Perception System based on Cooperative Fusion of Lidar and Cameras

Martin Dimitrievski[1], David Van Hamme[1], Wilfried Philips[1]
[1]Ghent University, IPI-imec,
St-Pietersnieuwstraat 41, B-9000 Gent, Belgium

***Abstract*—This paper proposes a novel sensor fusion method capable of detection and tracking of road users under nominal as well as in border cases of system operation. The proposed method is based on a sensor-agnostic Bayesian late fusion framework, augmented with an optional exchange of detector activation information between sensors, referred to as cooperative feedback. Experimental evaluation confirms that we obtain competitive detection and tracking performance in normal operation. The main benefit of the proposed method is in cases of sensor failure where, due to the probabilistic modeling, we observed significant improvements of both detection and tracking accuracy over the state of the art.**

***Index Terms*—cooperative fusion; camera; lidar; tracking;**

I. Introduction

With the new wave of autonomous vehicles expected to come on the market, perception algorithms are an important research topic in companies and academia. Higher levels of driving autonomy [1] require both high accuracy and system redundancy in cases of failures. To satisfy these requirements, prototype autonomous vehicles are equipped with multiple sensors. Information from different sensors must be combined to make driving decisions, referred to as sensor fusion. In literature, many approaches to sensor fusion have been proposed, validated on specialized benchmarks such as KITTI[2], nuScenes [3], Waymo [4], etc.

Sensor fusion approaches can be broadly divided into two categories: late fusion and early fusion. In late fusion, each sensor's data is processed independently into high level semantics (e.g., object candidates) which are then combined across sensors. In an early fusion system, data from multiple sources is aggregated prior to analysis which usually leads to higher precision than late fusion. However, early fusion has many practical disadvantages: higher bandwidth, models computationally intensive models, and system flexibility is reduced. In addition to these practical considerations, early fusion is more prone to the effect of domain shift. Therefore, late fusion remains very relevant to real-world deployment.

In this paper we propose a road user detection and tracking algorithm based on cooperative fusion between a lidar and multiple camera sensors, designed to be easily tuned to specific sensor configurations and environment conditions (figure 1). At the core is a sensor-agnostic Bayesian late fusion framework that can be used with any combination pre-trained detector, augmented with an optional exchange of detector activations between sensors, referred to as *cooperative feedback*. The main contribution of this paper is the theoretical model for cooperative fusion of detection confidence. A key benefit of the proposed method is the robustness to sensor failures. Moreover, we demonstrate that a probabilistic tracker built on this framework outperforms other state-of-the-art trackers in an on-line setting thanks to our rigorous treatment of object confidence.

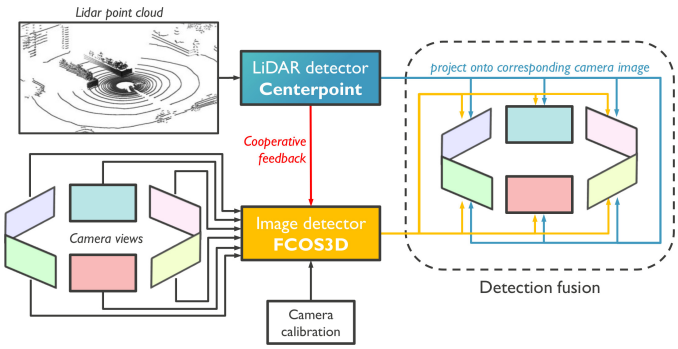

Figure 1. System diagram of the proposed cooperative fusion object detector.

The rest of the paper is organized as follows. In section §II we present the theoretical foundation of the fusion algorithm and give examples on how to estimate the models in practice. In section §III we demonstrate the effectiveness of the method through experimental evaluation. Finally, in section §IV we formulate the key conclusions and outline potential avenues to further develop and validate the method.

II. Proposed method

The goal of the sensor fusion is to combine detections from heterogeneous sensors into a scene description containing the estimated locations of road users. A detection $\mathbf{z}_l^{(k)}$ is a tuple containing the location $\mathbf{u}_l^{(k)}$ and size $\mathbf{s}_l^{(k)}$ of the detection in a sensor specific coordinate system as well as a reliability score, or an activation $a_l^{(k)}(\mathbf{u}_l^{(k)}, \mathbf{s}_l^{(k)})$. A road user $(\mathbf{r}, \mathbf{g})$ is a tuple of the road user's location $\mathbf{r}$ in world coordinates, and a feature vector $\mathbf{g}$. We model the scene in terms of occupancy of a 2D surface defined as binary-valued function $\mathrm{o}(\mathbf{r})$, where value 1 means occupied and 0 non-occupied [5]. In the following analysis we will assess the presence of road users by calculating the a posteriori probability of presence from their prior estimated probabilities and the likelihood of these supposed locations given the detections.

Because of the limited spatial resolution of sensors, we can only assess presence and absence hypotheses for a certain region centered around $\mathbf{r}$, e.g. by defining it as $H(\mathbf{r}, \mathbf{g}) \triangleq$

Final version: 10.1109/SENSORS52175.2022.9967331

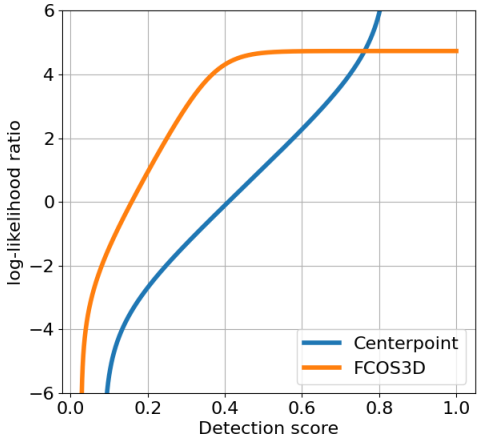

Figure 2. Log-likelihood ratios for the Centerpoint point cloud object detector (blue) and the FCOS3D camera object detector (orange) conditioned on detection activations for the nuScenes class "pedestrian".

$\max_{\mathbf{r}' \in \Omega(\mathbf{r},\mathbf{g})} \mathrm{o}(\mathbf{r}')$, which equals 1 if at least one road user is present in the state-space region $\Omega(\mathbf{r},\mathbf{g})$ centered around $\mathbf{r}$, and 0 otherwise. We denote the situation $H(\mathbf{r},\mathbf{g}) = 1$ as H_1 and $H(\mathbf{r},\mathbf{g}) = 0$ as H_0. We propose a practical method to estimate two likelihood functions $p_{A,U,S|H,R,G}(a_k, \mathbf{u}_k, \mathbf{s}_k | H_1, \mathbf{r}, \mathbf{g})$ for presence and $p_{A,U,S|H,R}(a_k, \mathbf{u}_k, \mathbf{s}_k | H_0, \mathbf{r}, \mathbf{g})$ for absence. To simplify the estimation, we can omit the dependency on the road user feature vector $\mathbf{g}$, aggregating instead the likelihood for any road user in the region regardless of shape, size or appearance. The shape of these functions can practically be learned by summarizing the activation in $\Omega(\mathbf{r})$ into a histogram $h(\alpha_k; \mathbf{u})$ of the a_k values which can serve as an approximation of $p_{A,U|H,R}(a_k, \mathbf{u}_k, \mathbf{s}_k | H_1, \mathbf{r})$. For absence, we use a training set of locations void of road users. In figure 2 we show such models of likelihood ratios for two object detectors [6] and [7] trained from data in the nuScenes [3] dataset.

Formally, the sensor fusion evaluates for any $\mathbf{r}$ the joint a posteriori log likelihood ratio for all sensors written as:

$$\ln \frac{p_{H|a^{(0)},\ldots,a^{(K-1)}}\left(H_1 | a^{(0)}(\mathbf{r}),\ldots,a^{(K-1)}(\mathbf{r})\right)}{p_{H|a^{(0)},\ldots,a^{(K-1)}}\left(H_0 | a^{(0)}(\mathbf{r}),\ldots,a^{(K-1)}(\mathbf{r})\right)}, \tag{1}$$

where superscripts 0 to $K-1$ indicate sensor index. If this log ratio is positive we should conclude road user presence, else absence. If the detectors are conditionally independent given $\mathbf{r}$, we can use Bayes' rule to calculate this joint log-ratio from the individual sensor likelihood ratios $\mathrm{llr}^{(k)}(a(\mathbf{r}))$ as

$$\ln \frac{p_{H(H_1;\mathbf{r})}}{p_{H(H_0;\mathbf{r})}} + \sum_{k=0}^{K-1} \mathrm{llr}(a^{(k)}(\mathbf{r})). \tag{2}$$

The first term is the log prior ratio which can be computed from the prevalence of pedestrians in the dataset. During tracking, the posterior ratio of the previous time step can be used as an estimate of the prior for the current time step. We can write this in a recursive form: $\mathrm{llr_t}(\mathbf{r}) \leftarrow \mathrm{llr_{t-1}}(\mathbf{r}) + \sum_{k=0}^{K-1} \mathrm{llr}(a_{t-1}^{(k)}(\mathbf{r}))$.

Computing the posterior in equation (2) for every location $\mathbf{r}$ is intractable. Instead we first determine a set of likely candidate locations, and only compute the posterior ratio for those locations. This is done in regions near individual or matched sensor detections. Matching between detections becomes difficult when each sensor outputs many detections. On the other hand, setting a high detection threshold will suppress weak evidence. We propose to address this shortcoming by allowing individual sensors to communicate with each other in a sensor-agnostic format, a concept we will call *cooperative feedback*.

After applying the object detector, each sensor shares their activation map $a^{(k)}(\mathbf{u},\mathbf{s})$ to the other sensors. This information allows the receiving sensor to locally adjust its processing according to what the source sensor perceives in that region. Concretely, we preemptively boost the log-likelihood ratio of each weak camera detection in vicinity of a strong lidar detection, with the amount of boosting determined by the amount of overlap between the activations, expressed by the Jaccard index: $\mathrm{llr}(a^{(cam)}(\mathbf{u},\mathbf{s})) \leftarrow \left(\mathrm{llr}(a^{(lidar)}(\mathbf{u}_i,\mathbf{s}_i)) - \mathrm{llr}(a^{(cam)}(\mathbf{u}_j,\mathbf{s}_j))\right) J(\mathbf{u}_i,\mathbf{s}_i,\mathbf{u}_j,\mathbf{s}_j)$. The effect of this boosting is that object candidates that would otherwise be sub-threshold will be selectively passed on to the Bayesian hypothesis evaluation.

The tracking of the spatial coordinates is performed similarly by using the Bayesian framework. For each individual hypothesis we spawn a particle filter which models the uncertainty of measured positions $\mathbf{u}$ and shapes $\mathbf{s}$ of the true road user's position and shape $\mathbf{r}$, $\mathbf{g}$ respectively. Ignoring the variation of activation scores within a local neighborhood $\Omega(\mathbf{r},\mathbf{g})$, the posterior distribution of $\mathbf{r}$ is given by the following recursion:

$$p_{R_t,H|U_t^{(0)},\ldots,U_t^{(K-1)}}\left(\mathbf{r}_t, H_1 | \mathbf{u}_t^{(0)},\ldots,\mathbf{u}_t^{(K-1)}\right) = p_{R_t,H}(\mathbf{r}_t, H_1) \prod_{k=0}^{K-1} p_{U_t^{(0)},\ldots,U_t^{(K-1)}|H,R_t}\left(\mathbf{u}_t^{(0)},\ldots,\mathbf{u}_t^{(K-1)} | H_1, \mathbf{r}_t\right), \tag{3}$$

where the first term is the prior and the second term is a product of the K sensor models. The sensor models are generally considered to be known, either given by the sensor manufacturer. The prior, however, is estimated by applying a motion model $p_{R_{t-1},H|R_t}(\mathbf{r}_{t-1}, H_1 | \mathbf{r}_t)$ to the posterior from the previous time step. Finally, the particle filter formulation defines the posterior as a set of particles: $p_{R_t,H|U}(\mathbf{r}, H_1 | \mathbf{u}) \approx \sum_{i=1}^{N_{pts}} w^{(i)} \delta_{\mathbf{r}^{(i)}}\left(d\mathbf{r}^{(i)}\right)$, where $w^{(i)}$ are particle weights that sum up to 1 and $\delta_{\mathbf{r}^{(i)}}\left(d\mathbf{r}^{(i)}\right)$ are particle positions, i.e. delta-Dirac mass located in $\mathbf{r}^{(i)}$. For more details on the tracker implementation we refer the reader to our previous work in [8], [9].

III. Experimental evaluation

We base our fusion system on the hardware of the nuScenes [3] vehicle: a sensor array of 7 heterogeneous sensors, see figure 1. The sensor array consists of one 3D lidar and 6 color cameras oriented in a radial pattern. The captured dataset is annotated for the presence of 8 classes of road users (car, truck, bus, trailer, construction vehicle, pedestrian, motorcycle and bicycle) as well as 2 classes of road infrastructure (traffic cone and barrier). The dataset contains 750 training sequences, 150 validation sequences and 150 testing sequences.

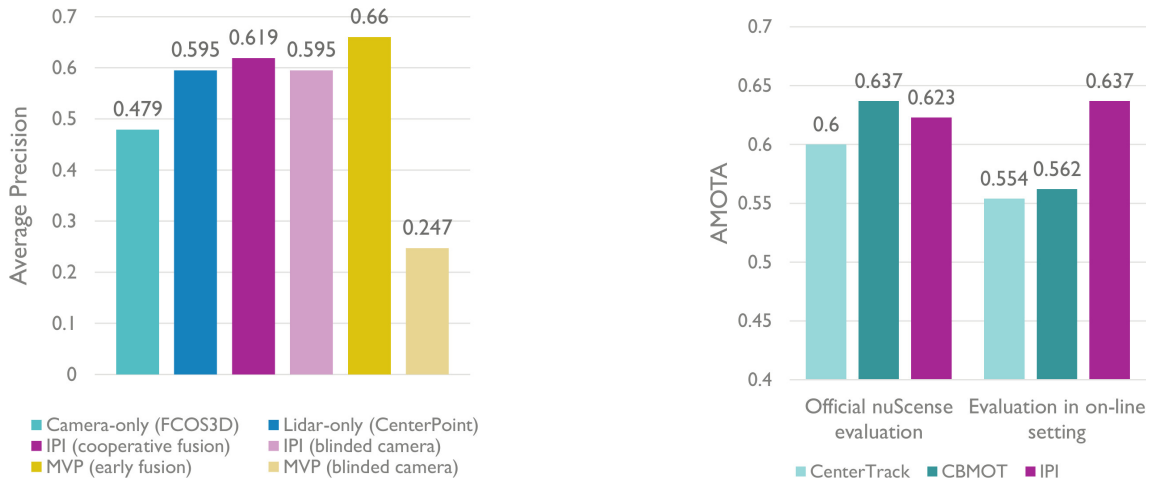

Figure 3. Results obtained by evaluating on the nuScenes validation dataset. Left: Average Precision for object detection; right: AMOTA for object tracking.

We perform object detection using the state-of-the-art lidar detector Centerpoint [6], and object detection in each of the 6 camera views using the state-of-the-art camera detector FCOS3D [7]. We use the provided sensor calibration matrices to project and match Centerpoint detections with camera detections by FCOS3D. In order to evaluate the detection accuracy under nominal circumstances, we use the default nuScenes detection and compute the detection mean Average Precision over the 150 validation sequences. On the left plot in figure 3 we show the cooperative fusion mAP compared to the individual lidar and camera detectors, as well as to a state-of-the-art early fusion method MVP [10]. The proposed fusion method outperforms both individual detectors which shows that probabilistic models are effective. However, as expected, it is outperformed by the early fusion method.

In a secondary experiment, simulating compromised camera operation, we deliberately disable the camera feed to both fusion methods simulating a hard camera failure. The proposed cooperative fusion method, in this case, shows the same performance to the lidar-only detector (mAP=0.595), but the precision of the early fusion method degraded far below the baseline (mAP=0.247). This experiment shows the fragile nature of early fusion when faced with out-of-domain input.

For evaluating the tracking performance, we use the nuScenes tracking dataset which contains labels for 7 classes of road users (bicycle, bus, car, motorcycle, pedestrian, trailer and truck). Tracking accuracy is measrured through the Average Multi-Object Tracking Accuracy (AMOTA) metric which averages over the MOTA [11] metric at different recall thresholds [3]. Using the same set of fused detection inputs, the proposed tracking algorithm was evaluated against CenterTrack [6] and CBMOT[12]. On the nuScenes validation sequences our tracking method, figure 3 right, achieves competitive results outperforming CenterTrack by 3.8% and falls behind CBMOT by 2.2%. However, as other authors [13] have discovered, the evaluation protocol of nuScenes performs post-processing of the submitted tracks, averaging track scores and filling-in missing track instances by looking at their past and future locations which puts on-line trackers at an unfair disadvantage [1]. Under this context, the evaluation removes the fine-grained track score information which any real-time tracker should compute at each time instance. We feel concerned that the computed AMOTA values in this way do not represent the true tracking performance in a real-world on-line application.

In order to measure the tracking performance in an on-line setting, we disabled the track score averaging and track instance interpolation blocks in the official evaluation script. We obtained the tracking results shown in the second block of bars in figure 3. In this on-line setting, our tracker significantly outperforms the state-of-the-art CBMOT (by 13.3%) and CenterTrack (by 14.9%). These results show the effectiveness of the detailed probabilistic modeling which our method employs at both detection as well as tracking. The tracking results were also evaluated on the test set and submitted to the nuScenes website where we reached an AMOTA score of 0.642.

IV. Conclusion and Future Work

In this paper we propose a probabilistic cooperative sensor fusion method for robust detection and tracking of road users. Our method significantly outperformed the state-of-the-art early-fusion MVP detector in simulated border cases of sensor failures. Compared to two state-of-the-art trackers, the proposed tracker shows competitive performance using the official nuScenes tracking evaluation protocol. However, when we evaluate in an on-line setting, the proposed tracker significantly outperformed the state-of-the-art. We are hopeful that in the future the nuScenes evaluation server will allow for a true on-line benchmarking. Our future research directions include re-training the detection CNNs to produce the likelihood information necessary for fusion in a Bayesian framework. In this manuscript we showed that the proposed method is capable of handling hard sensor failures by design, but the effect of various soft failure modes such as signal degradation remains to be seen.

[1] v1.1.9: https://github.com/nutonomy/nuscenes-devkit/blob/master/python-sdk/nuscenes/eval/tracking/loaders.py lines:144-168

REFERENCES

[1] S. I. Society of Automotive Engineers, "Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicle," 2018.

[2] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, June 2012, pp. 3354–3361.

[3] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," *CoRR*, vol. abs/1903.11027, 2019. [Online]. Available: http://arxiv.org/abs/1903.11027

[4] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine *et al.*, "Scalability in perception for autonomous driving: Waymo open dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2446–2454.

[5] S. Thrun, "Learning occupancy grid maps with forward sensor models," *Autonomous Robots*, vol. 15, no. 2, pp. 111–127, Sep 2003. [Online]. Available: https://doi.org/10.1023/A:1025584807625

[6] T. Yin, X. Zhou, and P. Krähenbühl, "Center-based 3d object detection and tracking," *CVPR*, 2021.

[7] T. Wang, X. Zhu, J. Pang, and D. Lin, "Fcos3d: Fully convolutional one-stage monocular 3d object detection," *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 913–922, 2021.

[8] M. Dimitrievski, P. Veelaert, and W. Philips, "Behavioral pedestrian tracking using a camera and lidar sensors on a moving vehicle," *Sensors*, vol. 19, no. 2, 2019. [Online]. Available: http://www.mdpi.com/1424-8220/19/2/391

[9] M. Dimitrievski, D. Van Hamme, P. Veelaert, and W. Philips, "Cooperative multi-sensor tracking of vulnerable road users in the presence of missing detections," *Sensors*, vol. 20, no. 17, 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/17/4817

[10] T. Yin, X. Zhou, and P. Krähenbühl, "Multimodal virtual point 3d detection," *NeurIPS*, 2021.

[11] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: The clear mot metrics," *EURASIP Journal on Image and Video Processing*, vol. 2008, no. 1, p. 246309, May 2008. [Online]. Available: https://doi.org/10.1155/2008/246309

[12] N. Benbarka, J. Schröder, and A. Zell, "Score refinement for confidence-based 3d multi-object tracking," *arXiv preprint arXiv:2107.04327*, 2021.

[13] Z. Pang, Z. Li, and N. Wang, "Simpletrack: Understanding and rethinking 3d multi-object tracking," *arXiv preprint arXiv:2111.09621*, 2021.