# Calibration-free match finding between vision and LIDAR

Egor Sattarov, Alexander Gepperth, Sergio Alberto Rodriguez Florez, Roger Reynaud

# Calibration-free match finding between vision and LIDAR*

Egor Sattarov[1], Alexander Gepperth[2], Sergio A. Rodriguez F.[1] and Roger Reynaud[1]

***Abstract*— We present a learning approach that allows to detect correspondences between visual and LIDAR measurements. In contrast to approaches that rely on calibration, we propose a learning approach that will create an implicit calibration model from training data. Our model can provide three functions: first of all, it can convert a measurement in one sensor into the coordinate system of the other, or into a distribution of probable measurements in case the transformation is not unique. Secondly, using a correspondence observation that we define, the model is able to decide if two visual/LIDAR measurements are likely to come from the same object. This is of profound importance for applications such as object detection or tracking where contributions from several sensors need to be combined. We demonstrate the feasibility of our approach by training and evaluating our system on tracklets in the KITTI database as well as on a small set of real-world scenes containing pedestrians, in which our method finds correspondences between the results of real visual and LIDAR-based detection algorithms.**

## I. INTRODUCTION

### A. Context of this work

This article is in the context of multisensory information processing, in particular vision and LIDAR. As these sensors take their measurements independently, it is a priori not clear whether two measurements originate from the same object (or more generally: from the same physical position). To find these correspondences, standard algorithms like object detection and tracking (i.e. DATMO) usually make use of a calibration procedure which allows to transform measurements of one sensor into the reference frame of the other. Such transformations are often quite sensitive [1] to the used measurement models (e.g., pinhole model for camera) and calibration parameters. Moreover, because of the very nature of the measured quantities, sometimes a one-to-one transformation does not even exist. This is for example the case when transforming 2D image points into a 3D coordinate system of a LIDAR device.

### B. Proposed approach

While calibration approaches are often quite precise, the calibration procedure itself is complex and error-prone and requires considerable expertise. Furthermore, a calibration procedure intrinsically depends on the common data representation (e.g. calibration pattern, features), and needs to be re-designed every time a change is made. On the other hand, it is often rather easy and cheap to obtain a large number of sample measurements from both sensors. Assuming the existence of such a sample database, we propose a simple method to extract an implicit calibration model between vision and LIDAR sensors. We pursue a data-driven approach where the statistics of each sensor are used to optimally project both measurements (i.e. object-level) onto a standardized representation format to which generic probabilistic methods can be applied. In this way, our approach is completely independent of the intrinsic characteristics of the measurements, and in particular of their dimensionality (i.e. n-d observations), leading to a strong reduction of design and re-design effort for the conception of multi-modal processing system in vehicles.

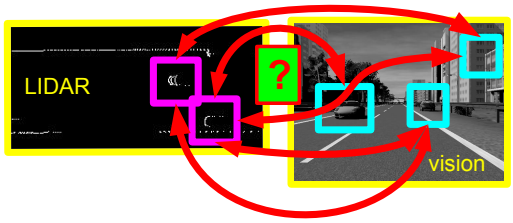


Fig. 1. Illustration of the multisensory correspondence problem: LIDAR (left) and visual (right) measurements, e.g., provided by independent object detection algorithms, "live" in completely different spaces and are thus very difficult to associate without applying prior knowledge.

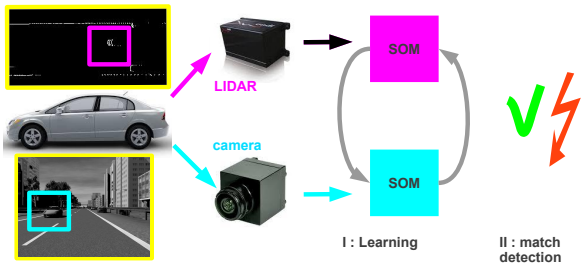


Fig. 2. Block architecture of the proposed correspondence detection method.

### C. Related work

State-of-the-art approaches for IV such as [2],[3] rely on the explicit need of a common frame where all sensors observations can be referenced (i.e. data alignment). This assumption greatly simplifies the association problem of multiple data sources (e.g. LIDAR, Radar, vision). However, in practice, a calibration procedure is required in order to precisely determine all sensors rigid-body transformations (i.e. extrinsic parameters) into the reference frame and their

*This work was supported by a PhD grant from the Digiteo consortium

[1]Egor Sattarov, Sergio A. Rodriguez F. and Roger Reynaud are with Université Paris-Sud and the Institute d'Electronique Fondamentale UMR 8622 CNRS, Orsay, France {egor.sattarov, sergio.rodriguez and roger.reynaud}@u-psud.fr

[2]Alexander Gepperth is with ENSTA ParisTech, 828 Blvd des Marchaux, 91762 Palaiseau Cedex, France alexander.gepperth@ensta-paristech.fr

uncertainties.

Recent works on the 3D sensor calibration have considerably simplified the procedure for determining the relative position of sensors using a set of natural features [4] or using a single observation of a set of calibration patterns ( covering different distances and orientations of the multi sensors field of view)[5].

Automatic calibration approaches can also infer the extrinsic parameters by the means of an optimization framework which registers sensors data in a common space (typically 2D/3D Cartesian space). Recently in [6],[7] and [8], online strategies were proposed to achieve data registration between a vision system and a ranging sensor by optimizing the extrinsics following a mutual information criterion of the sensing sources.

As an alternative to the classical approaches , the presented work is intended to perform multi-sensor data alignment through a probabilistic learning based framework. This approach not only provides a data alignment solution but also models the probability accorded to the observation transfer. Moreover, this method can provide an integrity measure of the data alignment using extrinsic parameters in a cross-validation scheme.

### D. Contributions and novelty

This article presents a new way of detecting multimodal correspondences for the important vision/LIDAR sensor combination that is becoming a standard in the intelligent vehicle domain. A main contribution of the used learning approach is that the "calibration" procedure is much simpler and can in fact be handled by a non-expert regardless of the precise type of measurements that are conducted. Furthermore, we show that the resulting data alignment is very computationally efficient and sufficiently accurate for most applications. Performing all experiments using the publicly available KITTI benchmark database adds significant credibility to our results.

## II. Methods

### A. Architecture overview

The complete model is composed of several components, as visualized in Fig. 2:

- LIDAR and vision sensors
- means to measure interesting quantities in both
- Self-organized Maps (SOM) for vision and for LIDAR, which learn to represent the inputs coming from the respective (synchronous) measurements
- an algorithm for learning a correspondence model between SOMs
- a module for deciding when two measurements correspond, based on the SOMs and the learned correspondence model

Within the scope of this article, we will use both actual measurements proposed by Honda experimental vehicle in form of object positions, and annotated tracklets from the KITTI database[9] as ideal data without noise. They are used separately. As we wish not to complicate the clean and simple algorithm we propose by details of unimodal processing in each modality.

### B. Model training

*1) Learning sensor statistics with self-organizing maps:* The self-organizing map algorithm, while originally proposed as a model cortical information processing, is a generative machine learning algorithm that aims to approximate the distribution of high-dimensional data, and to represent it in a topology-preserving way on a two-dimensional manifold. It is in fact quite related to K-Means[10] except that the preservation of topology makes it interesting for incremental learning scenarios.

SOM defines a fixed $N \times N$ grid of nodes ("neurons") $n_i$, each of which is associated with a so-called prototype vector $\vec{p_i}$. For a given input $\vec{x}$, each node gets assigned an activity $z_i$ based on the distance of its prototype to the input:

$$\begin{aligned} z_i &= d(\vec{x}, \vec{p_i}) \\ d(\vec{a}, \vec{b}) &= \sqrt{(\vec{a} - \vec{b})^2} \end{aligned} \tag{1}$$

As a distance measure, the euclidean distance is often used, and so shall we. In most cases, the calculation of activity is followed by a learning step where the prototypes are adapted to better fit the current input:

$$\begin{aligned} i^* &= \underset{i}{\operatorname{argmin}}\, z_i \\ \vec{p_i}(t+1) &= \vec{p_i} + \epsilon(t) G\left(i^*, i, \sigma(t)\right)\left(\vec{p}_{i^*} - \vec{p_i}\right) \end{aligned} \tag{2}$$

$G(i, j, \sigma = \exp\left(-\frac{d^2(i,j)}{2\sigma^2}\right))$ is a Gaussian with standard deviation $\sigma$ which is based on the euclidean distance between node $i$ and node $j$ on the two-dimensional grid of nodes. For faster convergence, the algorithm demands to gradually lower the learning rate $\epsilon(t)$ and their neighbourhood radius $\sigma(t)$ from initially large values $\epsilon_0,\ \sigma_0$ until the minimal values $\epsilon_\infty,\ \sigma_\infty$ are reached.

*2) Learning of conditional distributions between sensors:* Supposing the SOMs are trained using the algorithm described in Sec. II-B.1, correspondences between visual and LIDAR SOMs are detected using a simple probabilistic counting approach. Assuming that two sets of weights $w_{ij}^L,\ w_{ij}^V$ exist between nodes $i,\ j$ in visual and LIDAR SOMs, both are updated as follows for each simultaneously presented pair of visual and LIDAR measurements $\vec{x}^V,\ \vec{x}^L$:

$$\begin{aligned} \tilde{z}_i^X &= \begin{cases} 1 & \text{if } i = \underset{k}{\operatorname{argmin}}\, z_k^X \\ 0 & \text{else} \end{cases} \\ \tilde{z}_i^{\bar{X}} &= \begin{cases} 1 & \text{if } i = \underset{k}{\operatorname{argmin}}\, z_k^{\bar{X}} \\ 0.5 & \text{if } i \text{ is neighbour to } \underset{k}{\operatorname{argmin}}\, z_k^{\bar{X}} \\ 0 & \text{else} \end{cases} \\ w_{ij}^X &= w_{ij}^X + \tilde{z}_i^X \tilde{z}_j^{\bar{X}} \end{aligned} \tag{3}$$

where we have used a shorthand notation $X = L, V$ ($\bar{X}$ denoting the other modality, i.e., $L$ if $X = V$ and $V$ otherwise). After a sufficient amount of samples has been

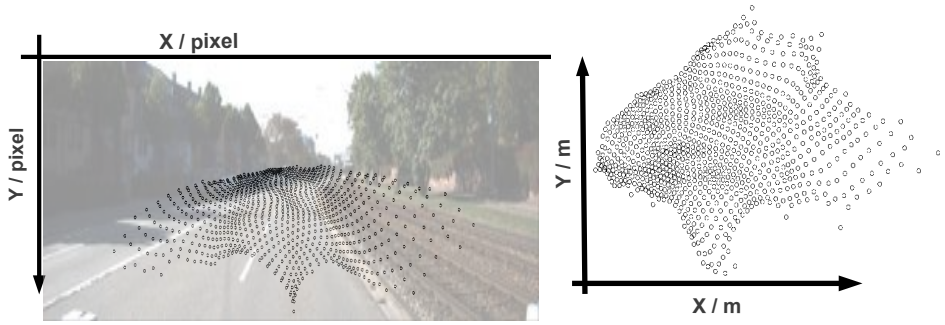


Fig. 3. Statistical models of sensory spaces acquired by self-organizing maps (SOM) for visual (left) and LIDAR sensors (right). The points represent the position of SOM prototypes in the space of each sensor. The local density of prototypes is guided by average local density of data points.

processed, we normalize the weight matrices in order to obtain normalized probabilities:

$$\Sigma_i^X = \sum_j w_{ij}^X$$
$$w_{ij}^X \to \frac{w_{ij}^X}{\Sigma_i^X} \tag{4}$$

It has to be noted that the visual and LIDAR measurements do **not** need to come from the same objects. Indeed, if this were the case, it would mean that we already know the correspondences we wish to identify. When working on a benchmark database like KITTI, this is the case but when training the system on recorded data not containing any annotations, evidently the correct correspondences are unknown except when there is always just a single object in sight. Therefore, the strategy we adopt is to present all *combinations* of visual and LIDAR measurements taken at a certain point in time (e.g. a single, synchronized image and LIDAR recoding, both for real sensors and in the case for KITTI) when learning conditional probability distribution between sensors. This assumes there is a sufficient amount of training data, because the "correct" correspondences will appear together far more often than random incorrect ones.

As we supposed that SOMs are already converged, we disable SOM learning during the whole phase of learning conditional distributions by setting $\epsilon(t) \equiv 0$ for both SOMs.

*3) Overall training procedure:* The overall training procedure is given in Alg. 1. It consists of a SOM training step and a step that determines conditional probabilities between the SOM representations of both measurements.

**Algorithm 1** Model Training: *Overview over the two-stage model training procedure consisting of learning distributions with SOMs, and learning multi-sensory conditional probabilities.*

**for** $t : 1 \to T_{\text{SOM}}$ **do**
  Draw a random image $i$ from $D_{\text{train}}$
  Draw random visual measurement $l$ $\vec{x}_{il}^V$ from $i$
  Draw a random image $i$ from $D_{\text{train}}$
  Draw a random LIDAR measurement $m$ $\vec{x}_{im}^L$ from $i$
  Update visual SOM with $\vec{x}_{il}^V$ acc. to Sec. II-B.1
  Update LIDAR SOM with $\vec{x}_{im}^V$ acc. to Sec. II-B.1
**end for**
Disable learning in SOMs by setting $\epsilon(t) \equiv 0$
**for** $t : 1 \to T_{\text{corr}}$ **do**
  Draw a random image $i$ from $D_{\text{train}}$
  **for** $(l, m) =$ all permutations of measurements **do**
    Feed visual SOM with $\vec{x}_{il}^V \to \vec{z}^V(t)$
    Feed LIDAR SOM with $\vec{x}_{im}^L \to \vec{z}^L(t)$
    Update $w_{ij}^L,\ w_{ij}^V$ acc. to Sec. II-B.2
  **end for**
  Normalize $w_{ij}^L,\ w_{ij}^V$ acc. to Sec. II-B.2
**end for**

## C. Unimodal detection of correspondences

After training is completed, the model can be used for detecting whether a given combination of visual and LIDAR measurements is likely caused by the same object. To this end, we develop a criterion that depends on a single parameter, the probability threshold $\theta$. Assuming that each measurement has generated activities $z_i^X$ in both SOMs, the criterion first computes a single binary measure $c^X = \{0, 1\}$ for each conditional probability matrix $w_{ij}^X$, using the shorthand notation $X = L, V$ for a certain modality, and "other" for the other one:

$$i^* = \underset{i}{\operatorname{argmin}}\, z_i^X \tag{5}$$
$$j^* = \underset{j}{\operatorname{argmin}}\, z_j^{\bar{X}} \tag{6}$$
$$P^{\bar{X}} = \left\{ j | w_{i^* j}^X > \theta \right\} \tag{7}$$
$$c^X = \begin{cases} 1 & \text{if } j^* \in P^{\bar{X}} \\ 0 & else \end{cases} \tag{8}$$

The two quantities $c^X$ express whether a best-matching unit (BMU) at position $i^*$ in $X$ can predict the best-matching unit at index $j^*$ in the *other* modality $\bar{X}$ based on the learned conditional probabilities. Given a best-matching unit in $X$, $\theta$ is used for selecting a set of nodes $P^{\bar{X}}$ with conditional probabilities that exceed $\theta$. If the BMU of $\bar{X}$ is an element of the selected set, we conclude that there is a match and set $c^X = 1$. Thus, the threshold $\theta$ governs the strictness of the matching: if it is high, only a small (or empty) set of nodes $P^{\bar{X}}$ will be selected and the probability of match

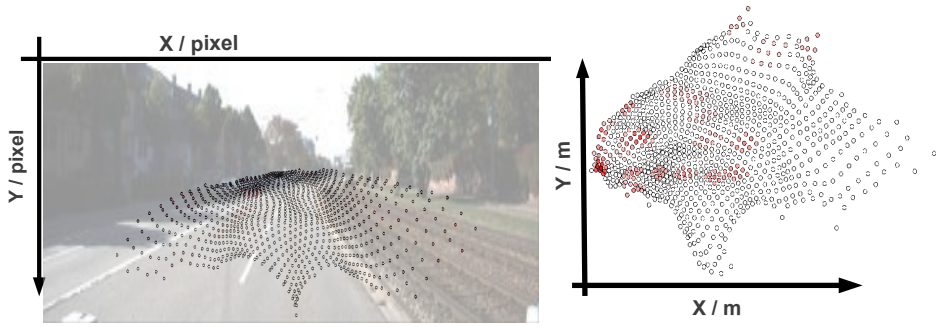


Fig. 4. Example of conditional probability distributions $P^X$ for vision (given a LIDAR node, left) and LIDAR (given a vision node, right). These distributions are used to detect correspondences.


Fig. 5. The KITTI database we use for our experiments is recorded from a moving car equipped with several cameras, a GPS device and most notably a Velodyne LIDAR device.

diminishes. On the other hand, if $\theta$ is low, the probability of match increases, up to the point where there will always be a match at $\theta = 0$. As it is often not necessary to detect all correspondences correctly but rather to exclude unlikely combinations, a more relaxed value of $\theta$ can help avoid missed correspondences while still being able of reducing the combinatorial space of correspondences.

### D. Fused correspondence detection

Apart from the unidirectional mutual sensor activity predictions, one can also use a cross-verified decision of improved quality. For that, the criterion of acceptance in eqn.(8) changes to:

$$w^X_{i^* j^*} \times w^{\bar{X}}_{j^* i^*} > \theta \tag{9}$$

$$w^X_{i^* j^*} + w^{\bar{X}}_{j^* i^*} > \theta \tag{10}$$

$$\sum_k w^X_{i^* k} z^{\bar{X}}_k \times \sum_k w^{\bar{X}}_{j^* k} z^X_k > \theta \tag{11}$$

where $z^X_i$ is again the activity at node $i$ in SOM X (which can be LIDAR or vision, whereas $\bar{X}$ represents the other modality), and the indices $i^*$, $j^*$ are the indices of the BMU's in both sensor's SOMs. The last eqn. (11) takes into account not only the BMU of each SOM, but also its neighbouring nodes plus their associated, learned conditional probabilities.

### E. Training and evaluation data

For the training stage and the evaluation of the proposed methods, two datasets are used:

- Dataset A is composed of annotated tracklets from the public KITTI benchmark database[9] (see also Fig. 5).
- Dataset B is composed of real detections captured from dash camera and four-layer lidar on-board an experimental platform. Visual pedestrian detections are obtained with the 'daimer' detector provided with the OpenCV vision library, and LIDAR ones are from connectivity based clustering.

From Dataset A: we employ the center positions of objects in 2D image coordinates as well as corresponding 3D laser coordinates as measured by a Velodyne laser scanner (i.e. tracklets). As the height-over-ground of a tracklet's center is often irrelevant for safety applications, we take a birds-eye perspective and just consider two of the three 3D coordinates, excluding height-over-ground. Due to the synchronized nature of visual and LIDAR recordings in the Dataset A, each tracklet can be assigned a unique visual image and therefore a corresponding LIDAR sweep. We use all types of objects provided by the database A, making the total number of considered tracklets 23497. For training the model, we use 70% of this data, performing a random split of available tracklets into train and test databases.

From Dataset B: we employ the center positions of objects in 2D image coordinates as well as corresponding 3D laser point cluster center positions. Vision-based objects and lidar-based detection were manually associated so as to obtain a ground-truth reference. This sequences is composed of pedestrians filmed in 9 short scenarios of about 2 minutes, making a total number of 8613 visual detections, and 5476 lidar detections. Due to the small size of this data base, we performed a cross-validation, that is, for each scenario the SOM are trained with 8 other scenarios and tested with the chosen one.

### F. Evaluation

In order to quantify the capacity of the trained model to identify visual/LIDAR correspondences, we use the test database as described in Sec. II-E. In order to prevent the SOMs from adapting during the evaluation phase, we set $\epsilon(t) \equiv 0$ for both SOMs.

Assuming a trained model (SOMs plus conditional probabilities), we process all images in the test database in a

**Algorithm 2** Evaluation: *Overview over the evaluation procedure.*

Disable learning in both SOMs by setting $\epsilon(t) \equiv 0$
**for** $i : 1 \rightarrow$ (images in $D_{\text{test}}$) **do**
  Draw image $i$ from $D_{\text{test}}$
  **for** $(l, m) =$ combinations of measurements **do**
    Feed visual SOM with $\vec{x}_{il}^{V} \rightarrow \vec{z}^{V}(t)$
    Feed LIDAR SOM with $\vec{x}_{im}^{V} \rightarrow \vec{z}^{L}(t)$
    Generate bin. measures $c^X$, $c^X$ acc. to. Sec. II-C
  **end for**
**end for**
Plot precision/recall curves

sequential manner. For each image, we present all combinations of visual and LIDAR measurements and compute the scores $c_L, \ c_V$ for each combination. A binary decision on the presence of a correspondence is taken according to eqn.(5). As this decision depends on a single threshold $\theta$ we can re-cast this evaluation in the form of a ROC analysis by varying $\theta$ in the interval $[0, 1]$ and measuring the precision/recall rates.

An overview over the complete evaluation procedure is given in Alg. 2.

## III. Experiments

### A. Organization of training and evaluation

Model training is performed in two steps: initially, the SOMs are trained independently of one another by drawing random samples from the train database, see Sec. II-E, and adapting each individual SOM according to Sec. II-B.1, with the input vector provided by the unimodal part of the drawn sample. Training parameters are: $N = 30$, $\epsilon_\infty = 0.01$ $\sigma_\infty = 1$, $\epsilon_0 = 0.6$, $\sigma_0 = \frac{N}{2}$. Neighbourhood radius and learning rate develop according to

$$\sigma(t) = \max\left(\sigma_\infty, \sigma_0 \exp(-\lambda_\sigma t)\right) \tag{12}$$

$$\epsilon(t) = \max\left(\epsilon_\infty, \epsilon_0 \exp(-\lambda_\epsilon t)\right), \tag{13}$$

with $-\lambda_\epsilon = 0.002$ and $\lambda_\sigma = 0.004$.

SOM training duration is limited to $T_{\text{SOM}} = 20000$ iterations. Subsequently, correspondences are trained according to Sec. II-C for another $T_{\text{corr}} = 20000$ iterations, randomly drawing *images* from the training database and feeding all possible combinations of visual/LIDAR measurements to the two SOMS as well as updating the two sets of weights $w_{ij}^{V}$, $w_{ij}^{L}$ based on the resulting SOM activities $z_j^X$, $X = L, V$. Evaluation is conducted according to Sec. II-F by iterating over all *images* in the test database and measuring precision/recall rates when presenting to the model all possible combinations of visual/LIDAR measurements in each image.

### B. Results

For KITTI base we first plot a separate ROC for LIDAR-vision and vision-LIDAR correspondence detection, given in Fig. 6. As can be expected, the LIDAR-vision-based correspondence detection gives better results, very likely because the vision-LIDAR transformation is one-to-one but not the other way round. We observe as well that performance is acceptable given that no prior knowledge was used at all but it is not an ideal ROC either. In a further experiment, we wish to back the claim made in Sec. I that the proposed method was able to handle arbitrary measurements without requiring explicit models. To this end, we repeat the previous experiment while tracklet width and tracklet height to the laser measurement, bringing up its dimensionality to 4. The ROCs obtained in this way are shown in Fig. 7. We see that the addition of additional information does not impair the ability of our system to detect correspondences. On the other hand, performance is not improved either, because the added information is irrelevant to the transformation to be computed. This experiment therefore shows that our model, due to the learning approach, is able to process very diverse types of measurements, and automatically extracts the information required for finding correspondences. Lastly, we evaluate the three fusion strategies proposed in Sec. II-D, which means that for a pair of visual and LIDAR measurements, there will now be only one decision on correspondence, not two as in previous experiments. The overall performance is shown in Fig. 8 and show that the fused decision outperforms any single unimodal one, boosting the already satisfactory

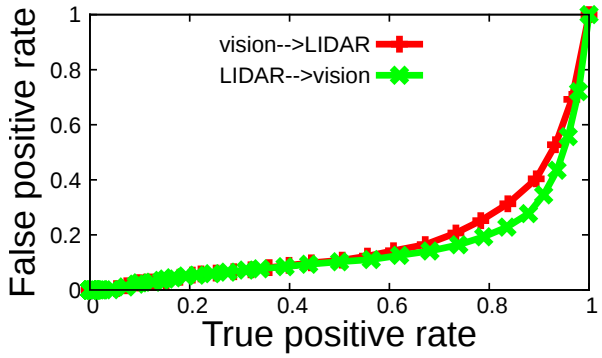


Fig. 6. ROCs for vision-LIDAR (red curve) and LIDAR-vision (green curve) correspondence detection. As can be expected, LIDAR-vision provides slightly better performance as the associated transformation is one-to-one.

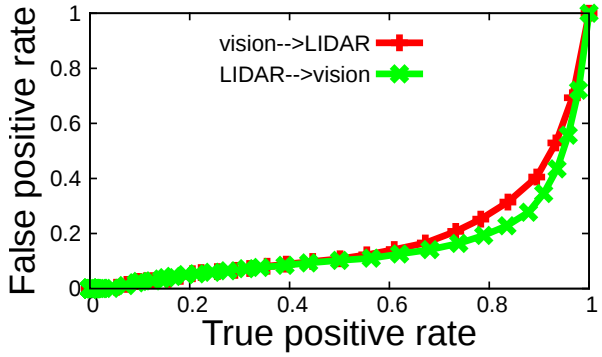


Fig. 7. ROCs for vision-LIDAR (red curve) and LIDAR-vision (green curve) correspondence detection, where laser measurements are augmented by object size. By comparison to Fig. 6, we observe that this irrelevant information is ignored.

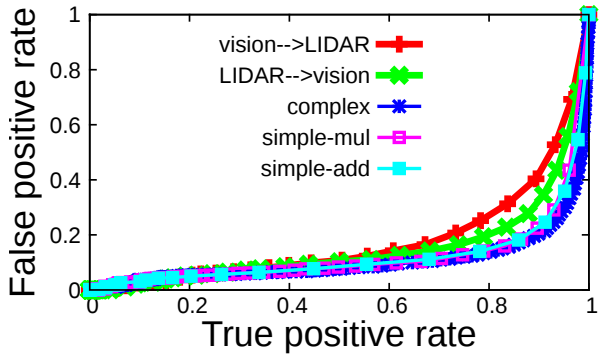


Fig. 8. ROCs for fused correspondence detection. The blue "complex" curve represents the cross-verified strategy of eqn.(11). The cyan "simple-add" curve corresponds to eqn. (10), and the violet "simple-mult" one to eqn. (9). It is apparent that all fusion methods outperform the unimodal ones (red and green curves).

performance even further.

For real sensors the ROCs are calculated using only complex activity predictions [11] as the most effective. The results are seen in Fig. [9]. One can observe low quality for unidirectional correspondences detections and very high quality for fused one. It can be explained by non-symmetrical detections nature and small number of detected objects per frame.

## IV. Discussion and conclusions

We have presented a learning approach to solve the problem of finding visual/LIDAR correspondences and validated its performance on a widely accepted benchmark database. In this section, we will review and justify the components of our model and outline principal conclusions and further research works.

### A. Model justification

The hybrid SOM-based architecture we propose here is based on two necessities: first, to have a generic model that will work with any kind of visual/laser measurements. This means that the model must be able to work regardless **what** is actually measured by each sensor. For a camera, this could be, e.g., pixel position of interest points, but also center position, size and identity if an object detection algorithm is used, or center position, size and speed if tracking is added. By using the self-organizing map architecture, every measurement is down-projected to a 2D image-like representation in a way that is statistically optimal and respects a certain topological constraint that allows to easily visualize and interpret a SOM's activity. For ensuring statistical optimality, we use a variant of the SOM model that has a well-defined energy function[11], which makes it actually very easy to detect measurement outliers that should be ignored.

Secondly, we want a model that will not fail even when the transformation between modalities is not one-to-one in both directions. To this end, we adapted a purely probabilistic approach, on top of the SOM mechanism, that will simply respond by a multi-peaked probability distribution in case there is inherent ambiguity due to non-unique transformations.

### B. Discussion of results

As seen in Sec. III, the quality of correspondence finding is very satisfactory given that we did not bring in *any* specific expert knowledge. In addition, the threshold $\theta$ allows us to smoothly change the behavior of the system, from a point where there are few correct correspondences but also few incorrect ones, to a point where there are many correct correspondences but also some incorrect ones. For example, for a multimodal tracking system a higher false positive rate can be acceptable if no correspondences are incorrectly rejected, since tracking can take into account past information and thus correct the occasional incorrect correspondence. Another very encouraging fact is that the quality of correspondence detection can be significantly improved by considering not only both unidirectional correspondences in isolation, but a fusion of both. As a proper fusion should be, it is indeed better-performing than any single contribution to it.

### C. Conclusion

We have shown that a learning-based approach can successfully solve the problem of multimodal correspondence detection, in particular between visual and LIDAR sensors. The only prerequisite is a collection of (unlabeled) data which is usually easy to obtain. No expert effort is required at all, and in particular no detailed models of the data acquisition process by the used sensors. The technique is very computationally efficient, and consumes no significant computational load, thus making it suitable for embedded operation. We hope to make this technique even more appealing by better exploiting the structure of conditional probabilities for even better-performing fusion strategies.

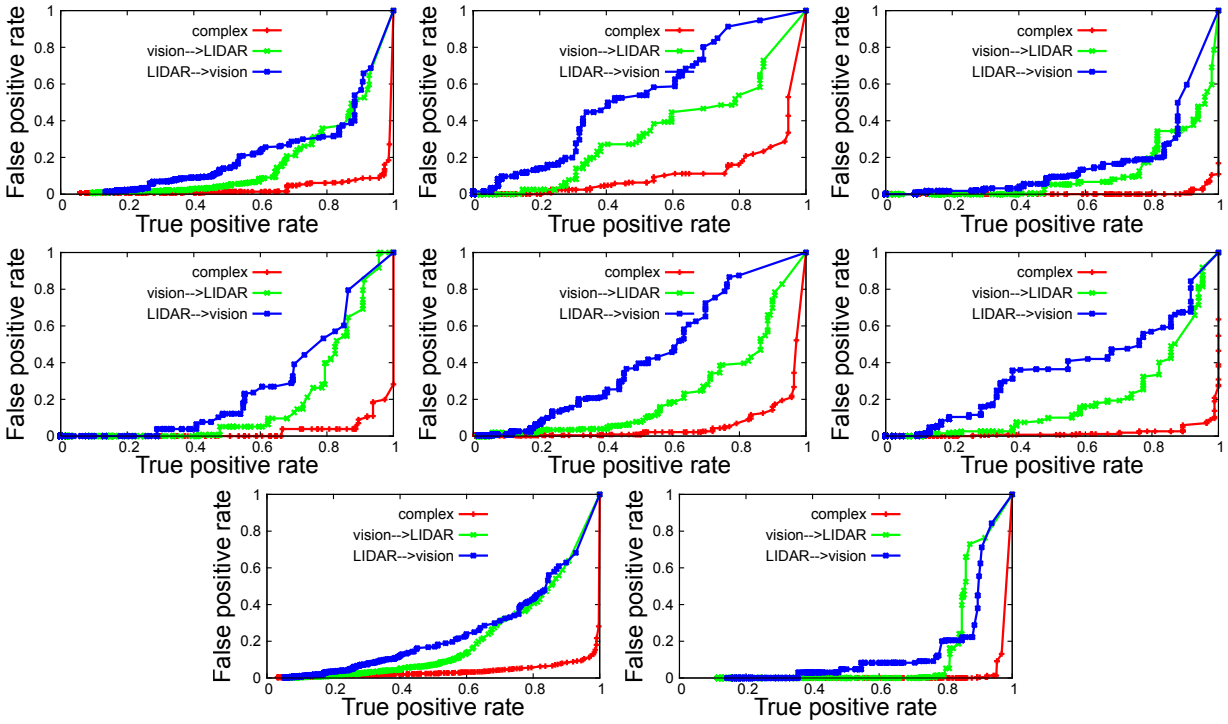


Fig. 9. ROCs for fused correspondence detection in the case of real sensors data. The red "complex" curve represents the cross-verified strategy of eqn.(11). unimodal ones (red and green curves)

## REFERENCES


[1] S. A. Rodriguez, V. Fremont, and P. Bonnifait, "Influence of intrinsic parameters over extrinsic calibration between a multi-layer lidar and a camera," in *IEEE 2nd Workshop on Planning, Perception and Navigation for Intelligent Vehicles*, vol. 1, 2008, pp. 34–39.

[2] H. Cho, Y.-W. Seo, B. Vijaya Kumar, and R. 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*, May 2014, pp. 1836–1843.

[3] S. A. Rodriguez, V. Fremont, P. Bonnifait, and V. Cherfaoui, "Multi-modal object detection and localization for high integrity driving assistance," *Machine Vision Applications*, vol. 1, pp. 1–18, 2011.

[4] D. Scaramuzza, A. Harati, and R. Siegwart, "Extrinsic self calibration of a camera and a 3d laser range finder from natural scenes," in *Intelligent Robots and Systems, 2007. IROS 2007. IEEE/RSJ International Conference on*, Oct 2007, pp. 4164–4169.

[5] A. Geiger, F. Moosmann, O. Car, and B. Schuster, "Automatic calibration of range and camera sensors using a single shot," in *International Conference on Robotics and Automation (ICRA)*, 2012.

[6] G. Pandey, J. R. McBride, S. Savarese, and R. M. Eustice, "Automatic extrinsic calibration of vision and lidar by maximizing mutual information," *Journal of Field Robotics*, 2014. [Online]. Available: http://dx.doi.org/10.1002/rob.21542

[7] J. Levinson and S. Thrun, "Automatic online calibration of cameras and lasers," in *Robotics: Science and Systems*, 2013.

[8] A. Napier, P. Corke, and P. Newman, "Cross-calibration of push-broom 2d lidars and cameras in natural scenes," in *Robotics and Automation (ICRA), 2013 IEEE International Conference on*, May 2013, pp. 3679–3684.

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

[10] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*. Prentice-Hall, 1981.

[11] T. Heskes, "Energy functions for self-organizing maps," in *Kohonen maps*, E. Oja and S. Kaski, Eds., 1999.