A THREE-LEVEL ARCHITECTURE FOR MODEL–FREE DETECTION AND TRACKING OF INDEPENDENTLY MOVING OBJECTS

Nicolas Pugeault[1], Karl Pauwels[2], Mark M. Van Hulle[3], Florian Pilz[4] and Norbert Krüger[5]

[1]*University of Surrey, U.K.*

[2,3]*Katholieke Universiteit Leuven, Belgium*

[4]*Aalborg Universitet Copenhagen, Denmark*

[5]*University of Southern Denmark, Denmark*

Keywords: Motion detection, 3D reconstruction, Egomotion, Independently moving objects.

Abstract: We present a three–level architecture for detection and tracking of independently moving objects (IMOs) in sequences recorded from a moving vehicle. At the first stage, image pixels with an optical flow that is not entirely induced by the car's motion are detected by combining dense optical flow, egomotion extracted from this optical flow, and dense stereo. These pixels are segmented and an attention mechanism is used to process them at finer resolution at the second level making use of sparse 2D and 3D edge descriptors. Based on the rich and precise information on the second level, the full rigid motion for the environment and for each IMO is computed. This motion information is then used for tracking, filtering and the building of a 3D model of the street structure as well as the IMO. This multi-level architecture allows us to combine the strength of both dense and sparse processing methods in terms of precision and computational complexity, and to dedicate more processing capacity to the important parts of the scene (the IMOs).

1 INTRODUCTION

Independently moving objects (IMOs), such as cars, bicycles and pedestrians represent a major source of hazard while driving, and detecting and tracking them is therefore a critical task for an automated driving system. Although this task is performed effortlessly by humans, its implementation in an artificial system demands the solving of several difficult vision problems. First comes the problem of actually detecting moving objects in the scene. Although there exist robust solutions for this problem in the case of a static camera (e.g., (Leibe et al., 2008)), the moving camera version is considerably more difficult. Recent approaches have used model–based detection with very good results (Leibe et al., 2008). One limitation of such approaches is the difficulty to detect IMOs early on. Far objects will be described by smaller patches in the image, that will be difficult to match with the model. Another limitation is that moving objects for which no model is provided will not be detected and will be simply ignored by the system. The present work presents a model-free detection mechanism, and is therefore more general. A second problem is tracking *multiple* objects simultaneously. Some methods use world knowledge to guide the tracking and simplify the problem (ground plane, perspective, see (Hoiem et al., 2007; Hoiem et al., 2008)). Third, combining early detection with real time processing enables the system to plan early and react appropriately to driving situations. There arises a dilemma, because early detection and interpretation of visual information requires high resolution images, which increases the processing load. For this reason, it is desirable to have an attention mechanism that allows the system to focus limited processing resources to moving objects, just as humans perceive regions corresponding to independent motion as highly salient (Rushton et al., 2007).

The problem of tracking multiple independent target simultaneously has been studied for a long time, e.g., Bar-Shalom reviewed early attempts in (Bar-Shalom, 1978). The key problem faced by multiple targets data association is the exponential complex-
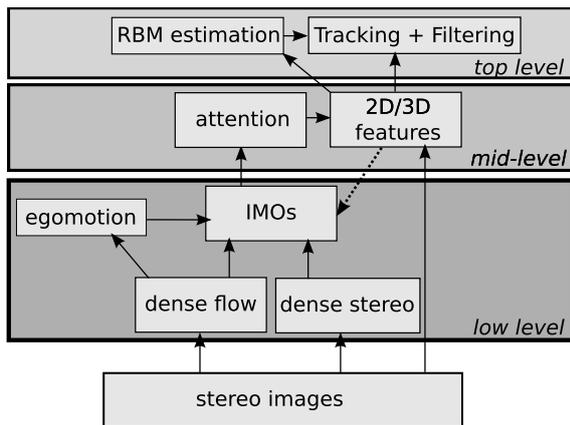
Figure 1: Illustration of the architecture presented in this paper.

ity of the exhaustive approach. A more sophisticated approach was proposed by Reid's Multi-Hypothesis tracking (MHT) (Reid, 1979). Cox (Cox, 1992) reviewed both approaches, focusing more on the data association problem, and extended the single target Probabilistic Data Association Filter (PDAF) to multiple targets, called Joint Probabilistic Data Association Filter (JPDAF). Betke et al. (Betke et al., 2000) proposed a multiple target tracking method with a complexity linear to the number of targets, that was applied on driving scenes recorded on American and German highways. The car detection is done either: (1) by image differentiation (only works with nearby cars); (2) a model-based approach driven by a two step search (first, the system looks for long edges, that are then verified against a car template). Comaniciu et al. (Comaniciu et al., 2003) focus on the issue of target's representation and localisation, in a kernel-based framework. Their approach dwells on the representation of 2D deformable objects whereas our focus is on the generation and use of 3D model information. Leibe et al. (Leibe et al., 2008) proposed a model–based system for coupled detection and tracking of multiple objects. Lourakis et al. (Lourakis et al., 1998) used a sparse set of points to detect independent motion from the residual planar parallax normal flow field. Sawhney et al. (Sawhney et al., 2000) combine both the epipolar and the shape consistency constraint to derive a sparse geometric solution for independent motion detection. Moreover, there exist on-line and commercial systems (e.g., (Bertozzi and Broggi, 1998)) for lane and obstacle detection, but those systems mainly make use of low resolution images to provide real–time processing, and could benefit from the multiple layer attention mechanism presented herein.

In this article, we present a three layer system that makes use of such an attention mechanism, illustrated in Fig. 1. The first layer combines multiple visual cues (dense stereo and optic flow) to segment the image motion into parts that result from self–motion in the static environment, and parts that are due to the independent motion. From this detection stage, Regions of Interest (ROIs) describing the IMOs are defined. The second level extracts 3D features using stereopsis, processing the background at low resolution and the ROIs at high resolution. IMOs containing little structure are discarded at this stage. This is symbolised in Fig. 1 by the dashed arrow between the middle and lower levels. The upper level of our approach is concerned with simultaneously tracking and building a feature–based model of the IMOs (similarly to (Dissanayake et al., 2001; Montemerlo et al., 2002; Thrun et al., 2004)).

We have evaluated our data set on a very challenging publicly available data set (http://www.mi.auckland.ac.nz/eisats) for which ground truth of IMOs is provided. We can show that by making use of the multi-level architecture, we are able to increase performance significantly compared to just using a the first level mechanism. We can also show on examples, that we are able to compute the 3D motion as well as the street structure in a way that allows for the computation of the relative 3D motion of the IMO compared to the street structure.

The novel aspects of this work are: (i) *Model–free:* the system does not rely on object models but instead automatically builds IMO models; (ii) *Attention–based:* the system uses attention to allocate resources efficiently, and especially processes IMOs at higher resolution; and (iii) *Hybrid:* the proposed system makes use of a hybrid approach, where the dense layer provides model-free IMO detection, and the feature based layer is used to discard outliers and build and track IMOs' models. In this way, we can combine computational efficiency with reliable and precise motion and object modeling.

2 METHODS

In section 2.1 we present the lower level of our system, that uses dense image processing to identify IMOs in the image. In section 2.2 we present the feature based part of our system. Finally, in section 2.3 we present the IMO–based attention mechanism we implemented.

2.1 Detection of Independent Motion

Attributing image motion to moving objects is relatively straightforward when the observer is static, but

(a) optical flow   (b) dense stereo disparity

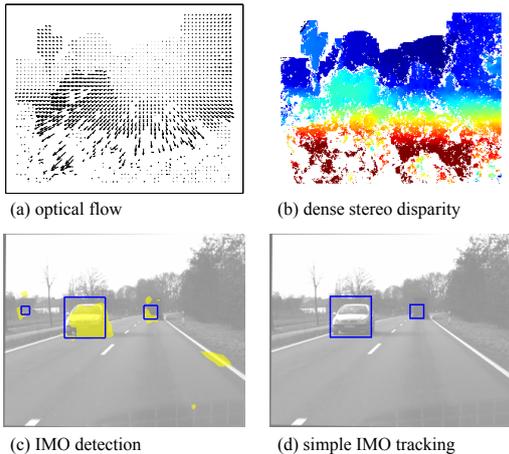(c) IMO detection   (d) simple IMO tracking

Figure 2: Illustration of the IMO detection level: a) the optical flow; b) the dense disparity; c) the low-level independent motion measure (yellow) and the detected IMOs after world-knowledge integration (blue boxes); d) the IMOs remaining after the simple tracking stage.

when the observer is moving, both moving objects and static environment generate image motion, and multiple cues must be combined to discriminate between both origins.

2.1.1 Visual Cue Combination

In our system, a pixelwise measure of independent motion is obtained on the basis of inconsistencies between the low-level cues dense optical flow (computed in the left image only), dense binocular disparity and camera motion, similar in spirit to (Thompson and Pong, 1990). We use phase-based techniques for the optical flow (Pauwels and Van Hulle, 2009) and disparity (Sabatini et al., 2007) estimation, and a continuous-time algorithm for the egomotion estimation (Pauwels and Van Hulle, 2006). The latter operates on the monocular optical flow field. Example optical flow and dense disparity results are shown in Figs. 2 a) and 2 b). These algorithms were chosen for their robustness to the nuisance factors typically encountered in real-world driving situations. In addition, GPU-based real-time implementations are available for the optical flow and disparity algorithms (Pauwels and Van Hulle, 2008). At each pixel $\mathbf{x} = (x,y)^{\mathrm{T}}$, a measure of independent motion, $I(\mathbf{x})$, is obtained by subtracting the optical flow component that results from the self-motion in the static environment from the estimated optical flow field $\mathbf{u}(\mathbf{x}) = (u_x, u_y)^{\mathrm{T}}$:

$$I(\mathbf{x}) = \|\mathbf{u}(\mathbf{x}) - d_D(\mathbf{x})A(\mathbf{x})\mathbf{t} - B(\mathbf{x})\boldsymbol{\omega}\| . \quad (1)$$

where $d_D(\mathbf{x})$ is the inverse depth obtained from dense disparity, and $\mathbf{t} = (t_x, t_y, t_z)^{\mathrm{T}}$ and $\boldsymbol{\omega} = (\omega_x, \omega_y, \omega_z)^{\mathrm{T}}$ are

the translational and rotational velocity of the moving observer, robustly estimated from $\mathbf{u}(\mathbf{x})$. Furthermore:

$$A(\mathbf{x}) = \begin{bmatrix} -1 & 0 & x \\ 0 & -1 & y \end{bmatrix}, \quad (2)$$

$$B(\mathbf{x}) = \begin{bmatrix} xy & -1-x^2 & y \\ 1+y^2 & -xy & -x \end{bmatrix} . \quad (3)$$

2.1.2 World-knowledge Integration

Spatiotemporal filtering is applied to the measure of Eq. (1) to reduce the noise. After thresholding, a number of independent motion blobs are obtained (yellow blobs in Fig. 2 c)). These blobs give a rough indication of an IMO's location. They are however quite noisy, with a significant number of false positives. To improve detection performance, these low-level detections are combined with world-knowledge. In particular, the ground-plane (as estimated from dense disparity) and expected object size, are used to obtain a more precise size and location estimate. In turn, the low-level cues are used to validate this box estimate (by evaluating the variance of the dense disparity within the box). The blue boxes in Fig. 2 c) correspond to the IMO-estimates obtained on the basis of the yellow blobs and the world-knowledge.

Finally, a simple tracking procedure is used that links the boxes across frames by connecting the detections to the track that is closest in 3D. Detections that are not consistently detected over a number of consecutive frames are removed. This reduces the number of false positives. The final detections after the tracking stage are shown in Fig. 2 d).

2.2 Feature Extraction and Tracking

In this section, we describe the second level of our system, that is composed of sparsely extracted features. The sparseness allows for the use of higher resolution (and therefore accuracy) than would be possible when processing each and every pixel.

2.2.1 Visual Primitives

The main features we use for describing the scene are local multi-modal edge and line descriptors called visual primitives (Krüger et al., 2004; Pugeault, 2008). These are extracted sparsely from the images along edges, and additional visual modalities, such as colour and phase, extracted at these locations are associated, to form a feature vector:

$$\boldsymbol{\pi} = (\mathbf{x}, \mathbf{t}, \phi, \mathbf{c}, \mathbf{f}) \quad (4)$$

where $\mathbf{x}$ is the 2D-primitive's position, $\mathbf{t}$ the local tangent to the contour, ϕ is the local phase, that encodes
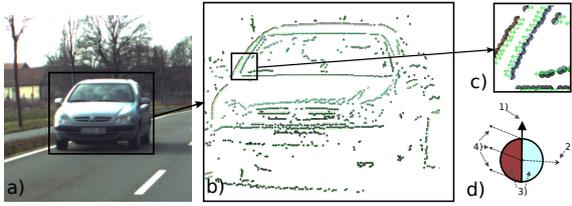
Figure 3: Illustration of the feature extraction level: a) one image; b) the 2D–primitives extracted; c) a detail of b); d) symbolic representation of a 2D–primitives, with 1-orientation, 2-phase, 3-colour, and 4-optical flow.

the contrast transition across the contour (e.g., bright to dark edge), **c** encodes colour information on both sides of the contour (sampled locally), and **f** encodes the local optical flow.

Such 2D-primitives are matched across stereo-pairs of images, allowing for the reconstruction of a 3D equivalent, called 3D-primitives, and described by the following feature vector:

$$\mathbf{\Pi} = (\mathbf{X}, \mathbf{T}, \Phi, \mathbf{C}) \qquad (5)$$

where $\mathbf{X}$ is the 3D-primitive's position, $\mathbf{T}$ the 3D tangent to the contour, Φ is the phase, and $\mathbf{C}$ encodes colour information of the 3D-contour. This process is illustrated in Fig. 3, where a) shows a detail of the image, b) the extracted 2D-primitives and c) a magnified detail. In d), the symbolic representation of primitives is illustrated, with 1) indicating orientation, 2) phase, 3) colour, and 4) optical flow. Moreover, SIFT features (Lowe, 2004) are also extracted from the images to allow for more robust matching.

2.2.2 Rigid Body Motion (RBM) Estimation

The motion of the camera and of the IMOs is evaluated using correspondences of 3D-primitives and SIFT features across time. In this case, because we consider only vehicles, we restrict ourselves to Rigid Body Motions (RBMs). The mathematical formulation of the RBM that we use is from (Rosenhahn et al., 2001), and has three advantages: First, the motion is optimised in 3D space; second, it allows for solving the motion jointly for different kind of constraint equations that stem from different type of image features (in this case, local edge descriptors and SIFT); third, it minimises the error directly in SE(3), and therefore does not require additional measures to handle degenerate cases. As been shown in (Pilz et al., 2009), a combination of heterogeneous features (edges and SIFT features) leads to an improved robustness and accuracy of the RBM estimate. Outliers are discarded using RANSAC (Fischler and Bolles, 1981).

2.2.3 Tracking and Filtering

All 3D-primitives are tracked using independent Kalman Filters (Kalman, 1960). The prediction stage is provided by the estimated motion. The position uncertainty of the 3D–primitives is re-projected in the image domain, into a 2×2 covariance matrix. Using this covariance matrix we estimate the likelihood for the 3D-primitive to find a match at each location by a normal distribution combined with a uniform distribution (that expresses the chance for a correct 3D–primitive not to be matched). We will write the fact that a primitive $\mathbf{\Pi}_i$ that predicts a primitive $\hat{\mathbf{\Pi}}_{i,t}$ at time t is matched (as described above) as $\mu_{i,t}$ and evaluate its likelihood as:

$$p[\mu_{i,t}] = \frac{e^{\frac{1}{2}(\Delta x)_{i,t}^{\top} \Sigma_{\Delta,i,}^{-1}(\Delta x)_{i,t}}}{(2\pi)\sqrt{|\Sigma_{\Delta,i,t}|}} + \beta \qquad (6)$$

The matrix $\Sigma_{\Delta,i,t} = \hat{\Sigma}_{x,i,t} + \tilde{\Sigma}_{x,i,t}$ is the sum of the re–projected position uncertainty for both the predicted ($\hat{\Sigma}_{x,i,t}$) and the observed ($\tilde{\Sigma}_{x,i,t}$) primitives in this image. In this equation, $\beta = p[\bar{\mu}|\mathbf{\Pi}]$ Also, $(\Delta x)_t = \hat{x}_{t|t-1} - \tilde{x}_t$ is the difference between the position of the two re–projected primitives, where $\hat{x}_{t|t-1}$ is the predicted re-position and $\tilde{x}_t$ is the position of the potential match. If the confidence $p[\mu_{i,t}]$ is larger than the chance value $\gamma = p[\mu|\Pi]$, the the match is considered valid. Furthermore, the similarity between the primitives (in orientation, phase, and colour) is also considered, and matches with a too low similarity (lower than $\tau = 0.9$) are disregarded.

Moreover the confidence in the existence of the accumulated primitive is updated depending on how consistently concordant evidence has been found in the image. The probability is evaluated from

$$p[\mathbf{\Pi}_{i,t}|\bar{\mu}_{i,t})] = (1 + \kappa_{i,t})^{-1} , \qquad (7)$$

where κ is evaluated recursively

$$\kappa_{i,t} = \frac{\gamma}{p[\mu_{i,t}]}\kappa_{i,t-1} . \qquad (8)$$

with $\kappa_1 = p[\Pi]$ is the prior probability that a 3D–primitive is correct.

If an hypothesis' confidence $p[\mathbf{\Pi}_{i,t}|\bar{\mu}_{i,t})]$ falls below a threshold $\tau_{\min}$, then it is deemed erroneous and discarded; if it raises above a threshold $\tau_{\max}$, then it is deemed verified up to certainty, and its confidence is not updated any more. This allows for the preservation of features during occlusion. This is effectively a soft version of the classical *n*-scan strategy in tracking (Reid, 1979).

Based on this filtering process, at the third level a 3D model of the object becomes accumulated and a final decision for the acceptance or removal of the IMO is made.

sequence T03   sequence S06
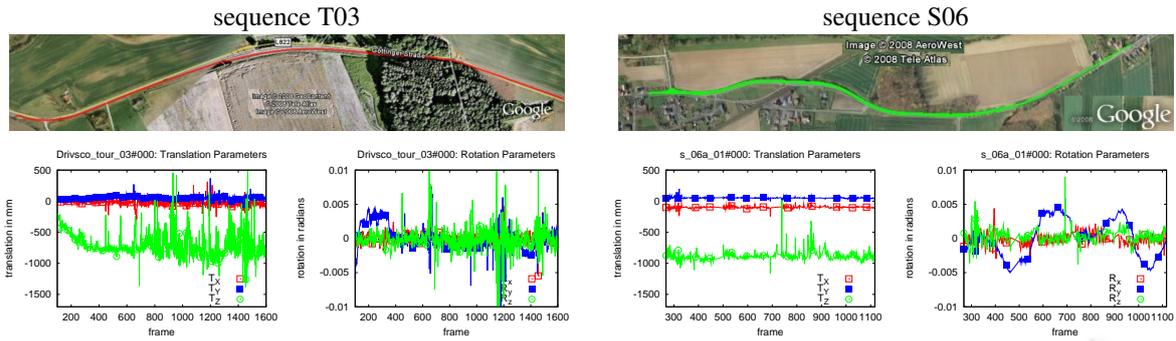

Figure 4: Top: The road on which our system was tested (from Google Earth) Bottom: Estimated vehicle egomotion: the estimated translation $T = (T_X, T_Y, T_Z)$, and the rotation provided as $R = \alpha\boldsymbol{a}$, where α is the rotation angle and $\boldsymbol{a}$ the axis.

2.3 Attention Mechanism

The complexity of operations at the feature level grows with the density of extracted features, therefore it is desirable to reduce the scale of processing or to downscale the images to achieve reasonable speed. On the other hand, when IMOs are very distant to the camera, they are represented by very small patches in the image, leading to low accuracy and robustness. In this work, we implemented a simple attention mechanism to optimise resource usage — for a review of attention mechanisms in computer vision, see, e.g., (Caputo and Lombardi, 1995).

The first, pre-attentive, layer of our system extracts IMOs from dense stereo and optical flow, on downscaled versions of the images. This is illustrated in Fig. 5a). The IMO extraction discussed in section 2.1 provides us with a list of attention vectors, of the form:

$$A_{i,t} = (x, y, w, h) \qquad (9)$$

where (x, y) designates the position of the ith IMO in the left image at time t, w and h the width and height of a corresponding region of interest (ROI). The position in the right image is given by the dense disparity estimate. The upper, feature–based layer use these ROIs to process the image at variable resolutions. Effectively, all ROIs are processed at twice the resolution as the background. The number of extracted 3D–primitives is assessed in each ROI, and ROIs with there are less than N 3D–primitives are discarded (experiments showed that $N = 20$ is an ideal value). In Fig. 5a) a discarded ROI is drawn as crossed in red, and the valid ones are drawn in green.

3 RESULTS

We have evaluated our system on two video sequences recorded on the two Drivsco road sequences

(a) IMO detection (b) left ROI (c) right ROI

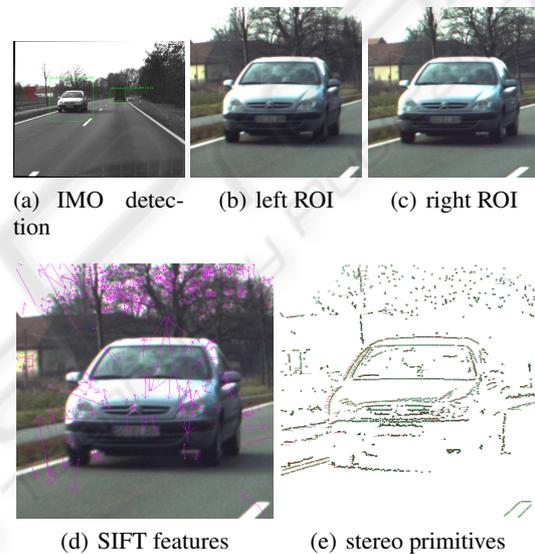(d) SIFT features (e) stereo primitives

Figure 5: Illustration of the attention mechanism: In a) the detected IMOs are denoted by bounding boxes in the image (red for deleted spurious IMOs, green for valid ones). In b-c), the region of interest (ROI) extracted from the detected IMOs are shown. In d), the red arrows indicate extracted SIFT features. In e), the red lines indicate stereo-correspondences of 2D–primitives.
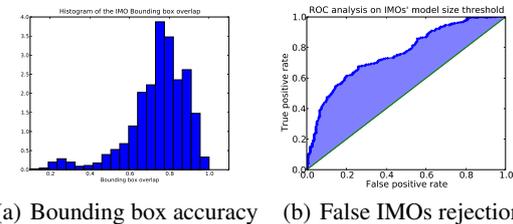
(a) Bounding box accuracy (b) False IMOs rejection

Figure 6: (a) Histogram of the bounding box accuracy. In this 1 represent a perfect overlap between the detected and hand–labelled bounding boxes. (b) ROC analysis of the IMO selection based on the number of 3D–primitives in the bounding box.

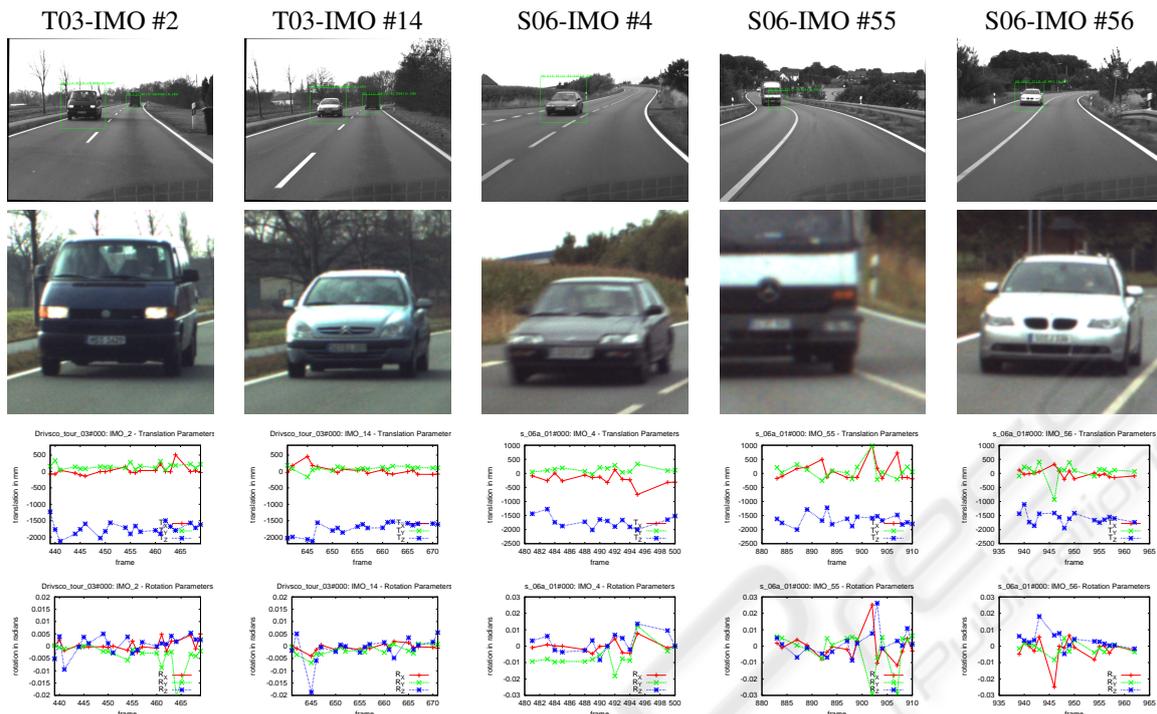| T03-IMO #2	T03-IMO #14	S06-IMO #4	S06-IMO #55	S06-IMO #56

Figure 7: Example IMOs, two of which were taken from the sequence T03, and three from the sequence S06. The first row shows the detected IMOs; the second the high resolution ROI; the third the estimated translation, provided as $T = (T_X, T_Y, T_Z)$; and the fourth the estimated rotation, provided as $R = \alpha \boldsymbol{a}$, where α is the rotation angle and $\boldsymbol{a}$ the axis.

T03 and S06 which are available from the web–page http://www.mi.auckland.ac.nz/eisats. The data sets being available from this web-page contain relevant data in connection with driver assistance systems. Hand–labeled ground truth for the individual IMOs is also given on this web–page. The images are taken from a pair of stereo cameras rigidly installed behind the front shield of a moving car — see Fig. 4. The baseline between cameras was 30 cm, the framerate 25Hz, and the images were undistorted and rectified. Each sequence contains 1500 stereo pairs of images. In Fig 4, the top row shows a bird eye view of the trajectory, taken from Google Earth.

We have evaluated the independent motion detection stage in terms of the precision/recall obtained on the entire (manually-labeled) sequence. A detection is considered correct if one third of the detected and labeled boxes overlap (intersection-over-union measure). If we do not include world-knowledge, and simply put boxes around the blobs, we obtain 0.15 precision and 0.62 recall (at maximum F-value). Clearly, this method suffers from many false positives. Adding world-knowledge greatly improves the results up to 0.88 precision and 0.80 recall. Finally, the simple tracking stage removes false positives at the cost of some recall. At maximum F-value we obtain 0.93 precision and 0.75 recall.

Moreover, we evaluated the accuracy of the generated bounding boxes using 2000 hand-labeled data images on two sequences. Fig. 6(a) shows a histogram of the overlap between the bounding boxes of correctly detected IMO's and the hand labeled ones. Fig 6(b) shows a ROC analysis (see, e.g., (Fawcett, 2006)) of the IMO rejection process based on the 3D structure present in the bounding box. Each point of the curve corresponds to a different threshold on the number of 3D-primitives required to validate the IMO. The axes show the false and true positive rates (ratio of remaining wrong and true IMOs, respectively). The diagonal represent the performance of the IMO detection of the first layer. Hence, the convexity of the curve shows that the selection process established at the two additional layers improves the classifier, and the horizontal part of the curve on the top–right corner shows that for low threshold (20 3D–primitives and more) false positives are discarded without any loss of true positives.

The RBM estimation provided a robust estimate, and RANSAC guarantees that at least 70% of the predictions have less than 3 pixels errors, which is sufficient for reliable tracking. The egomotion performance on both sequences is shown in Fig. 4, bottom
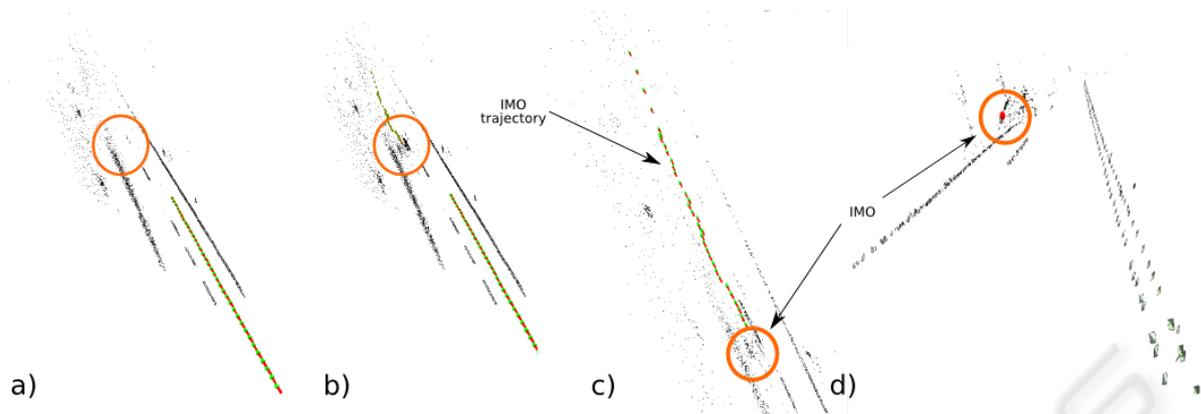
Figure 8: Illustration of a) the reconstructed road structure, b) and c) trajectories for the ego– and IMO–motion, and d) the reconstructed 3D model of the car with the lane structure.

row.

In the IMO case, some 40 SIFT features are extracted from the image patches, and between 10 to 20 of them are reliably matched both across stereo and time. Fig. 7 shows five of the detected and tracked IMOs, taken from the two sequences T03 and S06. Because this motion is computed visually, it effectively combines the vehicle's egomotion with the IMO's actual motion. As can be seen the motion estimates are qualitatively correct.

Fig 8 shows the reconstructed road for sequence T03 and IMO #4, with trajectories (string of green and red arrows) for both the egomotion and the IMO. Fig 8a) shows only the road and the egomotion trajectory. In Fig 8b) and c) show the road structure with the IMOs and their trajectory. Fig 8d) shows the reconstructed IMO and the reconstructed road structure. As can be seen, by the combination of all three levels, it is possible to give a qualitatively correct description of the ego–motion, the road structure, the 3D movement of the IMO as well as the relative position of the IMO to the road structure.

4 CONCLUSIONS

We presented a three level architecture that is composed of the following components: A dense, model free, IMO detection mechanism; an attention mechanism; and a sparse, feature tracking, motion estimation, and model building scheme. The combination of sparse and dense methods is shown to provide advantages in terms computational speed as well as precision: The attention mechanism guided by the low level allows for the processing of the motion at the highest available resolution, while processing the background at lower resolution. Finally, the system provides not only detection of the IMOs, motion and

trajectories and a feature based model of the IMO's shape.

REFERENCES

Bar-Shalom, Y. (1978). Tracking methods in multitarget environment. *IEEE trans. on Automatic Control*, AC-23:618–626.

Bertozzi, M. and Broggi, A. (1998). Gold: A parallel real-time stereo vision system for generic obstacle and lane detection. *IEEE TIP*, 7(1):62–81.

Betke, M., Haritaoglu, E., and Davis, L. S. (2000). Real–time multiple vehicle detection and tracking from a moving vehicle. *Machine Vision and Applications*, 12:69–83.

Caputo, G. and Lombardi, L. (1995). Attention mechanisms in computer vision systems. *Int. Workshop on Computer Architectures for Machine Perception*.

Comaniciu, D., Ramesh, V., and Meer, P. (2003). Kernel–based object tracking. *IEEE TPAMI*, 25(5):564–577.

Cox, I. J. (1992). A review of statistical data association techniques for motion correspondence. *IJCV*, 10(1):53–66.

Dissanayake, P., Newman, P., Durrant-Whyte, H., Clark, S., and Csorba, M. (2001). A solution to the simultaneous localisation and mapping (SLAM) problem. *IEEE trans. in Robotics and Automation*, 17(3):229–241.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874.

Fischler, R. and Bolles, M. (1981). Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Comm. of the ACM*, 24(6):619–638.

Hoiem, D., Efros, A. A., and Hebert, M. (2007). Recovering surface layout from an image. *IJCV*, 75(1):151–172.

Hoiem, D., Efros, A. A., and Hebert, M. (2008). Putting objects in perspective. *IJCV*, 80(1):3–15.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Jour. of Basic Engineering*, 82-D:35–45.

Krüger, N., Lappe, M., and Wörgötter, F. (2004). Biologically motivated multi-modal processing of visual primitives. *AISB Jour.*, 1(5):417–427.

Leibe, B., Schindler, K., Cornelis, N., and van Gool, L. (2008). Coupled object detection and tracking from static cameras and moving vehicles. *IEEE TPAMI*, 30(10):1683–1698.

Lourakis, M., Argyros, A., and Orphanoudakis, S. (1998). Independent 3d motion detection using residual parallax normal flow fields. In *ICCV*, pages 1012–1017.

Lowe, D. (2004). Distinctive image features from scale-invariant keypoints. *IJCV*, 1(1).

Montemerlo, M., Thrun, S., Koller, D., and Wegbreit, B. (2002). FastSLAM: A factored solution to the simultaneous localization and mapping problem. In *AAAI*.

Pauwels, K. and Van Hulle, M. (2006). Optimal instantaneous rigid motion estimation insensitive to local minima. *CVIU*, 104(1):77–86.

Pauwels, K. and Van Hulle, M. (2008). Realtime phase-based optical flow on the gpu. In *CVPR Workshop on Computer Vision on the GPU*.

Pauwels, K. and Van Hulle, M. (2009). Optic flow from unstable sequences through local velocity constancy maximization. *IVC*, 27(5):579–587.

Pilz, F., Pugeault, N., and Krüger, N. (2009). Comparison of point and line features and their combination for rigid body motion estimation. *Visual Motion Analysis, Springer LNCS 5604*.

Pugeault, N. (2008). *Early Cognitive Vision: Feedback Mechanisms for the Disambiguation of Early Visual Representation*. VDM Verlag Dr. Müller.

Reid, D. B. (1979). An algorithm for tracking multiple targets. *IEEE trans. on Automatic Control*, AC-24(6).

Rosenhahn, B., Krüger, N., Rabsch, T., and Sommer, G. (2001). Automatic tracking with a novel pose estimation algorithm. *Robot Vision*.

Rushton, S., Bradshaw, M., and Warren, P. (2007). The pop out of scene-relative object movement against retinal motion due to self-movement. *Cognition*, 105(1):237–245.

Sabatini, S., Gastaldi, G., Solari, F., Pauwels, K., Van Hulle, M., Diaz, J., Ros, E., Pugeault, N., and Krueger, N. (2007). Compact (and accurate) early vision processing in the harmonic space. In *VISAPP*, pages 213–220.

Sawhney, H., Guo, Y., and Kumar, R. (2000). Independent motion detection in 3d scenes. *IEEE TPAMI*, 22(10).

Thompson, W. B. and Pong, T. C. (1990). Detecting moving-objects. *IJCV*, 4:39–57.

Thrun, S., Liu, Y., Koller, D., Ng, A., Ghahramani, Z., and Durrant-Whyte, H. (2004). Simultaneous Localization and Mapping with Sparse Extended Information Filters. *IJRR*, 23(7–8):693–716.