

# Fusion Between Laser and Stereo Vision Data For Moving Objects Tracking In Intersection Like Scenario

Qadeer Baig, Olivier Aycard, Trung Dung Vu and Thierry Fraichard

**Abstract**—Using multiple sensors in the context of environment perception for autonomous vehicles is quite common these days. Perceived data from these sensors can be fused at different levels like: before object detection, after object detection and finally after tracking the moving objects. In this paper we detail our object detection level fusion between laser and stereo vision sensors as opposed to pre-detection or track level fusion. We use the output of our laser processing to get a list of objects with position and dynamic properties for each object. Similarly we use the stereo vision output of another team which consists of a list of detected objects with position and classification properties for each object. We use Bayesian fusion technique on objects of these two lists to get a new list of fused objects. This fused list of objects is further used in tracking phase to track moving objects in an intersection like scenario. The results obtained on data sets of INTERSAFE-2 demonstrator vehicle show that this fusion has improved data association and track management steps.

## I. INTRODUCTION

Perceiving or understanding the environment surrounding a vehicle is a very important step in driving assistance systems and for the functioning of autonomous vehicles. The task involves solving both simultaneous localization and mapping (SLAM) and detection and tracking of moving objects (DATMO) problems. While SLAM provides the vehicle with a map of the environment, DATMO allows the vehicle being aware of dynamic entities around, tracking them and predicting their future behaviors. If we are able to accomplish both SLAM and DATMO reliably in real time, we can detect critical situations to warn the driver in advance and this will certainly improve driving safety and can prevent traffic accidents. We use multiple sensors on the vehicle to observe the surrounding environment to solve these problems, since a single sensor can provide only a limited view of the environment.

A great deal of work has been done to solve these problems using different sensors, especially laser scanners [2], [12], [9], [5], [14]. In most of these works occupancy grids [4] framework has been used to represent the surrounding environment. Although using high resolution laser scanners (less than  $1^\circ$ ) we are able to obtain good maps of the environment [12]. However tracking moving objects in a complex intersection like scenario using only laser scanner is quite challenging due to temporary occlusions and many different directions of movements. So there is a strong need to use multiple sensors to solve tracking problem. But using multiple sensors inherently requires to perform

fusion between them at some appropriate level to get optimal results. The fusion can take place at three different levels: i) before objects detection, also called low level fusion, for example an occupancy grid is constructed for each sensor and fusion is performed to get a fused occupancy grid, ii) at objects detection level, output of each sensor is processed to extract lists of objects (or obstacles) in the environment and information about corresponding objects in these lists are fused to get fused list of objects, and iii) at track level, output of each sensor is processed to do tracking of moving objects and then fusion is performed on detected tracks (a track is a moving object detected consistently in few previous frames), fusion at this level is usually performed to confirm tracks detected by a primary sensor.

In this work we have developed a generic architecture for object detection level fusion between different sensors and tracking the fused list of objects (Figure 1). This architecture has two levels: the first level deals with pre processing of sensors data, detecting objects with properties specific to sensors and finally performing fusion between these lists of objects. The fusion at this level involves finding corresponding objects in the lists and merging their properties, this gives a fused list of objects with individual objects having more state information than their pre fusion versions. Second level deals with tracking of the fused objects. In this work we have used this architecture to perform fusion between objects detected by laser and stereo vision in the context of a European project INTERSAFE-2<sup>1</sup> on Volkswagen demonstrator. We perform laser processing to construct local grid map and to detect moving objects in the environment, the output of stereo vision processing consists of a list of 3D objects with classification information. After fusion, objects have position, dynamic state and classification information which result in more precise tracking results. Due to this fusion we also get good tracks for occluded or transparent objects for laser sensor.

Fusion between laser and stereo vision on other two levels (pre detection and track levels) has some limitations, for example occupancy grids constructed for stereo vision for low level fusion have many false positives and many iterations need be done to refine them [6], moreover due to high depth uncertainty for stereo sensor most of the weight is given to laser occupancy grid [1]. Baltzaksi et al. [3] have used this low level fusion technique for indoor robot navigation. Similarly at track level, due to small field of view and limited range of stereo vision, there are many miss

Authors are with University of Grenoble1 & INRIA Rhône-Alpes, Grenoble, France, (e-mails: Qadeer.Baig@imag.fr, Olivier.Aycard@imag.fr, Trung-Dung.Vu@imag.fr and Thierry.Fraichard@inria.fr)

<sup>1</sup><http://www.intersafe-2.eu>

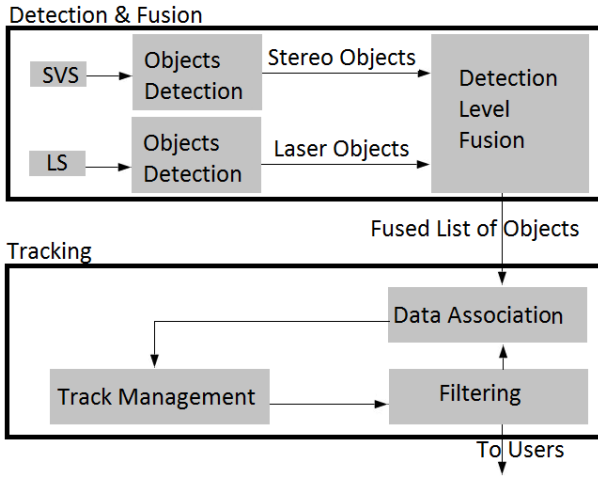


Fig. 1. Architecture of the perception system (SVS=Stereo Vision Sensor and LS=Laser Sensor).

detections because moving objects quickly go out of effective range before they are confirmed as tracks.

The rest of the paper is organized as follows. In next Section we present the demonstrator used for this work and sensors installed on it. We summarize our work [2] on laser processing to build a map of the environment, localize our ego vehicle inside this map and detect moving objects in Section III. In Section IV we introduce the stereo-vision processing done by University of Cluj. In Section V we detail our work on fusion, with tracking in Section VI. Experimental results are reported in Section VII. We conclude this work in Section VIII.

## II. EXPERIMENTAL SETUP

The Volkswagen demonstrator vehicle used to get datasets for this work has multiple sensors installed on it. It has a long range laser scanner (Lidar) with a field of view of  $160^\circ$  and a maximum range of 150 m. It has 161 laser beams called channels and resolution of  $1^\circ$ . Each channel can detect a maximum of two targets with vertical divergence of  $3^\circ$ . A channel detects two targets because each laser beam scans two planes: horizontal and another with a vertical divergence of  $3^\circ$ , if the object detected in the second plane is not the same as that of first plane then it is reported as second target. In this work we use nearest target only. Other sensors installed on this demonstrator include a stereo vision camera, four short range radars (SRR) one at each corner of the vehicle and a long range radar (LRR) in front of the vehicle (Figure 2). Our work in this paper is only concerned with the processing and fusion of Lidar and stereo vision data. For both of these sensors we have following two reference frames:

- Stereo (with origin on the ground plane exactly below the laser scanner, z-axis pointing in the direction of driving and x-axis towards right).
- Laser (with origin fixed on the laser scanner, y-axis pointing in the direction of driving and x-towards right).

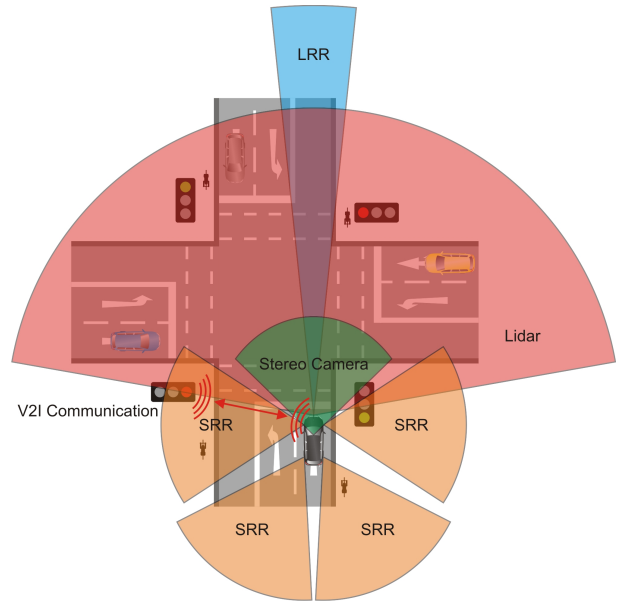


Fig. 2. Sensors installed on the demonstrator vehicle

These frames of reference are fixed w.r.t each other and their transformation matrices are known.

## III. LASER PROCESSING

In this section we summarize our laser data processing [2] used to detect moving objects, with details of improvements done for this fusion work. This process consists of following steps: first we construct a local grid map and localize the vehicle in it, then using this map we classify individual laser beams in the latest data frame as belonging to moving or static parts of the environment, finally we segment this laser scan to make objects from individual laser hit points.

### A. Environment Mapping & Localization

We have used incremental mapping approach based on laser scan matching algorithm to build a local vehicle map. Based on occupancy grid representation the environment is divided into two dimensional lattice of rectangular cells and we keep track of probabilistic occupancy state of each cell. We build a grid map of  $90m \times 108m$  with each cell having dimensions of  $0.3m \times 0.3m$ . Environment mapping is essentially the estimate of posterior probability of occupancy  $P(m | x_{1:t}, z_{1:t})$  for each cell of grid  $m$ , given observations  $z_{1:t} = \{z_1, \dots, z_t\}$  from time 1 to time  $t$  at corresponding known poses  $x_{1:t} = \{x_1, \dots, x_t\}$ , here  $z_t = \{P_i\}$  where  $P_i = (x = r_i \cos \theta_i, y = r_i \sin \theta_i)$  is the impact point of  $i$ th laser beam (from its polar coordinates  $r_i, \theta_i$ ) for  $\forall i \leq 161$  expressed in laser frame of reference, and  $x_t = (x, y, \theta)$  is the vehicle pose. To solve localization problem we have used importance sampling based particle filter [11]. A total of 300 particles are used. For the given previous pose  $x_{t-1}$  and current odometry information  $u_t = (\nu, \omega)$  (translational and rotational velocities) we sample different possible positions of the vehicle from the motion model  $P(x_t | u_t, x_{t-1})$ . Then we compute the probability of each position using laser data

and a sensor model. The pose of the particle getting highest probability is taken as true pose.

### B. Moving & Static Parts Distinction

By projecting the latest laser data onto the local map constructed so far, the impact points are classified as static or dynamic. The points observed in free space are classified as dynamic whereas the rest are classified as static. More precisely, if we represent the local grid map at time  $t$  as  $M^t = \{m\}$  where  $m$  is a grid cell and if  $M^t[P_i]$  gives the occupancy probability of the grid cell corresponding to the laser impact point  $P_i$  then we classify the laser impact points into following two types:

- $MovingPoints = \{P_i \mid M^t[P_i] < 0.5\}$
- $StaticPoints = \{P_i \mid M^t[P_i] \geq 0.5\}$

### C. Laser Objects Extraction

Finally we perform segmentation to extract objects from these laser impact points. We define an object as:

$$\acute{O}_L = \{P_n \mid dist(P_n, P_{n-1}) < S_{thr}\}$$

Here  $P_i$  is the impact point as defined above,  $dist(P_n, P_{n-1})$  is the euclidean distance between two adjacent points, and  $S_{thr}$  is the segment threshold distance which is equal to 2.0 meters in our experiments. An object is marked as dynamic if  $\acute{O}_L \cap MovingPoints \neq \phi$ , Finally we calculate the polar coordinates of center of gravity (centroid)  $(r_L, \theta_L)$  of each object using Cartesian coordinates of its constituting points as:  $r_L = \sqrt{\hat{x}^2 + \hat{y}^2}$  and  $\theta_L = atan2(\hat{y}, \hat{x})$  where  $\hat{x} = \sum_i x/n$  and  $\hat{y} = \sum_i y/n$  for  $\forall P_i(x, y) \in \acute{O}_L$  and  $n = |\acute{O}_L|$ .

### D. Laser Processing Output

The output of laser processing step at time  $t$  consists of a local grid map  $M^t$ , and a list of detected moving objects  $L_{objects}^t = \{O_L\}$  where  $O_L = (\acute{O}_L, r_L, \theta_L)$  is moving object with centroid information. Grid map is only used to display on the screen whereas list of dynamic objects is used further for fusion. The results of laser processing are shown in Figure 3.

## IV. STEREO VISION PROCESSING

Another team from University of Cluj-Napoca Romania working on INTERSAFE-2 project has performed the stereo vision processing [7], [8]. Their output of stereo vision processing consists of a list of objects detected in each frame of the stereo images. For each object in the list we are given 3D coordinates of the four corners of the lower rectangle of the object cuboid in stereo frame of reference, the height of the top rectangle and the class of the object (pedestrian, pole vehicle etc), but no information about the dynamic state of the object are available, objects can be dynamic or static. From this data we can easily calculate the coordinates of the eight corners of the cuboid.

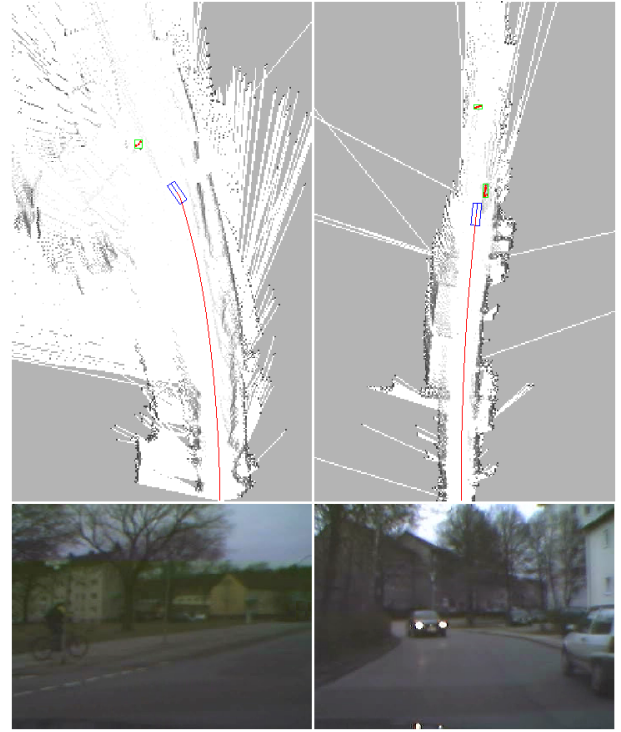


Fig. 3. Mapping and moving objects detection results. Detected moving objects (a bicycle in left image and two cars in right image) are shown as green rectangles.

### A. Pre-Fusion Processing

The first step before performing fusion between laser data and stereo vision objects is to project stereo objects onto the laser plane using transformation matrices to achieve the common spatial reference. For object level fusion between laser and stereo we need to represent vision objects by their centroids. We take this centroid as the middle point of the front line segment of object rectangle. Since laser points also belong to the front end of objects, so this centroid gives better results than the object rectangle center. We calculate polar coordinates of the centroid for each vision object (in a similar way as explained for laser objects) to make the representation compatible to the laser objects for fusion.

### B. Stereo Vision Processing Output

The output of stereo vision processing at time  $t$  consists of a list of objects  $V_{objects}^t = \{O_V\}$  where  $O_V = (r_V, \theta_V, class)$ .  $r_V$  and  $\theta_V$  are the polar coordinates of the object centroid.

## V. LASER AND STEREO DATA FUSION

In this section we give details of object detection level fusion between laser and stereo vision sensors. Two lists of objects are input to the fusion process: list of dynamic objects detected by laser and represented as centroid points, and list of objects (static or dynamic) detected by stereo vision represented as points along-with classification information. We believe that an object detection level fusion between these two lists can complement each other thus giving more

complete information about the states of objects in the environment. This fusion process consists of following two steps:

#### A. Object Association

In this step we determine which stereo objects are to be associated to which laser objects from the two object lists, using nearest neighbor technique. The positional uncertainty of an object given by stereo vision increases with depth, so we have defined a distance threshold function based on the depth of the stereo object from the origin as:

$$V_{thr} = 5 * \frac{r_V}{20}$$

$V_{thr}$  is the uncertainty in position of an object detected at a distance of  $r_V$  by stereo vision. Here 5 (meters) is the maximum depth uncertainty for an object detected at a distance of 20m for this work. Stereo objects beyond this distance are ignored because the effective range of stereo is limited to 20m for this work. A stereo object  $O_V^i$  is associated to a laser object  $O_L^j$  if  $dist(O_V^i, O_L^j) < V_{thr}$  and  $O_L^j$  is closest to  $O_V^i$ .

#### B. Position information fusion

This step works on the pair of objects associated with each other in the previous step and fuses their position (range and bearing) information. We model the position uncertainty using 2D Gaussian distribution for both objects. Suppose  $P_L = [r_L, \theta_L]^T$  is the centroid position of laser object and  $P_V = [r_V, \theta_V]^T$  is the centroid position of associated stereo vision object. If  $X$  is the true position of the object then the probability that laser detects this object at point  $P_L$  is given as:

$$P(P_L|X) = \frac{e^{-\frac{(P_L-X)^T R_L^{-1} (P_L-X)}{2}}}{2\pi \sqrt{|R_L|}}$$

and similar probability for stereo object is given as:

$$P(P_V|X) = \frac{e^{-\frac{(P_V-X)^T R_V^{-1} (P_V-X)}{2}}}{2\pi \sqrt{|R_V|}}$$

Here  $R_L$  is the 2X2 covariance matrix of range and bearing uncertainty calculated from the uncertainty values provided by the vendor. Whereas  $R_V$  is the covariance matrix for stereo vision and depends on the depth of the object from origin. To calculate this matrix empirically for different ranges (with a difference of 2 meters) we manually associate vision objects with corresponding laser objects. Considering the laser object position as the mean true position of the object we are able to calculate their difference for both range and bearing values for vision objects. Repeating this process for different ranges in different data sets we are able to collect data for calculating this matrix.

Using Bayesian fusion the probability of fused position  $P = [r_F, \theta_F]^T$  is given as:

$$P(P|X) = \frac{e^{-\frac{(P-X)^T R^{-1} (P-X)}{2}}}{2\pi \sqrt{|R|}}$$

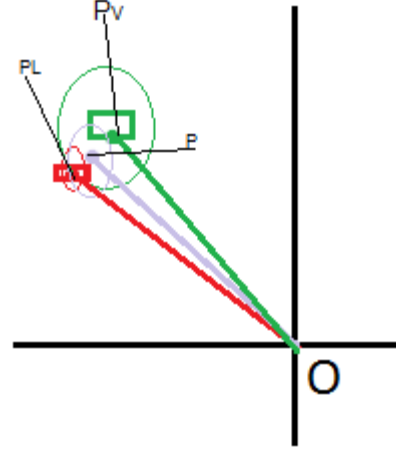


Fig. 4. Fusion process: red color shows the position uncertainty of laser object, green color for corresponding stereo object and violet the fusion result of the two.

where  $P$  and  $R$  are given as:

$$P = \frac{P_L/R_L + P_V/R_V}{1/R_L + 1/R_V}$$

and

$$1/R = 1/R_L + 1/R_V$$

respectively. This process of fusion is shown in figure 4.  $P$  is taken as the position of the fused object.

The result of this fusion process is a new list of fused objects. This list also has all the laser objects which could not be associated with stereo objects and all the stereo objects which could not be associated with some laser objects. We keep unassociated stereo objects because they may correspond to dynamic objects which may not have been detected by laser in current frame due to occlusion as explained next.

Figure 5 shows an interesting fusion scenario, objects shown in cyan color are the objects detected by stereo vision whereas the objects shown by light violet rectangles are the laser detected objects, red dots are raw laser impact points. An oncoming car that was being detected by laser until last scan is now occluded by a cyclist whereas stereo system was able to detect this car. So this miss detection by laser will be filled in by stereo during fusion hence giving a smooth track. The increased position uncertainty with depth for stereo vision objects can also be seen in the figure (green ellipses).

#### C. Fusion Output

The output of fusion process consists of fused list of objects  $F_{objects}^t = \{O_F\}$  where  $O_F = (r_F, \theta_F, class, DynamicState, SensorCount)$ . For each object we have position (centroid) information, dynamic state information (dynamic or unknown, unknown for unassociated stereo objects), classification information and a count for number of sensors detecting this object.

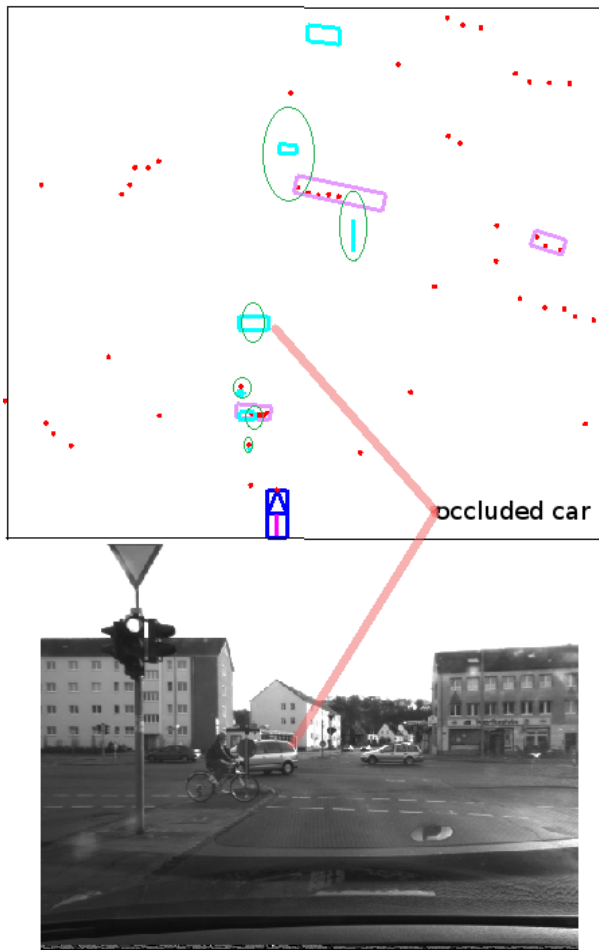


Fig. 5. Laser and Stereo vision objects fusion, see the text for details.

## VI. TRACKING

In general multi objects tracking problem is quite complex. It includes the definition of tracking methods, association methods and maintenance of objects currently present in the environment. Usually Bayesian filters are used to solve this problem which require the definition of a specific motion model of tracked objects to predict their positions in the environment. Using the prediction and observation update combination, new position of each object is estimated. In the following we explain the components of our tracking module.

### A. Data Association

This step consists of assigning new objects of fused list to the existing tracks. Since in the current work we are more concerned with tracking multiple objects in an intersection like scenario so it is important to choose a more effective technique of data association. We have used MHT [10] approach to solve the data association problem. An important optimization that we have achieved here due to fusion process mentioned above is related to classification information provided by stereo vision. While generating hypotheses we ignore all those hypotheses which involve

objects from different classes. For example a hypothesis trying to involve a pedestrian with a vehicle in a track will be ignored, this significantly reduces the number of hypotheses. To further control the growth of tracks trees we need to use some pruning technique. We have chosen the N-Scans pruning technique to keep the tracks trees to a limit of  $N$ .

### B. Track Management

In this step tracks are confirmed, deleted or created using the  $m$ -best hypotheses resulting from the data association step. New tracks are created if a new track creation hypothesis appears in the  $m$ -best hypothesis. A newly created track is confirmed if it is updated by objects detected in current frames after a variable number of algorithm steps (one step if the object was detected by both laser and stereo vision otherwise in three steps). This implies that the spurious measurements which can be detected as objects in the first step of our method are never confirmed. To deal with non-detection cases, if such a hypothesis appears (which can appear for instance when an object is occluded by another one) tracks with no new associations are updated according to their last position, for them next filtering stage becomes a simple prediction. In this way a track is deleted if it is not updated by a detected object for a given number of steps.

### C. Filtering

Since in an intersection like scenario there may be different types of objects (vehicles, motor bikes, pedestrians etc) moving in different directions using different motion modes, a single motion model based filtering technique is not sufficient. To address the tracking problem in this scenario we have used an on-line adapting version of Interacting Multiple Models (IMM) filtering technique. The details of this technique can be found in our other published work [13]. We have seen that four motion models (constant velocity, constant acceleration, left turn and right turn) are sufficient to successfully track objects on an intersection. We use four Kalman filters to handle these motion models. Finally the most probable trajectories are computed by taking the most probable branch and we select one unique hypothesis for one track tree.

## VII. RESULTS

Fusion and tracking results are shown in figures 6 and 7 along-with the images of corresponding scenarios. Figure 6 shows an interesting intersection scenario for fusion. Here left image shows tracking results based only on laser data, car behind the cyclist was occluded in last few frames giving insufficient impact points to be detected as a moving object. Right image shows tracking with fusion, car was also partially detected by stereo vision and was successfully tracked in the fused results. Figure 7 shows a similar scenario, the ego vehicle is waiting for the signal, a truck turning left, a cyclist and a pedestrian crossing the road in opposite directions are being tracked. Although truck in this scenario is partially occluded by the cyclist but due to fusion it has been tracked successfully. Table I shows empirically observed statistics of missed tracks for three data sets.

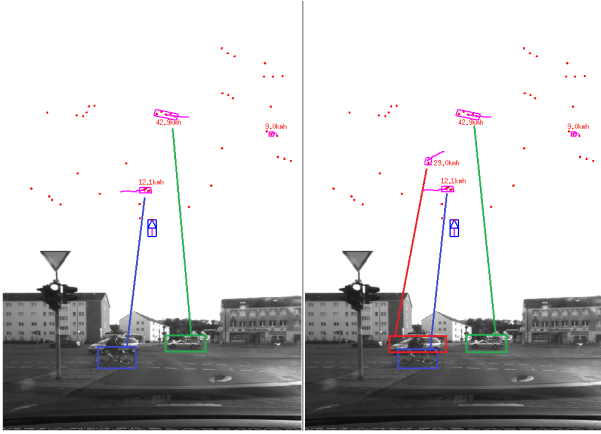


Fig. 6. Fusion and tracking results. Left(laser only): car occluded by cyclist is not being tracked. Right(laser and stereo): with fusion car was successfully tracked.



Fig. 7. Tracking results for a truck, a pedestrian and a cyclist.

TABLE I  
STATISTICS OF MISSED TRACKS WITH AND WITHOUT FUSION FOR THREE DATA SETS.

| Data Set | Missed tracks without fusion | Missed tracks with fusion |
|----------|------------------------------|---------------------------|
| 1        | 10                           | 7                         |
| 2        | 12                           | 8                         |
| 3        | 7                            | 4                         |

## VIII. CONCLUSION

In this paper we have presented our approach for fusion between laser scanner and stereo-vision at the object detection level as opposed to pre-detection or track level. We have demonstrated how the fused objects have more state information than their pre-fusion versions. This fusion has improved the data association and track management steps in the tracking phase. Experimental results on a Volkswagen demonstrator vehicle in the scope of the European project INTERSAFE-2 show the effectiveness of the work presented in this paper.

## IX. ACKNOWLEDGEMENTS

This work was conducted within the research project INTERSAFE-2. INTERSAFE-2 is part of the 7th Framework Programme, funded by the European Commission. The partners of INTERSAFE-2 thank the European Commission for supporting the work of this project.

## REFERENCES

- [1] Q. Baig and O. Aycard. Low level data fusion of laser and monocular color camera using occupancy grid framework. In *International Conference on Control, Automation, Robotics and Vision (ICARCV)*, Singapore, December 2010.
- [2] Q. Baig, TD. Vu, and O. Aycard. Online localization and mapping with moving objects detection in dynamic outdoor environments. In *IEEE Intelligent Computer Communication and Processing (ICCP)*, Cluj-Napoca, Romania, August 2009.
- [3] H. Baltzakis, A. Argyros, and P. Trahanias. Fusion of laser and visual data for robot motion planning and collision avoidance. *Mach. Vision Appl.*, 15(2):92–100, 2003.
- [4] A. Elfes. *Occupancy grids: a probabilistic framework for robot perception and navigation*. PhD thesis, Carnegie Mellon University, 1989.
- [5] D. Hähnel, R. Triebel, W. Burgard, and S. Thrun. Map building with mobile robots in dynamic environments. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2003.
- [6] D. Murray and J. Little. Using real-time stereo vision for mobile robot navigation. In *Autonomous Robotics*, 2000.
- [7] S. Nedeveschi, R. Danescu, T. Marita, F. Oniga, and S. Bota. Stereovision-based sensor for intersection assistance. In *Advanced Microsystems for Automotive Applications 2009*, pages 129–163. Springer Berlin Heidelberg, 2009.
- [8] S. Nedeveschi, T. Marita, R. Danescu, F. Oniga, and S. Bota. On-board stereo sensor for intersection driving assistance. architecture and specification. In *IEEE Intelligent Computer Communication and Processing (ICCP)*, pages 409–416, Cluj-Napoca, Romania, August 2009.
- [9] E. Prassler, J. Scholz, and P. Fiorini. Navigating a robotic wheelchair in a railway station during rush hour. *Int. Journal on Robotics Research*, 18(7):760–772, 1999.
- [10] D. B. Reid. A multiple hypothesis filter for tracking multiple targets in a cluttered environment. Technical Report D-560254, Lockheed Missiles and Space Company Report, 1977.
- [11] S. Thrun, W. Burgard, and D. Fox. *Probabilistic Robotics (Intelligent Robotics and Autonomous Agents)*. The MIT Press, September 2005.
- [12] TD. Vu, O. Aycard, and N. Appenrodt. Online localization and mapping with moving objects tracking in dynamic outdoor environments. In *IEEE International Conference on Intelligent Vehicles*, 2007.
- [13] TD. Vu, J. Burlet, and O. Aycard. Grid-based localization and local mapping with moving objects detection and tracking. *International Journal on Information Fusion*, Elsevier, 2009. To appear.
- [14] C.-C. Wang. *Simultaneous Localization, Mapping and Moving Object Tracking*. PhD thesis, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, April 2004.