Three-Level Early Fusion for Road User Detection

Rudi Lindl and Leonhard Walchshäusl

BMW Group Research and Technology, 80992 Munich, Germany
email: {rudi.lindl, leonhard.walchshaeusl}@bmw.de

This paper deals with a novel three-level sensor fusion approach in order to detect and track cars and pedestrians. The underlying perception system is composed of a far infrared imaging device, a laser scanner and several radar sensors, which operate integrated into a BMW sedan. At three different levels fusion is applied to approach the generation of a robust and accurate description of the area in front of the vehicle. Based on this environment perception a preventive safety application is outlined, which autonomously brakes in case of an inevitable accident.

1 Introduction

Statistic evidence of the European Union shows that accidents resulting in fatalities or serious injuries are caused to the highest percentage by collisions of cars with vulnerable road users. This fact points to the urgent need for active and passive automotive safety systems as a significant contribution to the overall road safety.

For this purpose the Preventive and Active Safety Applications project (*PReVENT*), an European automotive industry activity co-funded by the Sixth Framework Programme of the European Commission (EC), was established. Within the *PReVENT* subproject *COMPOSE*, one conceptual application aims at collision mitigation of cars by means of autonomous braking in case of inevitable pedestrian accidents or rear-end collisions in urban areas.

However, an erroneous application of emergency braking caused by false alarms would greatly impede road safety improvement not lastly due to the major setback such an incident would represent for driver acceptance. Therefore, an active autonomous intervention in the process of driving requires an outstanding degree of perception performance, particularly with regard to accuracy, availability and robustness.

Current off-the-shelf single sensor approaches can hardly fulfil these challenging demands. Accordingly, the potential of a multi sensor system in combination with a novel three-level early fusion approach is researched in this paper.

1.1 Related Work

Takizawa et al. [TYI04] proposes a fusion method for the detection of vehicles. Lidar data and image features are combined to a fusion vector which is classified by a principle component analysis. Although detected vehicles are tracked by a kalman filter, fusion is only utilized at a single level (within the classification process). A sensor fusion architecture based on bayesian networks is offered by Kawasaki [KK04]. A Bayesian network describes the fusion system in a causality model, which makes the fusion algorithm easy to understand. The proposed architecture and algorithm was tested with a perception system composed of millimeter wave radar plus vision sensor for vehicle tracking.

1.2 Overview

This publication focuses on a novel three-level early fusion approach based on only slightly pre-processed sensor data. In chapter 2 we briefly give a taxonomy on different sensor fusion techniques. The envisaged safety application is presented in chapter 3. In the following chapter 4 the sensor configuration and the resulting sensor data is discussed. Chapter 5 is dedicated to the novel three-level early fusion approach. After a short motivation with respect to early fusion 5, section 5.1 gives an overview of the tracking cycle and explains the three levels of fusion in more detail. Finally, the last section gives an overview on the system architecture and implementation details of the fusion system.

2 Sensor Fusion Taxonomy

Sensor Fusion comprises a very wide domain and one has to deal with many varieties. Elmenreich [Elm02] proposes an universal definition:

> "Sensor Fusion is the combining of sensory data or data derived from sensory data such that the resulting information is in some sense better than would be possible when these sources were used individually."

There are several ways to categorize sensor fusion approaches like regarding the point in time when the fusion is performed (see figure 1) or considering the interaction role between two sensors (see Brooks [BI97]).

2.1 Time based taxonomy

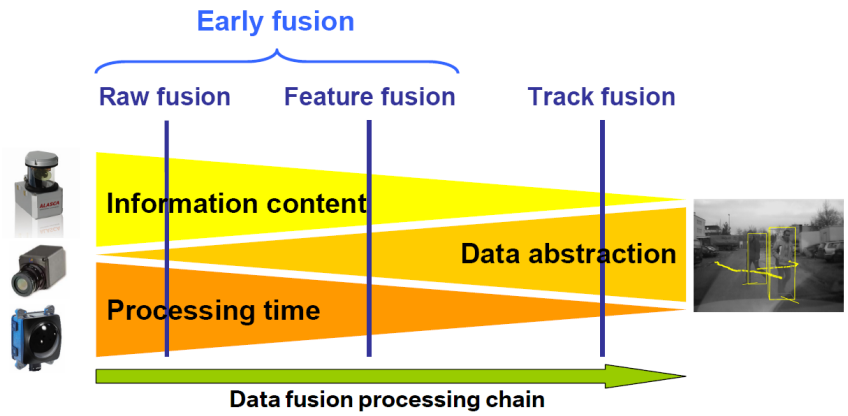


Figure 1: Time based fusion taxonomy.

Raw-data fusion In early or raw-data fusion systems data provided by multiple and even diverse sensors is combined at an early stage of the data processing chain. In addition, a joint data interpretation with respect to a common model basis is performed. Data from one sensor is assessed with regard to the relevance of its information, always in the context of data provided by other sensors.

Feature fusion Feature fusion combines various features such as edges, corners, segments or positions. These features are generated by a preprocessing which acts independently of the other sensors.

Track fusion In track-based or decision fusion approaches several sensor data streams are processed independently from each other until the level of object data is reached. Based on these independent results a common decision has to be made. At this point object-data rather than sensor data is combined.

2.2 Interaction based taxonomy

Complementary fusion If two sensors work independently from each other the fusion is called complementary. For example, two sensors surveying the environment in two non overlapping areas work in a complementary fashion.

Cooperative fusion In cooperative fusion systems multiple sensors are working together in a tight and coupled manner. Take two cameras for 3D reconstruction based on stereo computer vision algorithms as an example.

Competitive fusion Competition is introduced in a fusion system if sensors are operating redundant that is to say two or more sensors estimate the same object property. Strategies have to be introduced in order to solve the conflicts which arise if sensors disagree about object properties.

3 Collision Mitigation Application

The target application of the demonstration system is collision mitigation by means of autonomous braking. According to a german accident analysis [Bun01], most of the accidents with vulnerable road users happen in urban areas on straight, unprotected roads. Therefore, the collision mitigation application will mainly focus on the prementioned scenarios.

The basis for the envisaged autonomous braking is a probabilistic situation assessment. Only if an inevitable collision is detected, the system engages the brakes autonomously. Subject to the condition that a *perfect* environment perception would be educible, this would have high potential to attenuate or even prevent accidents, since machines are capable of reacting much faster and more efficiently than human drivers,

4 Perception System

The central challenge for many advanced driver assistance systems is an adequate perception of the vehicle's environment, a high degree of reliability and last but not least a high degree of measurement precision. One of the key factors to meet these requirements is a multi sensor perception system which is explained in the following section.

4.1 Sensor Configuration

BMW has set up an experimental car with the following sensor configuration (see figure 2) to research the potential of multi-sensor perception.

Concentrating on the surveillance of the area in front of the vehicle, these cooperative sensors, which operate on the basis of distinct physical principles, complement each other both in effective range and spatial accuracy.

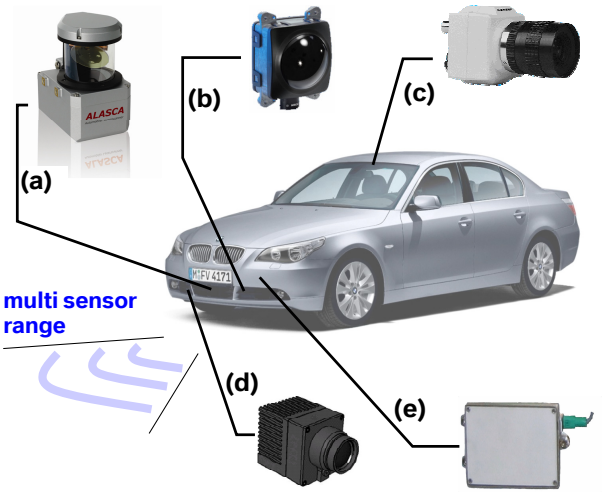


Figure 2: BMW experimental car equipped with the following sensor configuration: (a) laser scanner, (b) long range radars, (c) grey-scale camera, (d) far infrared camera, (e) short range radars.

The usage of a far infrared (FIR) sensor guarantees both perception at bad lighting conditions and straightforward vehicle and pedestrian detection since they have a characteristic signature regarding their temperature (exhaust system respectively uninsulated body parts as head and limb). As most pedestrian scenarios covered by the experimental vehicle are situated in the area to the right side of the road, this sensor is mounted at the right of the frontal bumper. Long and short range radar sensors are surveying the environment ahead providing a seamless transition in distance and field of view resolution. Moreover, a laser scanning (lidar) device is mounted beneath the number plate to enhance the detection and tracking quality for both pedestrians and vehicles. The visual grey-scale cameras are currently used for supervising and controlling purposes only.

4.2 Sensor Data

In the following a short survey of the different types of sensor data in combination with their preprocessing is given.

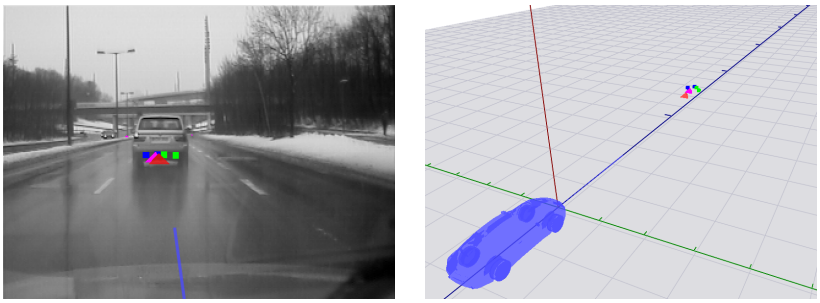
4.2.1 Radar

The radar sensors provide information about the relative position $\vec{p}_r$ and relative velocity $\vec{v}_r$ of an object (see figure 3(a) and 3(b)). Accordingly, a radar measurement M_r is defined as following:

$$M_r = (\vec{p}_r, \vec{v}_r) \tag{1}$$

4.2.2 Lidar

The lidar sensor is capable of providing up to 1400 reflection points of the scanned environment. In order to reduce the cost of computation in the subsequent tracking process correlated raw measurements are aggregated to single lines. Several connected lines

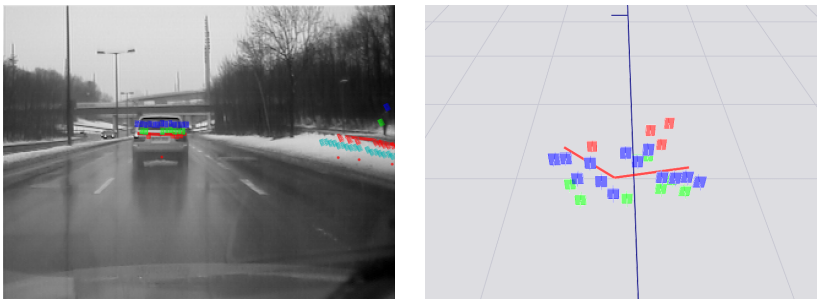


(a) (b)

Figure 3: Radar reflection of a vehicle. Green and blue boxes represent short range radar reflections. Red and magenta triangles are long range radar reflections. (a) Radar reflections projected into grey-scale image. (b) Radar reflections within the virtual 3D environment.

$l_1, l_2, \ldots, l_n$ are combined to segments (see figure 4(b)). With respect to the nomenclature of graph theory this segment represents a simple path. A lidar segment measurement M_l is defined as following:

$$M_l = (l_1, l_2, \ldots, l_n) \tag{2}$$



(a) (b)

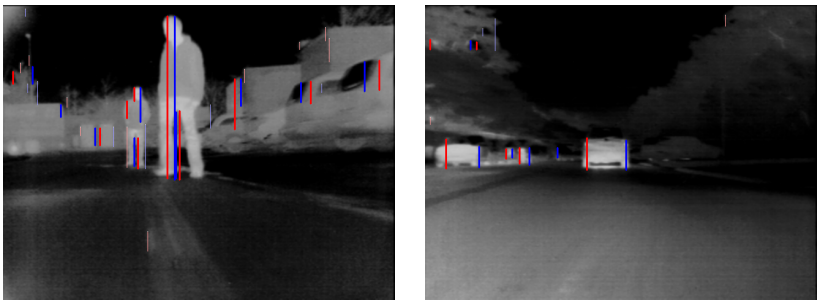
Figure 4: Reflections and segment data of a vehicle generated by a four-layer lidar sensor. Green, red, yellow and blue boxes represent the lidar echo at different layers. The red line illustrates the result of the preprocessing (segment data). (a) Lidar responses projected into grey-scale image. (b) Lidar reflections and segments within the virtual 3D environment.

4.2.3 Far infrared

Vertical edges ($\hat{E}_+$ and $\hat{E}_-$) with positive respectively negative gradient are extracted from the far infrared image by a sobel operator (see figure 5). A subsequent coarse pre-classification step rejects irrelevant edges. Within the early fusion processing (see section 5) a common three-dimensional sensor data description is necessary. Therefore, a projection converts image plane edges $\hat{E}$ into their corresponding three-dimensional representation E. Accordingly, the height H of an image edge $\hat{E}$ is calculated. A far infrared measurement

M_{f+}, respectively M_{f-}, is defined as following:

$$M_{f+} = (E_+, H_+) \quad (3)$$
$$M_{f-} = (E_-, H_-) \quad (4)$$



(a) (b)

Figure 5: Edges with positive (red lines), respectively negative (blue lines), gradient extracted from a far infrared image. (a) Edges of a pedestrian. (b) Edges of a vehicle.

5 Three-Level Early Fusion System

In contrast to track-based fusion (see section 2) "early fusion" combines data provided by multiple and even diverse sensors at an early stage of the data processing chain and performs a joint data interpretation with respect to a common model basis (cf. [WLVT06]). In doing so, signatures of various sub-threshold findings in the data processing chain may interfere constructively and thereby contribute to an above-threshold result to form a distinctive, well-recognized object instantiation. Thus, an increase of robustness, reliability and consistency in the environment perception is expected as the input from an individual sensor can be processed in view and with the help of the other sensors.

To come up with this early fusion demand we enhanced several of the basic tracking steps mentioned in the following.

5.1 Levels of Fusion

Generally speaking tracking can be performed by three circular steps namely time prediction, data association (data matching) and measurement update (correction) [FP02]. On top of this basic pattern we added further steps to cope with early fusion and multi-object demands(see figure 6).

Fusion is utilized at three different levels (hypotheses generation, classification and measurement update) of the tracking system (see highlighted steps of figure 6). Firstly, fusion during the hypotheses generation improves the system response time as the intial guess can be estimated more precisely. Secondly, the classification of objects is more robust if features of all available sensors are taken into account. Finally, fusion at the filtering level provides more precise and high available tracking results as redundant and complementary sensor data is combined. These extensions as well as the fundamental structure of the system are discussed in the subsequent sections.

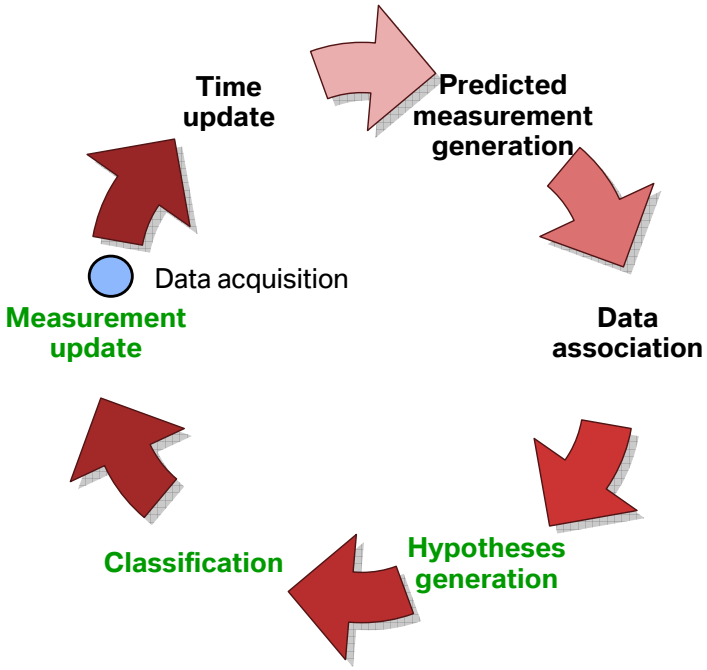


Figure 6: Overview of the tracking cycle. The cycle starts at the blue circle with the data acquisition. The emphasized components hypotheses generation, classification, and measurement update mark the three level of fusion.

Data acquisition: As most of the used sensors are working on different clock rates and time is crucial in collision mitigation applications, we preserve a high time resolution by a semi-asynchronous data acquisition. The actual data acquisition is done by polling every sensor for new data.

Time update: According to every objects' state (position, orientation, velocity, etc.) at the previous cycle, these states have to be estimated for the current time. As an example, this can be performed on the basis of the objects' underlying dynamic models.

Predicted measurement generation: In the previous step for every object an updated representation (state) with regard to the current time is generated. These estimated states are the basis for the following step, which predicts what each sensor would measure under the assumption that every objects' state was correctly estimated.

Data association: The next step within the aforementioned tracking cycle is the data association that extracts and assigns corresponding pairs of real and predicted measurements. An algorithm from Hopcroft [HK73] for maximum matchings in bipartite graphs is used for this purpose.

Hypotheses Generation: A priority goal of the hypoth-eses generation is a direct and complete detection of all so far untracked and possibly relevant objects in the sensors' ranges. Thereto, a high error of second kind is consciously taken into account. Usually, a subsequent classification procedure as well as an observation of the objects over time can select and eliminate irrelevant assumptions. To limit the cost of

computation, the hypotheses generation focuses on salient and unmatched sensor data.

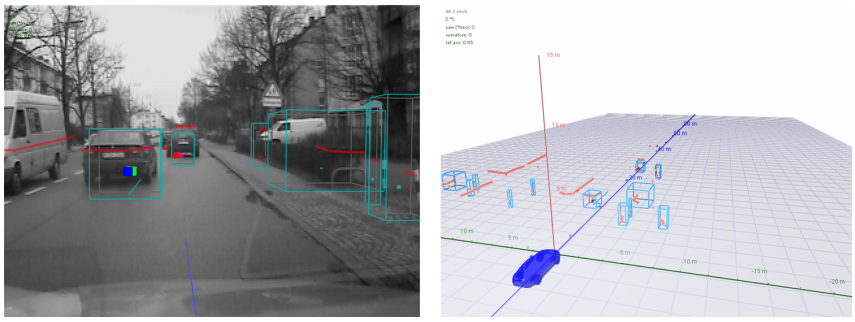


Figure 7: Example for hypotheses generation and aggregation. The cyan boxes represent new instantiated hypotheses. An aggregation of a lidar segment M_l and radar responses M_r instatiate the car hypothesis ahead resulting in an initial guess for both hypothesis' dimension and velocity.

Currently, unmatched salient measurements where new assumptions are placed, are radar responses M_r, lidar segments M_l within certain dimensions and pairs M_f of vertical edges from the far infrared imaging device. An aggregation step tries to combine overlapping hypotheses to one hypothesis. The initial state for this new hypothesis is composed of all measurements from all involved sensors (compare the illustration 7). Therefore, the oscillating phase caused by the Extended Kalman Filter may be shortened.

Classification: Classification occurs at three different phases in the data processing pipeline (see figure 8). According to the particular demands for the recognition, adequate and adapted classifiers are utilized.

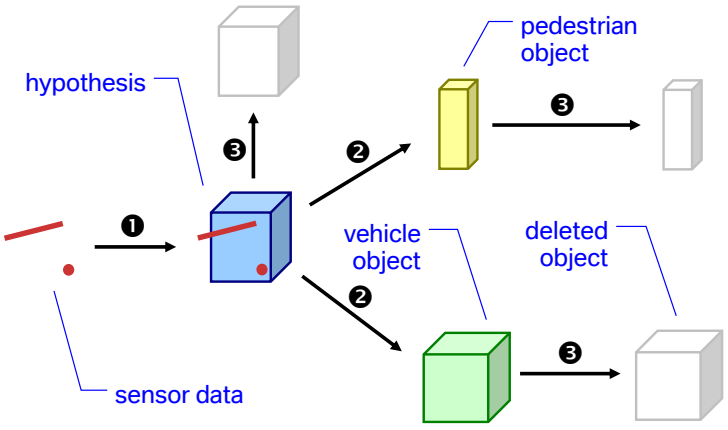


Figure 8: Classification phases of an object life cycle. (1) coarse pre-classification, (2) hypotheses classification, (3) object revalidation.

1. Hypotheses are initialized on salient and unmatched sensor data. A first coarse pre-classification step rejects impractical assumptions by checking if the width or height of a hypothesis lies below a threshold θ_d. This simple criterion ensures

a very efficient processing resulting in a high throughput. Furthermore, a high error of second kind is consciously taken into account since a direct and complete detection of relevant objects is mandatory.

2. Relevant objects (pedestrians and motorcars) are determined by a recognition process on active hypotheses. For that reason certain state components of a hypothesis are taken as feature input for a decision tree. The feature vector is composed of the hypothesis age h_a, velocity $\vec{h_v}$, dimension h_w, variance h_{σ^2} and the existence of adjacent far-infrared image edges. These features are derived from different sensor measurements.

3. Classified objects and hypotheses are checked for their confidence values in regular time intervals. This step is performed by an object revalidation process. Hypotheses and classified objects are removed from the virtual environment, if they are no longer supported by sensor data over a certain period of time.

Measurement update: A conventional Extended Kalman Filter (EKF) (see [WB95] for instance) has been chosen since it handles the nonlinearities of this application quite well. For every assigned pair of real and predicted measurement, which has been calculated before, a measurement update on the underlying object is performed. In doing so, the information of several measurements enhance the states by updating the objects' state values and furthermore, lowering the estimation error covariances. Thereby, for each assigned sensor data a measurement update step is conducted before the next cycle starts with the object's state prediction in time. With the notation of [BW95] the equations at time step k of the EKF's measurement update can be written as

$$K_k = P_k^- H_k^T (H_k P_k^- H_k^T + R_k)^{-1} \quad (5)$$

$$\hat{x}_k = \hat{x}_k^- + K_k(y_k - h(\hat{x}_k^-)) \quad (6)$$

$$P_k = (I - K_k H_k) P_k^- . \quad (7)$$

The specific term $h(\hat{x}_k^-)$ has already been evaluated during the calculation process of predicted measurements and thus equation (6) can be written as

$$\hat{x}_k = \hat{x}_k^- + K_k(y_k - y_k^-) \quad (8)$$

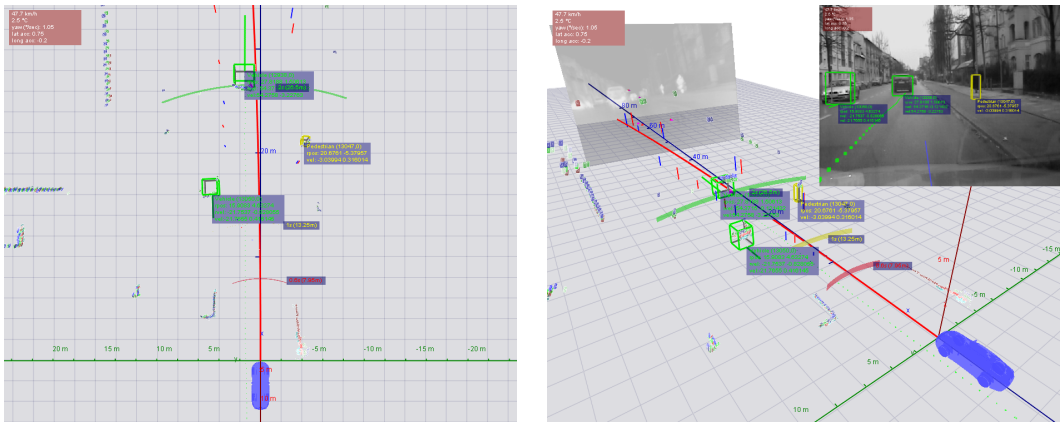
for every pair (y_k, y_k^-) of measurement and predicted measurement, matched by the data association process. As all sensor data is projected into the 3D global world coordinate system, the entries of the Jacobian H_k can be easily deduced from the underlying object-model without any further complex and time consuming calculations.

5.2 Implementation Details

A cyclic top-down architecture has been implemented to facilitate the detection, classification and tracking of relevant road users over time. The real world vehicle surroundings and the sensor configuration are reflected by a virtual environment, which is modeled as a hierarchical scene-graph structure [BW95], ensuring centralized data access and efficient spatial dependency processing. Furthermore, a vector-quaternion-scalar (VQS) [RH94] representation has been chosen in order to achieve coordinate system transformations

between the entities of the scene-graph. The topological object modeling is based on a winged-edge representation [BGZ02]. Essential tasks, like sensor coordinate transformation, clipping or occlusion testing can be easily performed and adapted for a specific sensor, since both the topological and the spatial modeling is widely-used in computer graphics.

To allow an efficient graph traversal as well as a decoupling of algorithm and data portions, the so called Visitor Design Pattern [BMRS96] has been used extensively. The visualization component (see figure 9) was implemented in OpenGL. This renders an acceleration by a 3D graphics adapter possible and consequently disburdens the central processing unit.



(a) (b)

Figure 9: Screenshots of our demonstrator *iFuse* visualizing the virtual automotive environment. (a) Bird's eye view with own car (blue), sensor data and detected objects. (b) Same scene as (a) from an isotropic perspective with far infrared visualisation and gray-scale camera projection.

6 Conclusion

This paper proposed a novel three-level early fusion approach to detect and track cars and pedestrians in real-time. Early fusion is applied at three different levels of a common tracking approach. Firstly, fusion during the hypotheses generation has shown to improve the system response time as the intial guess can be estimated more precisely. Secondly, the classification of objects is more robust since features of all available sensors are taken into account. Finally, fusion at the filtering level provides more precise and high available tracking results as redundant and complementary sensor data is combined. However, it has to be considered that due to the high amount of raw sensor data real-time demands are difficult to preserve.

7 Further Work

The classification quality and the system response time can be further improved by utilizing more complex classifying algorithms like neuronal networks or suport vector machines. Further research is needed in order to evaluate the suitability of these algorithms with

respect to multi-sensor demands. In addition, an increased set of object types like trucks, cyclists or motor-cyclists will improve the granularity of the perception system and could allow for the conceptual implementation of other applications. Within the hypothesis generation step, the potential of a Kalman filtered measurement aggregation has to be evaluated. Object dependent filtering techniques like a particle filter will be applied in order to achieve a more robust and granular tracking. In addition to these improvements, an extensive evaluation of the perception system is planned.

8 Acknowledgement

The three-level fusion approach presented in this publication is part of the main results achieved in the *COMPOSE*-project which is an application-driven subproject of the *PReVENT* Integrated Project, an automotive initiative co-funded by the European Commission's Sixth Framework Programme for active road safety. *COMPOSE* aims at collision mitigation and protection of vulnerable road users by (semi-) automated braking and to this end develops robust and reliable environment perception systems. The one which is based on a novel three-level early fusion approach is presented in this paper.

References

[BGZ02] H.J. Bungartz, M. Griebel, and C. Zenger. Einfuhrung in die Computergraphik. *Wiesbaden: Vieweg*, 2002.

[BI97] Richard R. Brooks and S. Sitharamar Iyengar. Real-time distributed sensor fusion for time-critical sensor readings. In *Optical Engeneering*, volume 36, pages 767–779, March 1997.

[BMRS96] Frank Buschmann, Regine Meunier, Hans Rohnert, and Peter Sommerlad. *Pattern-Oriented Software Architecture: A System of Patterns*, volume 1. John Wiley and Sons Ltd, 1996.

[Bun01] Statistisches Bundesamt. Verkehrsunfälle, Dezember und Jahr 2001. Technical Report 8, Statistisches Bundesamt, 2001. Reihe 7.

[BW95] B.D. Allen Gary Bishop and Greg Welch. Tracking: Beyond 15 minutes of thought: Siggraph 2001 course 11. Technical report, University of North Carolina at Chapel Hill, 1995.

[Elm02] Wilfried Elmenreich. *Sensor Fusion in Time-Triggered Systems*. PhD thesis, Technische Universität Wien, 2002.

[FP02] David A. Forsyth and Jean Ponce. *Computer Vision: A Modern Approach*. Prentice Hall, 2002. FOR d 02:1 1.Ex.

[HK73] John E. Hopcroft and Richard M. Karp. An $n^{5/2}$ algorithm for maximum matchings in bipartite graphs. *SIAM J. Comput.*, 2(4):225–231, 1973.

[KK04] Naoki Kawasaki and Uwe Kiencke. Standard platform for sensor fusion on advanced driver assistance system using bayesian network. In *2004 IEEE Intelligent Vehicles Symposium Proceedings*, pages 250–255, University of Parma, Italy, June 2004. IEEE.

[RH94] Warren Robinett and Richard Holloway. The visual display transformation for virtual reality. Technical Report TR94-031, University of North Carolina at Chapel Hill, Department of Computer Science, Chapel Hill, NC, USA:, 10 1994.

[TYI04] Hiroomi Takizawa, Kenichi Yamada, and Toshio Ito. Vehicles detection using sensor fusion. In *Proceeding of IEEE Intelligent Vehicles Symposium 2004*, pages 238–243, Parma, Italy, June 14-17 2004. IEEE.

[WB95] Greg Welch and Gary Bishop. An introduction to the kalman filter. Technical Report TR95-041, University of North Carolina at Chapel Hill, Department of Computer Science, Chapel Hill, NC 27599-3175, 1995.

[WLVT06] Leonhard Walchshäusl, Rudi Lindl, Katrin Vogel, and Thomas Tatschke. Detection of road users in fused sensor data streams for collision mitigation. In Jürgen Valldorf and Wolfgang Gessner, editors, *Proceedings of the 10th International Forum on Advanced Microsystems for Automotive Applications (AMAA'06)*, pages 53–65, Berlin, April 2006. VDI/VDE/IT.