2020 IEEE International Conference on Robotics and Automation (ICRA)
31 May - 31 August, 2020. Paris, France

Low Latency And Low-Level Sensor Fusion For Automotive Use-Cases

Matthias Pollach*, Felix Schiegg* and Alois Knoll*
*Chair of Robotics, Artificial Intelligence and Real-time Systems, Technical University Munich, Munich, Germany
Email: matthias.pollach@tum.de, felix.schiegg@tum.de, knoll@in.tum.de

***Abstract*—This work proposes a probabilistic low level automotive sensor fusion approach using LiDAR, RADAR and camera data. The method is stateless and directly operates on associated data from all sensor modalities. Tracking is not used, in order to reduce the object detection latency and create existence hypotheses per frame. The probabilistic fusion uses input from 3D and 2D space. An association method using a combination of overlap and distance metrics, avoiding the need for sensor synchronization is proposed. A Bayesian network executes the sensor fusion. The proposed approach is compared with a state of the art fusion system, which is using multiple sensors of the same modality and relies on tracking for object detection. Evaluation was done using low level sensor data recorded in an urban environment. The test results show that the low level sensor fusion reduces the object detection latency.**

***Index Terms*—sensor fusion, object detection, Bayesian networks**

I. Introduction

Every year more than 1.25 million people die in traffic accidents and roughly 20 million are seriously injured. The main cause of traffic accidents is human failure, with a rate of more than 90% [7]. Over the last decade, politics and industry have invested a significant effort to reduce the number of accidents. One outcome is advanced driver assistance systems (ADAS), which help decrease the number and the severity of accidents by alerting and supporting the driver [11]. These figures are only considering traffic but they are representative for the significant impact automation offers for society.

One of the main challenges for any automated or autonomous vehicle is to perceive the environment with low latency. A key enabler for this task is the detection of objects proximate to the vehicle. Fusing data from different sensors and different sensor modalities allows to achieve a reliable and low latency object detection.

There are various approaches towards sensor fusion. One popular method is to fuse the output of smart sensors. Such sensors have major computational capabilities and signal processing included, in order to detect objects. This is state of the art in currently available and upcoming ADAS systems [11]. Nevertheless, it is essential to process all available information to increase the detection confidence while reducing the latency. The time delay until an object is perceived by any AD system is of great interest. The use of smart sensors cannot guarantee that all captured information is used for the decision making process because each sensor filters data without taking information from other sensors into account. Some smart sensors rely on object tracking for object detection which negatively impacts their detection latency. At the same time it offers benefits like the improved handling of partial or entire object occlusions. Directly coupling the sensors to a centralized fusion system without performing object detection on sensor level has the potential to reduce detection latency and avoids losing valuable information before the fusion step, increasing object detection confidence. This concept is generally referred to as a low level sensor fusion concept and helps overcome specific weaknesses of individual sensor technologies.

The main contribution of this work is a a low level sensor fusion for object detection in a road vehicle environment with reduced detection delay in comparison to smart sensors. The results are compared to a commercial state-of-the-art light detection and ranging (LiDAR)-based sensor fusion solution. The test data was captured in urban environments at different times of day guaranteeing a variety of different weather and lighting conditions. The proposed fusion approach uses LiDAR, radio detection and ranging (RADAR) and camera data at a low level, which enables the object detection to be based on all available information while decreasing detection delay.

A combined temporal and spatial association is proposed and evaluated because the sensor setup of the test vehicle is not synchronized in time. In addition, a probabilistic fusion network processing the associated data is described and used for evaluation.

II. Related Work

A. Object detection

Object detection has experienced significant improvements over the past years using two-stage and single stage-detectors. Single-stage detectors like YOLO [6] or SSD [5] do not have a separation of object detection and classification. The achievable results of these architectures are trailing in accuracy when compared to two-stage architectures. The two-stage networks are more complex and use a object detection or object proposal and a classification step. The two stage detectors were introduced with R-CNN [1] and experienced improvements in various ways. Some of theses well known improved networks are Fast R-CNN [2], Faster R-CNN [3] or Mask R-CNN [4] to name some of the many developments.

978-1-7281-7395-5/20/$31.00 ©2020 IEEE

B. Low level sensor fusion

Sensor fusion is a well understood problem and has been an active field of research for decades. The fusion of data from multiple sources, generally consists of three main steps: The spatio-temporal registration, the alignment and the association of data to corresponding targets, followed by the state estimation and prediction step [15]. Sensor fusion can happen on different levels, depending on the available information. The levels range from a low level of information, like unprocessed sensor data, to a high level information, like objects. Examples of low level sensors fusion have been a field of research for automated robotic applications for decades [21] and use a variety of different approaches, ranging from basic mathematical concepts as in [15] to fuzzy logic [37]. Fusion is an enabler for latency reduction, pattern preservation and information gain and offers the possibility to reduce bandwidth issues in embedded platforms.

III. Proposed approach

A. Sensor Specific Pre-Processing

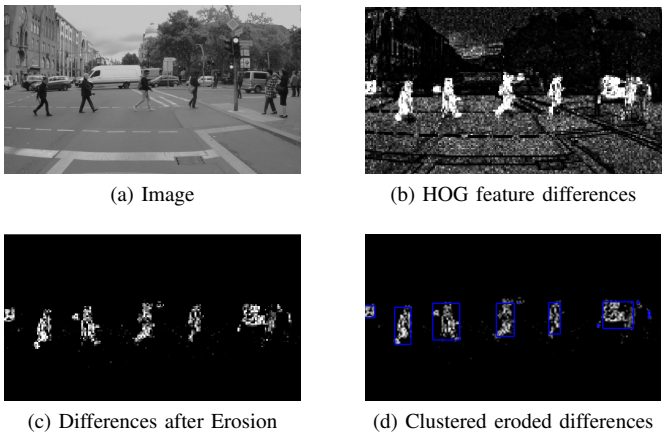
The sensor data is processed, associated and fused. These steps are sensor specific and the following describes how the data is pre-processed before it is associated and fused.

1) Camera Processing: A global shutter grayscale camera is used as a basis throughout this work. There is a large variety of different image processing techniques and applications, this present work only uses HOG features from the camera, in order to demonstrate the advantages of the proposed low level sensor fusion concept on LiDAR and RADAR. The camera data is pre-processed and serves the purpose of demonstrating the possibility to associate camera data with other sensor modalities.

A detailed description of HOG features can be found in [33]. The focus of this work is to detect vulnerable road users and vehicles. The implementation is based on the results found in [10]. This parameterization is not optimized because the output of the camera is only used to demonstrate the general possibility to fuse camera data with other sensor modalities. However, a main difference to the regular use of HOG features is that they are not directly used for object detection. Instead, the eroded difference of HOG features of two consecutive image frames is used. These differences are clustered and a bounding box for each cluster is calculated. This approach is referred to as HOG frame to frame difference(F2FD) throughout this work. Static objects cannot be detected whenever the car is not in motion, which is a clear limitation of this method. Nevertheless, the object detection latency requirements are less crucial whenever the vehicle is not in motion. The object detection for this type of scenario is based on LiDAR and RADAR only. As an improvement to this method, the camera could switch to a neural network based object detector.

2) LiDAR Processing: The LiDAR data measurements are clustered for further processing using the density based clustering algorithm DBSCAN [12]. This choice is based on the possibility to parallelize the algorithm for large amounts of data while being deterministic. In addition, DBSCAN is robust against noise and outliers while handling an unknown number of clusters. Following the proposed parameterization from [12], a minimum number of points $MinPts = 4$ is chosen. This reduces the loss of object detection-relevant information before the association step.

The computed clusters are represented by 3D bounding boxes, which are created for each cluster to enable an efficient association process.



(a) Image (b) HOG feature differences

(c) Differences after Erosion (d) Clustered eroded differences

Fig. 1. HOG F2FD Implementation: In (a) the raw camera image is shown. (b) depicts the HOG feature differences without any further processing. In (c) the HOG F2FD after erosion is depicted. Figure (d) depicts the bounding boxes, which are based on clusters resulting from the DBSCAN used on the data depicted in (c).

3) RADAR Processing: The RADAR sensor provides a list of targets, which is the result of the peak detection algorithm applied to the 2D Fourier transformed signal [36]. Each target is associated with a SNR value, a radial velocity vector, a distance and angle measurement. In a first step, RADAR targets are clustered to identify targets originating from the same physical object. However, RADAR targets are substantially different from LiDAR point clouds. Consequently, the parameterization needs to be adapted. For the present use case the parameters were set to $MinPts = 1$ and $epsilon = 2.75$. Clusters with a single target are referred to as single targets in the following. The clusters containing two or more targets are associated with the mean SNR value of all measurements of the respective cluster.

In many cases, it is not possible to group RADAR targets, as there may only be one target per object. Furthermore, there are targets captured by the sensor, which do not correspond to any real world object. Those targets are often referred to as ghost targets. All targets and clusters are provided to the association.

B. Proposed Association Method

The focus is a real time fusion before tracking which is a contrast to many state-of-the-art association methods. The proposed approach associates data from multiple sensor modalities running asynchronously from 2D and 3D representations.

Temporal offsets result in spatial offsets, whenever an object moves or the ego vehicle is in motion. The proposed approach

combines spatial and temporal alignment for association. The input are measurements from individual sensors, which are already associated with a bounding box.

Each bounding box has an area of A_n. For the 3D case, the same simplification is valid for the given use case because all objects move on the ground. Therefore, following the concept of an occupancy grid based fusion approach, 3D bounding boxes are reduced to 2D bounding boxes for the purpose of association. No relevant information for the association process is lost while the efficiency is increased simultaneously. The overlapping ratio $overlap_k$ of two bounding boxes is referred to as $\triangle A$. The overlap ratio for bounding boxes is computed as follows:

$$overlap = \frac{\triangle A}{min\left(A_1, A_2\right)} \tag{1}$$

Additionally, the distance $distance_k$ between the geometric centers C_n of bounding boxes is computed, relative to the size of the bounding boxes. The diagonal $diag_n$ expansion of a bounding box is determined based on its width and length.

A weighted $distance_k$ between two centers is computed according the following equation:

$$distance_k = \frac{d\left(C_1, C_2\right)}{max\left(diag_1, diag_2\right)} \tag{2}$$

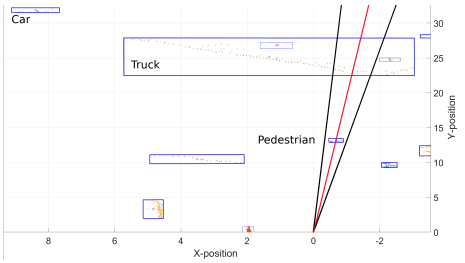
In addition to the geometric relationship, it is necessary to take the age of the data into account during association. The fastest sensor, running at f_{max} and introducing a maximum time difference between measurement frames of t_{fast}, is used as a reference for associating the data. Any other sensor measurements are available at a frequency of f_{Sensor_n} and have a time difference between frames of t_{delay}. The older the data is, that has to be associated with the newest incoming data, the smaller the threshold for association gets. This results in a more tolerant association for increasing time discrepancy between available sensor measurements, accounitng for motion of dynamic objects and the ego vehicle. The computed $overlap_k$ and $distance_k$ are used to compute association thresholds according to the following:

$$t_{overlap_k} = \alpha \cdot \frac{t_{fast}}{t_{delay}} \frac{\triangle A_k}{min\left(A_1, A_2\right)} \tag{3}$$

$$t_{dist_k} = \beta \cdot \frac{t_{delay}}{t_{fast}} \frac{d\left(C_1, C_2\right)}{max\left(diag_1, diag_2\right)} \tag{4}$$

Data is associated, if the center distance of two bounding boxes is smaller than the threshold t_{dist_k} or if the overlap ratio is larger than the threshold $t_{overlap_k}$. In this case α and β are tuning parameters to adjust the threshold for the desired application. When using this approach, the association process is timed on the fastest sensor. Whenever a new set of measurements from the fastest sensor is available, the system has to check, whether any other sensor has provided data in the meantime. Assuming that the individual sensor frequencies and latencies are known and assuming that all objects have a limited range of velocity, it is at all times possible to determine the worst case association

(a) Detected Objects in 2D Plane



(b) Orthographical View

Fig. 2. Association of 2D and 3D Data: In (a) a bounding box is shown. The corresponding bounded 3D LiDAR data is visualized in (b) and shows the resulting association using the projected camera rays in the orthographical perspective. The camera rays associated to the left and right edge of the bounding box are depicted in black, the ray corresponding to the middle of the bounding box is shown in red. Only data originating from the pedestrian is associated.

error, which determines the location uncertainty of an object. As a consequence of this proposed association method, the estimated object size will vary depending on the age of data, but the influence of data age on wrongly assigning measurements is minimized.

The prerequisite for the association process is the access to data in one reference coordinate system. Representing many measurements with a bounding box allows to efficiently associate data, originating from the same source. Depending on the sensor modality, the bounding box is either 3D or 2D. In order to associate data from 2D and 3D space, a perspective projection, according to the camera pin hole model, has to be used. This allows to associate camera data with 3D data.

This proposed low level sensor fusion approach projects image data to the 3D space, using the same model as a 3D to 2D projection would use.

Two conditions have to be met, so that the computation of the projection from 2D to 3D data space is executed. The existence of an event in 3D space is one criteria and the other is that the event is in the FOV of the camera. The algorithm computes the azimuth angle of the start and end position of the bounding box. When looking at the 3D data from an orthographical perspective, the azimuth angles are used to associate camera data with the closest captured object in 3D space. Any other objects, being within the azimuth angles, are not associated.

C. Proposed Fusion Network

Statistical inference is defined as the process used to create hypotheses about underlying distributions and their parameters, by analyzing data. There are different approaches towards inference and even various different schools, such as the ones mentioned in [17]. In order to represent probabilistic knowledge, it is a common approach to use numerical representations. A graphical representation is more intuitive and human readable. A graph structure, in which the nodes represent propositional variables and edges represent dependencies, is a common model of a joint probability distribution. Depending on the choice of graph, dependencies and independencies can be visualized using undirected or directed graphs. For the use case of object detection, Bayesian and Markov networks were analyzed, as those are well understood and popular methods for uncertainty management. Factor graphs can be seen as a further development and combination of those approaches [25]. The fact that the Bayesian approach allows to capture conditional independencies is advantageous. A more detailed comparison of Bayesian and Markov networks can be found in [17] and [27]. These analysis suggest that the Bayesian network is the best choice for the desired fusion network. Exact inference is chosen as a tool for evaluation purposes, as it is a deterministic and simple inference method.

The proposed low level sensor fusion approach targets embedded hardware and the fusion network is required to run in real-time. Bayesian network allow for efficient inference implementations, where exact and approximate implementations are available.

The implemented Bayesian network uses a tree structure and leaf nodes chosen based on expert knowledge. The existence probability of an object detection increases when associated data from multiple sources is available. The Bayesian network has three sensor modalities contributing to the existence probability hypothesis. Those sensor technologies are in the present use-case LiDAR, camera and RADAR. The network is designed in such a way that more sensors of different modalities can be added. The possibility to expand 2nd layer nodes by additional parameters and sensors is given. In addition, the network allows to combine low level sensor data and object level data, in case smart sensors are part of a sensor set.

In the case of LiDAR, the leaf nodes are chosen as LiDAR clusters (LC) and noise is assumed to be removed by the DBSCAN.

The RADAR nodes are chosen as single RADAR target (RT), RADAR target cluster (RC) and RADAR SNR (SNR). The SNR is used to determine the quality of the RADAR data. The SNR is normalized based on the maximum expected SNR value, depending on the used RADAR.

For the camera two leaf nodes are used. The camera SNR value depends on the magnitude of a F2FD cluster and the noise in the corresponding F2FD data frame. The HOG node depends on the size of the detection (SIZE) and the SNR of the detection.

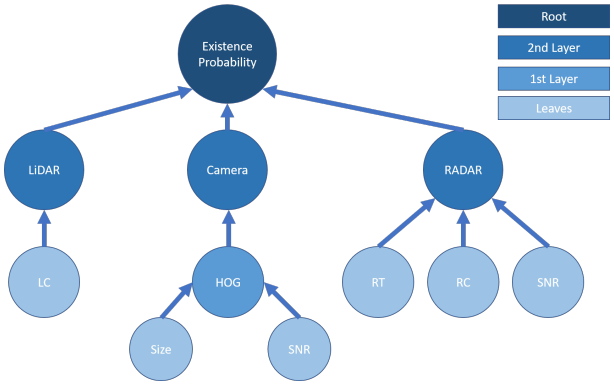


Fig. 3. Implemented Bayesian Network: The three sensor modalities contribute to the decision of the existence probability. The leaf nodes show the information provided to the network.

Computing the existence probability of the root node is implemented according to the following steps:

1) Initialize all leaf nodes
2) Compute next higher level node distributions depending on leaf nodes
3) Compute next higher level node distributions depending on previous layer nodes
4) Check if root node is reached, if not go to step 3
5) Output root node hypothesis

The proposed low level sensor fusion approach initializes the leaf nodes based on the output of the association step. It is possible that only data from one sensor modality is associated to an object, whereas it is likely that data from all sensor modalities is available for an object. Based on this input, the higher level layers of the tree structure are computed. This leads to a probability distribution for each individual sensor modality in the higher-level layer, which is used to compute the existence probability of the root node. Missing values for leaf nodes, caused by missing sensor measurements, are treated as zero-values.

IV. Experiments

A. Sensor Setup

The data recording was performed by a car equipped with LiDAR, camera and RADAR sensors in an urban environment. The mounting positions of all devices are known and the sensor setup has been calibrated in advance. The extrinsic and intrinsic calibration matrices are known. The recording vehicle uses the following setup for data capturing purposes:

- **360°Ibeo Lux LiDAR sensors:** In total, six Ibeo Lux LiDAR sensors are mounted around the car, resulting in a 360° coverage of the vehicle environment. Each LiDAR has four laser beams, a horizontal field of view (FOV) of 110° and runs at $12.5Hz$.
- **Front RADAR:** This is a short range RADAR operating at a frequency of $24GHz$ with a FOV of 100°. The sensor returns targets at a frequency of $30Hz$.

- **Front camera:** The used camera is a global shutter grayscale camera with a resolution of 752x480 at $25Hz$. The FOV is $67°$.
- **High precision LiDAR sensor:** A Velodyne HDL-64 using 64 beams is mounted, but it is not used in this present work, as it is non automotive grade. It serves as a reference to determine ground truth.

The presented algorithms were implemented based on data from all LiDAR sensors, data from the RADAR and data from the front facing camera. The proposed low level sensor fusion approach is evaluated in the area, where all three sensor modalities overlap. This is the case for the FOV of the front facing camera. The area in front of the car is the most critical and has to be reliably monitored at all time. This includes varying environmental conditions, for example, darkness, rain or fog. However, the recorded data set does not cover any situations with snow, rain or fog.

1) Association implementation: The sensors are not synchronized and run at different frequencies, which results in temporal offsets between the sensor measurements. The camera sensor runs at a frequency of $f_C = 25Hz$, the RADAR at $f_R = 30Hz$ and the LiDAR at $f_L = 12.5Hz$. The available data sets only provide data from one camera and one RADAR. The only sensor covering $360°$ of the environment is the LiDAR. Following the proposed association scheme, the temporal parameters for the association process are chosen as constants. The worst case delay that can occur is caused by the LiDAR and is $t_{delay} = 80ms$. It is reasonable to allow a more sensitive threshold or respectively smaller overlaps for associating data, as this results in the assumption, that the object is larger than it is in reality. The maximum sensor frequency resulting in $t_{fast} = 33.3ms$ is used. The worst case delay assumption is made throughout the entire association process. The parameters α and β are chosen accordingly.

The first step is the association of 3D bounding boxes. This happens under the assumption that all relevant objects have two degrees of freedom. This is feasible, because any road user is connected to the ground. By performing an orthogonal projection to the ground plane, the bounding box overlap is computed. The camera data is associated with LiDAR and RADAR bounding boxes, projected to the ground plane.

B. Recorded Scenarios

The available data set includes a variety of different use cases, ranging from city highways to busy urban street crossings. Due to limitiation given by the prototypical installation of the sensor set, no data from harsh environmental conditions is available. The algorithms have to handle data originating from various kinds of objects in the environment. In this present work, urban use cases are in the focus. For this purpose, vehicles and pedestrians are the two object classes being evaluated. Consequently, the three following scenarios were chosen for evaluating the proposed sensor fusion approach. The maximum distances listed in the tables results from the limited camera resolution which did not allow for ground truth determination at greater distances.

The busy urban crossing scenario (I) includes a large variety of different vehicles and pedestrians, as well as bicycles. This scene was chosen due to the large amount of road users and the presence of partly occluded objects. The test vehicle is approaching an intersection and stopping at a red light while vehicles are passing by at velocities of around $50km/h$ and pedestrians are crossing the street. Different viewing angles and occluded road users can be observed in this scene. The duration of the scenario is 42 seconds.

The red light scenario (II) includes only vehicles passing by, while the test vehicle is approaching a red light and waiting at the stop line. This scenario was chosen because different vehicle types are present. A very small car can be investigated in this scene, as well as a bus. The presence of a bridge pillar occluding vehicles until they enter the intersection allows for a suitable evaluation of the object detection delay. The duration of the scenario is 34 seconds.

The night scenario (III) was recorded at twilight with challenging lighting conditions. All vehicles have their lights turned on and pedestrians are present on sidewalks. The test vehicle approaching vehicles are using two lanes and occlude each other. Additionally, pedestrians are present, who are not actively lit by a light source. The duration of the scenario is 40 seconds.

TABLE I
SCENARIOS

Category	Number	Max distance (m)	Min distance (m)
(I) Urban Crossing			
Pedestrian	7	45	5
Vehicle	27	50	15
Bicycle	2	45	34
(II) Red Light			
Pedestrians	-	-	-
Vehicles	10	50	8
Bicycles	-	-	-
(III) Night			
Pedestrians	8	25	3
Vehicles	19	50	2
Bicycles	-	-	-

C. Comparison of Fusion Approaches

The proposed low level sensor fusion (LLSF) approach is compared with a state-of-the-art LiDAR-based sensor fusion system, which is referred to as fusion box (FB). Identical unprocessed sensor data is played back to both systems and the results are simultaneously recorded. For a fair comparison, the LLSF only returns object detections if LiDAR data is available. The FB is based on the fusion of six LiDAR sensors. The evaluation is based on the above described scenarios, which have independent ground truth information for each sensor modality available. The data from all sensors is manually annotated to create the ground truth data. Pedestrian and vehicle detections are analyzed separately to allow a detection performance evaluation for these classes.

The proposed LLSF system operates on single frames and does not apply any tracking. A positive object detection occurs when the output of the Bayesian network is above a threshold of 0.5. In contrast, the FB uses an object tracker as well as object trajectory prediction. This is an advantage for the FB when comparing the metrics in case of partly or fully occluded objects. Nevertheless, a tracking algorithm can be easily added to the output of the LLSF system.

The three chosen scenarios represent typical urban use cases during night and day. That includes road users at various distances, captured in different angles and difficult lighting conditions. Table II shows the result for the analyzed scenarios.

TABLE II
OVERALL NUMBER OF OBJECT DETECTIONS

	TP	FN	FP	F1-Score
Vehicles				
LLSF	5763	48	24	0.9938
FB	5584	227	32	0.9768
Pedestrians				
LLSF	865	33	5	0.9786
FB	852	46	64	0.9394

The average percentage of road users being detected by more than a single sensor modality, in this case LiDAR, is 94.38% for vehicles and 85.63% for pedestrians. These objects are detected with a higher confidence in comparison to objects detected by LiDAR only. For the present evaluation The sensor FB introduces an average detection delay of $T_{delay} = 176.5ms$ for vehicles and $T_{delay} = 268.28ms$ for pedestrians, whereas the LLSF introduces a delay of $T_{delay} = 92.65ms$ for vehicles and $T_{delay} = 105.92ms$ for pedestrians.

In figure 4 the number of objects detected by the different fusion approaches is depicted for 100 measurement frames of scenario (II).

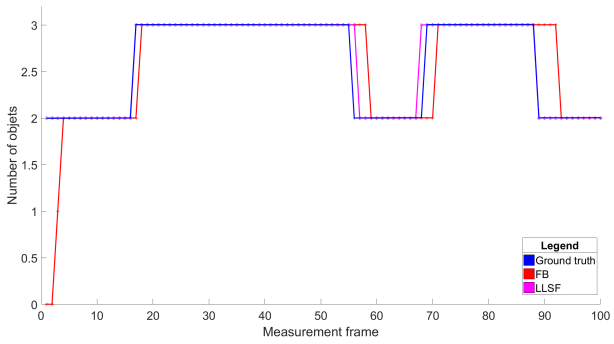


Fig. 4. Fusion Approach Comparison: The number of objects detected per frame by the two fusion approaches is compared to ground truth. In this example, the time delay introduced by the FB can be seen at the beginning.

V. CONCLUSION

The proposed low level sensor fusion uses data provided by LiDAR, RADAR and camera to improve object detection before tracking by benefiting from the fact that objects are seen by multiple sensors. A centralized fusion approach operating on low level sensor data enables the reduction of detection latency. To efficiently use this approach, different techniques like clustering are necessary to reduce the computational load of handling low level sensor data for association.

A single physical object can be captured by multiple sensor modalities. As soon as data from more than one sensor technology can be associated with the same object, the detection hypothesis is improved. The data association approach allows the use of asynchronous sensor setups. The test results show that the proposed approach significantly reduces the detection latency. The state of the art tracking-based solution performs better for partially or fully occluded objects. It is important to consider the main conceptual differences of the compared sensor fusion systems. The FB uses only one sensor modality, whereas the proposed low level sensor fusion approach uses three sensor modalities to detect objects. Additionally, the FB uses object tracking and object trajectory prediction. None of those methods is implemented in the proposed approach, as it only operates on single measurement frames.

The proposed low level sensor fusion approach detects objects with a decreased delay when entering the sensors' FOV. The quicker the detection is, the faster a system can react and avoid or mitigate a crash. In greater distances LiDAR measurements are sparser and object detection latency increases with the FB.

The test results show that the great majority of vehicle detections benefit from the low level sensor fusion approach in terms of detection latency and the availability of data from multiple sensor. Objects in greater distance provide less dense LiDAR measurements, which leads to increased detection latency for the FB or no detection. The LLSF is able to compensate this effect by using data from other sensor modalities.

The FB generates more false positives for each scenario which can partially be explained by the usage of trackers. Once an object is detected, it is not immediately dropped, even though it might not be visible anymore. The proposed low level sensor fusion does not have this feature. However, introducing this capability could improve the performance in situations with temporarily occluded objects.

Future Work

The implemented Bayesian network was used to demonstrate the working principle of the sensor fusion. However, the network is partially expert-based and little optimization work has been done. This should be improved in the future to achieve better object detection hypothesis. Additionally, possibilities to expand the network should be considered. Another aspect that should be investigated in the future is the possibility to train classifiers which directly operate on associated multi modal sensor data. The proposed object detection has the potential to eliminate the object detection step in state-of-the-art neural networks while providing a richer features space in comparison to single modality based inputs only.

REFERENCES

[1] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, The IEEE International Conference on Computer Vision (ICCV), October, 2017

[2] Girshick, Ross, Fast R-CNN, IEEE Conference on Computer Vision and Pattern Recognition, 2014

[3] Ren, Shaoqing and He, Kaiming and Girshick, Ross and Sun, Jian, Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, The IEEE International Conference on Computer Vision (ICCV), 2015

[4] He, Kaiming and Gkioxari, Georgia and Dollar, Piotr and Girshick, Ross, Mask R-CNN, The IEEE International Conference on Computer Vision (ICCV), October, 2017

[5] Wei Liu1, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu , Alexander C. Berg1, SSD: Single Shot MultiBox Detector, European Conference on Computer Vision, 2016

[6] Joseph Redmon and Ali Farhadi, Yolo9000: better, faster, stronger, arXiv preprint, 2017

[7] Association for safe international road travel Road safety facts "https: //www.asirt.org/safe-travel/road-safety-facts/, 2019. [Online; accessed 15-September-2019].

[8] Barzilay and Szolovits. Exact inference in bayes nets pseudocode. http://courses.csail.mit.edu/6.034s/handouts/spring12/ bayesnets-pseudocode.pdf, 2012. [Online; accessed 15-September-2019].

[9] Hyunggi Cho, Young-Woo Seo, BVK Vijaya Kumar, and Ragunathan Raj Rajkumar. A multi-sensor fusion system for moving object detection and tracking in urban driving environments. In *Robotics and Automation (ICRA), 2014 IEEE International Conference on*, pages 1836–1843. IEEE, 2014.

[10] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893 vol. 1, June 2005.

[11] Martin Horn Daniel Watzenig. Automated driving. In *Safer and More Efficient Future Driving*. Springer International Publishing, 2017.

[12] Martin Ester, Hans peter Kriegel, Jrg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. pages 226–231. AAAI Press, 1996.

[13] K. C. Fuerstenberg, K. C. J. Dietmayer, and V. Willhoeft. Pedestrian recognition in urban traffic using a vehicle based multilayer laserscanner. In *Intelligent Vehicle Symposium, 2002. IEEE*, volume 1, pages 31–35 vol.1, June 2002.

[14] Daniel Göhring, Miao Wang, Michael Schnürmacher, and Tinosch Ganjineh. Radar/lidar sensor fusion for car-following on highways. In *Automation, Robotics and Applications (ICARA), 2011 5th International Conference on*, pages 407–412. IEEE, 2011.

[15] David L. Hall and Sonya A. H. McMullen. Mathematical techniques in multisensor data fusion (artech house information warfare library), 2004.

[16] John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979.

[17] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988.

[18] N. Kaempchen, M. Buehler, and K. Dietmayer. Feature-level fusion for free-form object tracking using laserscanner and video. In *IEEE Proceedings. Intelligent Vehicles Symposium, 2005.*, pages 453–458, June 2005.

[19] Nico Kaempchen and Klaus Dietmayer. Data synchronization strategies for multi-sensor fusion. In *Proceedings of the IEEE Conference on Intelligent Transportation Systems*, pages 1–9, 2003.

[20] Nico Kämpchen. Feature-level fusion of laser scanner and video data for advanced driver assistance systems, 2007.

[21] C. Tarin, H. Brugger, R. Moscardo, B. Tibken and E. P. Hofer, Low level sensor fusion for autonomous mobile robot navigation, IMTC/99. Proceedings of the 16th IEEE Instrumentation and Measurement Technology Conference (Cat. No.99CH36309), Venice, 1999, pp. 1377-1382 vol.3.

[22] D. Kellner, M. Barjenbruch, K. Dietmayer, J. Klappstein, and J. Dickmann. Instantaneous lateral velocity estimation of a vehicle using doppler radar. In *Proceedings of the 16th International Conference on Information Fusion*, pages 877–884, July 2013.

[23] P. Kmiotek and Y. Ruichek. Representing and tracking of dynamics objects using oriented bounding box and extended kalman filter. In *2008 11th International IEEE Conference on Intelligent Transportation Systems*, pages 322–328, Oct 2008.

[24] S. Lange, F. Ulbrich, and D. Goehring. Online vehicle detection using deep neural networks and lidar based preselected image patches. In *2016 IEEE Intelligent Vehicles Symposium (IV)*, pages 954–959, June 2016.

[25] H. A. Loeliger, J. Dauwels, J. Hu, S. Korl, L. Ping, and F. R. Kschischang. The factor graph approach to model-based signal processing. *Proceedings of the IEEE*, 95(6):1295–1322, June 2007.

[26] M. Mahlisch, R. Schweiger, W. Ritter, and K. Dietmayer. Sensorfusion using spatio-temporal aligned video and lidar for improved vehicle detection. In *2006 IEEE Intelligent Vehicles Symposium*, pages 424–429, 2006.

[27] Ljubo Mercep. *Context-Centric Design of Automotive Human-Machine Interfaces*. PhD thesis, Technische Universität München, 2014.

[28] J. Sankaran and N. Zoran. Tda2x, a soc optimized for advanced driver assistance systems. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2204–2208, May 2014.

[29] R. Schubert, C. Adam, M. Obst, N. Mattern, V. Leonhardt, and G. Wanielik. Empirical evaluation of vehicular models for ego motion estimation. In *2011 IEEE Intelligent Vehicles Symposium (IV)*, pages 534–539, June 2011.

[30] Neil Scicluna and Christos-Savvas Bouganis. Fpga-based parallel dbscan architecture. In *International Symposium on Applied Reconfigurable Computing*, pages 1–12. Springer, 2014.

[31] Qi Yue Shaobo Shi and Qin Wang. Fpga based accelerator for parallel dbscan algorithm. *Computer Modelling & New Technologies*, 18(2):135–142, 2014.

[32] Z. Taylor and J. Nieto. Motion-based calibration of multimodal sensor extrinsics and timing offset estimation. *IEEE Transactions on Robotics*, 32(5):1215–1229, Oct 2016.

[33] Grace Tsai. Histogram of oriented gradients. *University of Michigan*, 2010.

[34] B. Duraisamy, M. Gabb, A. Vijayamohnan Nair, T. Schwarz and T. Yuan, Track level fusion of extended objects from heterogeneous sensors, 2016 19th International Conference on Information Fusion (FUSION), Heidelberg, 2016, pp. 876-885.

[35] Rui Xu and D. Wunsch, II. Survey of clustering algorithms. *Trans. Neur. Netw.*, 16(3):645–678, May 2005.

[36] J. Zheng, T. Su, W. Zhu, X. He and Q. H. Liu Radar High-Speed Target Detection Based on the Scaled Inverse Fourier Transform, in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 8, no. 3, pp. 1108-1119, March 2015.

[37] S. Blank, T. Foehst and K. Berns, A fuzzy approach to low level sensor fusion with limited system knowledge, 2010 13th International Conference on Information Fusion, Edinburgh, 2010, pp. 1-7.