, M. Havlena, T. Pajdla. Measuring camera translation by the dominant apical angle. CVWW 2008. Moravske Toplice, Slovenija. Feb. 2008.
[7] H. Bakstein, |T. Pajdla, K. Cornelis, B. Leibe, M. Havlena, A. Torii. OmniCamera Tracking. DIRAC Deliverable D6-2, 2007.
[8] Z. Kukelova and T.Pajdla. A Minimal Solution to the Autocalibration of Radial Distortion. CVPR 2007.
[9] Z. Kukelova, T. Pajdla. Two Minimal Problems for Cameras with Radial Distortion. OMNIVIS 2007 in conjunction with ICCV 2008. Rio de Janeiro 2008.
[10] J. Byroed, Z. Kukelova, T. Pajdla, K. Astroem. Fast and Robust Numerical Solutions to Minimal Problems for Cameras with Radial Distortion. Submitted to CVPR 2008.
[11] M. Havlena, T. Pajdla, K. Cornellis. Structure from Omnidirectional Stereo Rig Motion for City Modeling. to appear in International Conference on Computer Vision Theory and Applications (VISAPP'08), Funchal, Madeira, Portugal, January 2008.

[12] 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 this document)**

[13] D. Nister, O. Naroditsky, and J. R. Bergen. „Visual odometry". In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'04)*, 2004.
[14] Z. Zhu, T. Oskiper, O. Naroditsky, S. Samarasekera, H. Sawhney, and R. Kumar. "An Improved Stereo-based Visual Odometry System". In *Performance Metrics for Intelligent Systems (PerMIS'06)*, 2006.
[15] D. Nister. "A Minimal Solution to the Generalised 3-Point Pose Problem. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'04)*, 2004.
[16] A. Davison. "Real-Time Simultaneous Localization and Mapping with a Single Camera". In *International Conference on Computer Vision (ICCV'03)*, 2003.

12. Appendix 1 – Guided matching

The algorithm for guided matching using iterpolation of apical angles and ordered sampling.

1. Detect feature points x and y in the left and right images, respectively.
2. Set the search region to full image
3. Generate tentative matches p by finding mutually closest feature points, measured by the Euclidean distance in the feature space.
4. Compute relative camera motion E from p by ordeed sampling as in [2]:
    a. the orderd random sampling
    b. the orientation constraint on 5 points used to solve the minimal problem
    c. the orientation constraint on all reconstructed points
    d. the soft voting to find the most consistent model
5. Remove mismatches on epipolar lines caused by repetitive patterns
    a. Compute apical angles as in [6]
    b. Remove the tentative matches from p which have apical angles far from all their image neighbours
6. Stop if there are enough matches or no feature points left.
7. Construct search regions for feature points x and y
    a. Assign the apical angle to each unmatched feature point in images as the apical angle of its neares image neighbour.
    b. Use the apical angle to predict the position of the feature point in the other image.
    c. Set the search region for the corresponding feature point to a neighbourhood of the predicted point
8. Go to 3