# XDvision: Dense & Robust Outdoor Perception for Autonomous Vehicles

Victor Romero-Cano[1], Nicolas Vignard[2] & Christian Laugier[3]

***Abstract*— Robust perception is the cornerstone of safe and environmentally-aware autonomous navigation systems. Autonomous robots are expected to recognise the objects in their surroundings under a wide range of challenging environmental conditions. This problem has been tackled by combining multiple sensor modalities that have complementary characteristics. This paper proposes an approach to multi-sensor-based robotic perception that leverages the rich and dense appearance information provided by camera sensors, and the range data provided by active sensors independently of how dense their measurements are. We introduce a framework we call *XDvision* where colour images are augmented with dense depth information obtained from sparser sensors such as lidars. We demonstrate the utility of our framework by comparing the performance of a standard CNN-based image classifier fed with image data only with the performance of a two-layer multimodal CNN trained using our augmented representation.**

## I. INTRODUCTION

Environment perception is the first building block of any autonomous robotic system. Perception is performed by collecting measurements from sensors and then processing this information in order to generate knowledge about both the layout of the environment and the objects in it. Passive sensors such as cameras for example, are inexpensive and provide dense and rich appearance information. Active sensors such as lasers on the other hand, provide sparser depth and heading measurements. Point clouds generated by a laser or images recorded by a camera have been extensively used to detect generic objects in urban environments. The works in [12], [30] present approaches for generic object detection in lidar and camera data respectively. At a high level, these methods cluster sensor data into groups that individually correspond to objects in the environment. Although these one-sensor-modality-based methods have in general, a state-of-the-art performance, there are adverse environmental conditions that limit the usability of the data provided by the individual sensors they utilise. Airborne dust for example, can significantly increase the noise in depth measurements, whereas illumination variations make some portions of the image too bright or too dark and therefore unusable.

Multi-sensor data fusion is the most promising way for achieving all-weather-conditions perception [14]. It allows robotic systems to compensate for the weaknesses of a given sensor modality using the strengths of another complementary one. Sensor fusion methods can be classified according to the level of abstraction at which fusion is performed [15]. High level fusion methods perform estimation using each sensor modality independently and then approach the fusion of these estimates as a data association problem [27], [10], [7]. These approaches require independent estimation machinery for each sensor modality which makes them not only overly complex but also disregard raw information that could be valuable for the fusion process. There is an increasing number of estimation frameworks that make use of a lower level data representation where raw measurements from all sensing modalities are consider in the fusion process. The works in [23], [5], [29], for instance, perform multi-modal motion detection, semantic segmentation and sensor calibration respectively, based on per-pixel appearance and geometric features. They utilise unified and sound methods for processing all sensor modalities at once. All these works, except [5] require sensor measurements to densely cover the measured scene. Therefore there is a clear need of low- or pixel-level multi-modal fusion methods that transform raw sensor data into a common and dense representation that can eventually be processed by a unified recognition method. There are some works in the literature that approach this low-level fusion problem using non-parametric [11] or energy-based [17] approaches. They however require all sensor modalities to be similarly dense. This paper introduces a framework for low-level multi-sensor data fusion in the pixel space that is independent of the scene coverage of any of the sensors. Our framework outputs a new image-like data representation where each pixel contains not only colour but also other low level features such as depth and object IDs.

Our approach is generic so it allows for the integration of data coming from any active sensor into the image space. Additionally, it does not aim at tackling the object detection problem directly but it proposes a multi-modal-data representation from which object detection methods may benefit. In this paper we tackle the concrete problem of fusing images and sparse lidar returns, however, as explained before, the framework is amenable for the inclusion of any other sensor modality. The framework presented in this paper creates *XDimages* by extrapolating range measurements across the image space in a two-stage procedure. The first stage considers locally homogeneous areas given by a super-pixel segmentation while the second one further expands depth values by performing self-supervised segmentation of areas seeded by the range sensor. In summary, the main contributions of this paper are as follows:

[1]V. Romero-Cano is with Universidad Autónoma de Occidente, Cali, Colombia, e-mail: varomero@uao.edu.co

[2]N. Vignard is with Toyota Motor Europe (TME), Brussels 1140, Belgium, e-mail: nicolas.vignard@toyota-europe.com

[3]C. Laugier is with INRIA Rhone Alpes, Grenoble 38334, France, e-mail: christian.laugier @inria.fr

- The XDvision framework: an approach to sparse-feature densification in the image space,
- A new multi-sensor data representation: colour images augmented with dense depth and object hypothesis,
- An experimental validation that proves the advantages of our new data representation for the problem of object recognition.

## II. RELATED WORK

With the advent of self-driving cars, there has been in the robotics community an increasing interest in the development of robust perception systems that provide correct estimates even under different and challenging environmental conditions. Although some approaches to robust perception resort to statistical methods for dealing with data outliers [13], [2], the work presented in this paper belongs to the group that tackles the robust-perception problem by leveraging the complementarynature of pasive and active sensor modalities. Multi-sensor approaches to robotic perception, can be categorised according to the level at which the data from each sensing modality is fused in order to obtain the estimate of interest. According to [6], data fusion can be made at the level of symbolic estimates or *high level fusion*, at the level of features or *medium level fusion*, or at the level of raw data or *low level fusion*.

*High level fusion* methods fuse estimates obtained by independently processing the data from each sensor modality. The work in [18] for example, uses a combination of monocular camera and lidar for detecting and tracking pedestrians in urban environments. The system uses a Kalman Filter-based approach to segment and track objects from the lidar data. These objects are then projected onto the image in order to extract Regions of Interest (ROI). ROIs in the image space are finally fed to a classification scheme composed by several Support Vector Machine (SVM) classifiers. Other works that follow this high-level data fusion pipeline can be found in [7], [26], [27].

*Medium level fusion* methods extract features from each sensor's raw data and then fuse this medium level representation. The works in [21], [22] extract features such as texture statistics from image-based detections and, shape and location features from stereo-vision depth measurements. All these features are then used to learn a Probabilistic Graphical Model (PGM) particularly tailored for the problem of simultaneous tracking and classification. Other works that follow this medium-level data fusion pipeline can be found in [5], [19], [31], [25].

*Low level fusion* methods on the other hand explore the complementary relationships between passive and active sensors at the pixel level. The approaches in [23], [17], [28] follow this intuition but require the fused modalities to have similar coverage densities. Additionally, low level fusion fosters the development of recognition approaches that use an improved and unified version the multi-modal information in the object recognition task [9]. Our proposed XDvision framework provides a procedure for fusing lidar and image data independently of the lidar data's density. We show how this low level fusion improves object recognition in urban environments.

## III. THE XDVISION FRAMEWORK

This section introduces the XDvision framework, a method that provides a new data structure for the representation of multi-modal sensor information in the image space. The framework's pipeline is given in Figure 1.

We call an instance of our data structure an *XDimage*. It corresponds to an augmented camera image where individual pixels contain both appearance and geometric information. The first and more challenging problem to be solved in order to build XDimages is that of densifying sparse point cloud data provided by active range sensors. In our approach we extrapolate depth information using a two-steps procedure as follows:

1) Extend depth values projected onto individual pixels to neighbouring pixels that have similar appearance. This step is described in Section III-A.
2) Obtain range-based object hypothesis.
3) Extrapolate range measurements in order to account for entire objects. This step is described in Section III-C.

### A. Depth densification via super-pixel-guided extrapolation

In order to build XDimages, range measurements are first projected on the image space. These sparse depth measurements are locally extended using Simple Linear Iterative Clustering (SLIC) [1]. SLIC is a simple and parallelisable method, based on k-means clustering, for decomposing an image into a regular grid of visually homogeneous regions or so-called super-pixels. As a result, SLIC super-pixels provide a regular grouping of image pixels according to their distance both spatially and in the colour space. We utilise this super-pixel segmentation for two complementary tasks. First, the super-pixels are used to assign depth values to all of the pixels within super-pixels with at least one range measurement. The super-pixel segmentation along with the object hypotheses explained in the following section, are then used for computing an accurate initialisation of object-wise appearance models that will guide our final extrapolation/segmentation stage which is explained in III-C.

### B. Object hypothesis generation

In order to generate object hypothesis from the lidar sensor, the occupancy grid provided by the Hybrid Sampling Bayesian Occupancy Filter (HSBOF) [16] is utilised. The occupancy grid is thresholded and connected components analysis is then used to get the final object hypotheses. Note that these hypotheses are defined on the ground/grid space, thus they are 2D and do not convey object-height information. We propose the use of interactive image segmentation in order to add a third -height- dimension to our object hypotheses as shown in Section III-C. Note as well that in this work we consider objects as clusters of data rather than high-level abstractions with a semantic description. Our object hypotheses could encompass not only moving objects

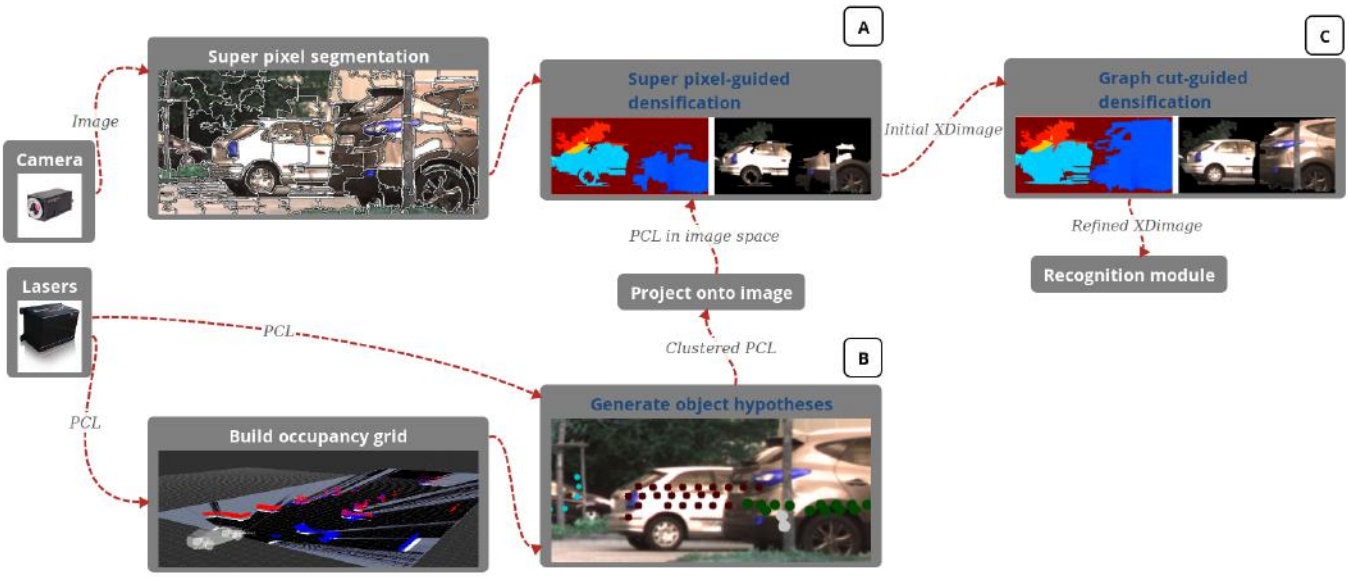


Fig. 1. The XDvision framework.

such as cars and pedestrians but also static ones, including trees, walls or green areas.

### C. Graph cut-guided densification

Due to the sparse nature of range measurements, the super-pixel guided depth densification step presented in Section III-A may in cases not be enough. In many applications, lidar returns tend to be concentrated at the bottom of objects or very sparcely distribted. As a result of this and the fact that objects have heterogeneous appearances, super-pixels with range measurements do not always cover entire objects. In order to further extend depth measurements so that depth estimates are available for entire object hypotheses, we propose a self-supervised segmentation procedure based on the graph-cuts algorithm for interactive foreground/background segmentation, also known as *grab-cut* [24]. This Section starts with a summary of the graph cut model applied to image segmentation [3]. It then explains how this model was extended to perform interactive foreground extraction by [24], and how we use it as a self-supervised method for performing lidar-aided object segmentation.

In interactive segmentation we aim at segmenting out the foreground from the background based on foreground and background appearance models that are known a-priori. These models can be assembled from user input or as in our methodology from an initial lidar-based segmentation $T_F$. We start by formulating an energy function that encodes the trade-off between a good pixel-wise segmentation and spatial coherence. A good segmentation will follow the distributions provided by our initial segmentation but it will also enforce spatial smoothness. This trade-off can be captured by an energy function of the form:

$$E\left(\mathbf{x}, w, \mathbf{z}\right) = U\left(\mathbf{x}, w, \mathbf{z}\right) + V\left(\mathbf{x}, \mathbf{z}\right), \tag{1}$$

where $\mathbf{x} \in \{0, 1\}$ represents the segmentation output, $w = \{h_B\left(z_i\right), h_F\left(z_i\right)\}$ corresponds to the colour distributions for background and foreground parametrised via Gaussian Mixture Models (GMMs). The term $U$ measures the fit of the segmentation $\mathbf{x}$ to the the data $\mathbf{z}$, given the model $w$. In this work we propose to initialise the parameter $w$ based on the initial segmentation provided by our super-pixel-guided depth extrapolation module presented in Section III-A. Finally, $V$ is a smoothness term that encourages nearby pixels to have the same label. The optimal segmentation is obtained by jointly optimising $w$ and $\mathbf{x}$ as follows:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \min_{w} E\left(\mathbf{x}, w, \mathbf{z}\right). \tag{2}$$

Optimisation is performed by iteratively updating the segmentation $\mathbf{x}$ using graph cuts and the parameters $w$ using the Expectation Maximisation (EM) algorithm [8]. Algorithm 1 provides the complete pseudo-code of our XDvision framework. In this algorithm, lines 1 to 6 perform the preprocessing which include gathering the multi-modal raw data, building the 2D occupancy grid, generating object hypothesis and projecting the clustered lidar points on the image space. Lines 1 to 6 constitute the core of the approach which includes obtaining the initial depth densification and image segmentation, initialising the background and foreground appearance models and obtaining the final dense depth map and segmentation.

## IV. EXPERIMENTS

This section presents the experimental evaluation of our XDvision framework in terms of per-object depth extrapolation and the enhanced recognition capabilities it induces. The effect on recognition performance of our XDimages is evaluated by comparing the recognition performance of a CNN classifier based on images only, and based on our

**Algorithm 1** The XDvision framework

```
Collect camera image
Collect PCL
Build occupancy grid
Obtain object hypotheses by clustering occupancy grid
Perform super-pixel segmentation
z ← Project clustered PCL onto image plane
for Each object hypothesis do
    T_F ← Perform super-pixel-based segmentation/depth extrapolation
    h_F(z_i) ← Initialise foreground model
    h_B(z_i) ← Initialise background model
    for sweep = 1 − 5 do
        Update x given current w = [h_F, h_B] using graph cuts
        Update w given current x using EM
    end for
end for
```

XDimages. For the super-pixel segmentation we used the GPU implementation by [20] of the super-pixel method in [1]. Finally, for the graph-cut-based segmentation step, we utilised the CPU implementation of the GrabCut algorithm available in the OpenCV library [4].

Our multi-modal CNN is composed of two independent CNNs that converge in one fully connected layer and a softmax classifier. There is one layer for color patches and another for depth-map patches. We learn each layer by fine-tuning a pre-trained Alexnet CNN. Since our per-sensor CNNs were pretrained on colour images, colour patches from our dataset can be used directly. On the other hand depth-map patches are normalised, padded and then converted into a jet color space before being used for fine-tuning the depth layer.

### A. Experimental setup

Training was performed using the KITTI dataset, in particular, its object detection benchmark. The training part of the dataset, which is conformed by a total $7480$ images, contains a representative number of *cars*, *cyclist*, *pedestrians* and *vans*, within other object categories which all together were considered as *other*-object class. The dataset contains both stereo-vision and velodyne data. For training, we used the only the left images from the stereo pairs and the velodyne points that fall within the camera's field of view. In this section we provide a quantitative evaluation of our XDframework based on the recognition performance it induces. The validation dataset is a subset of the KITTI dataset which provides semi-dense velodyne lidar data.

*1) Depth extrapolation:* Our depth extrapolation method accounts for most of the objects of interest in the scene as long as they have been detected by the range sensors. Once lidar range measurements are projected onto the image space, only a $20\%$ of the surface of the object hypotheses is covered in average. Our super-pixel guided segmentation procedure extends depth coverage to an average of $50\%$. Finally, our graph-cut based densification step increases coverage to an average of $85\%$. Figure 2 illustrates these results.

Fig. 2. Two instances (columns) of our depth extrapolation results. The top row shows the clustered lidar points projected on the image. In the middle and bottom rows, the original image has been masked with the depth extrapolation output. Cooler (blue) colours represent smaller depth values. The middle row represents the super-pixel-based extrapolation result whereas the bottom row shows the final graph cut-based output

### B. Training of the multi-modal classifier

In order to learn the parameters of the multi-modal CNN presented in [9], we extract per-object depth maps by projecting velodyne lidar returns on the image plane and performing depth extrapolation as explained in Section III-A. Subsequently, depth maps are rendered into depth images by normalising all depth values onto the range $0 - 255$ and then applying a jet colourmap as proposed by [9]. Figure 3 illustrates the obtained depth images.

### C. Recognition performance

In order to test the recognition performance induced by the use of our XDvision representation, we employed a multi-modal deep-learning classification framework similar to the one in [9]. Tables I, II, III and IV show confusion matrices that measure the recognition performance for colour-only, colour and dense depth without fusion, colour and sparse depth without fusion and our XDimages respectively. The KITTI dataset contains instances of car, cyclist, pedestrian, van, person sitting, truck and misc. The misc category corresponds to objects with very low frequency of occurrence such as motorcycles or trolleys. In our experiments we considered

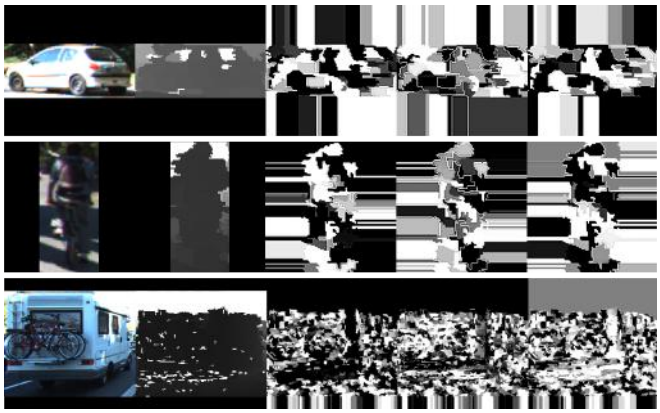

Fig. 3. Image and depth preprocessing. Depth maps are normalised and then converted into RGB by applying a jet colourmap. From left to to right: RGB image patch, depth map and three channels from the depth image rendered as in [9]

the classes *Car*, *Cyclist*, *Pedestrian*, *Truck* and *Van* as they are the most representative and we grouped the remaining classes into a general one that we labelled *Other*.

TABLE I

CONFUSION MATRIX FOR COLOUR-ONLY CLASSIFICATION (%)

| Act.\Pred. | Car | Cyclist | Pedestrian | Truck | Van | Other |
|---|---|---|---|---|---|---|
| **Car** | 96.1 | 0.1 | 0.3 | 0 | 0.6 | 2.9 |
| **Cyclist** | 10.1 | 58.2 | 24.1 | 0 | 0 | 7.6 |
| **Pedestrian** | 1.3 | 1.3 | 96 | 0 | 0.5 | 0.9 |
| **Truck** | 13 | 0 | 1.9 | 53.7 | 27.7 | 3.7 |
| **Van** | 43.7 | 0 | 2.1 | 2.1 | 47.9 | 4.2 |
| **Other** | 10.7 | 4.1 | 5.1 | 0.7 | 2.1 | 77.3 |

TABLE II

CONFUSION MATRIX FOR MULTI-MODAL CLASSIFICATION (%) USING IMAGE AND DENSE POINT CLOUD WITHOUT FUSION

| Act.\Pred. | Car | Cyclist | Pedestrian | Truck | Van | Other |
|---|---|---|---|---|---|---|
| **Car** | 96.3 | 0.1 | 0.5 | 0 | 0.6 | 2.5 |
| **Cyclist** | 5.1 | 69.3 | 14.1 | 0 | 0 | 11.5 |
| **Pedestrian** | 0.4 | 0.9 | 95.2 | 0.4 | 0.4 | 2.7 |
| **Truck** | 11.1 | 0 | 1.9 | 66.7 | 18.4 | 1.9 |
| **Van** | 23.2 | 0 | 2.1 | 1.4 | 67.0 | 6.3 |
| **Other** | 8.9 | 2.8 | 3.1 | 0.2 | 1 | 84 |

The confusion matrices in Tables I and III show the increase in recognition performance when colour and dense depth are fed to the classifier. The major source of uncertainty is the class *Other* which contains instances that in some cases share similarities with the classes car, cyclist or pedestrian. The third image row in figure 3 shows an instance of our class *other*.

Although our XDimages allow for multi-modal classification results which are better than the colour-only case, for the kitti dataset in particular, the recognition improvement due to our sensor fusion approach compared with using combined colour and velodyne (dense point cloud without depth densification) is negligible. The following hipotheses can be formulated from these results:

TABLE III

CONFUSION MATRIX FOR MULTI-MODAL CLASSIFICATION (%) USING SENSOR-FUSION OUTPUT

| Act.\Pred. | Car | Cyclist | Pedestrian | Truck | Van | Other |
|---|---|---|---|---|---|---|
| **Car** | 96.3 | 0.1 | 0.5 | 0 | 0.8 | 2.3 |
| **Cyclist** | 6.3 | 68.4 | 13.9 | 0 | 0 | 11.4 |
| **Pedestrian** | 0 | 0.4 | 96.1 | 0.4 | 0.4 | 2.7 |
| **Truck** | 11.1 | 0 | 1.9 | 66.7 | 18.4 | 1.9 |
| **Van** | 21.8 | 0 | 2.1 | 1.4 | 68.4 | 6.3 |
| **Other** | 8.6 | 3.1 | 3.1 | 0.3 | 1.5 | 83.4 |

TABLE IV

CONFUSION MATRIX FOR MULTI-MODAL CLASSIFICATION (%) USING COLOUR AND SPARSE POINT CLOUD WITHOUT FUSION

| Act.\Pred. | Car | Cyclist | Pedestrian | Truck | Van | Other |
|---|---|---|---|---|---|---|
| **Car** | 90.7 | 1 | 1.43 | 0.6 | 5.23 | 1.03 |
| **Cyclist** | 1.91 | 62.07 | 34.12 | 0.38 | 1.13 | 0.38 |
| **Pedestrian** | 0.34 | 9.46 | 90.09 | 0 | 0 | 0.11 |
| **Truck** | 8 | 0 | 2.29 | 64.39 | 22.1 | 3.21 |
| **Van** | 3.78 | 1.3 | 1.74 | 3.96 | 57.02 | 5.19 |
| **Other** | 13.64 | 7.73 | 2.86 | 3.11 | 8.5 | 64.16 |

- The denser the range data is, the closer the depth map obtained from the raw depth measurements gets to the actual object silhouette.
- The denser the range data is, the more sensitive our sensor fusion approach becomes to calibration errors. That is, more pixels belonging to the background are used to initialise our foreground models in the segmentation stage.

Therefore we have a fusion-based method that achieves an improvement over colour-only classification using cheap LIDAR sensors that is comparable with the improvements obtained by combining images and dense point clouds, which come from expensive laser sensors such as the velodyne.

## V. CONCLUSION

The work presented in this paper addresses the problem of extending colour images with sparse range data at the pixel level. To this end, we developed a framework for densifying sparse range data in the image space. Our framework provides a methodology for creating extended images independently of the density of the range sensor at hand. We adapted and combined two powerful segmentation techniques such as SLIC and Graph Cuts into a hyrarchical methodology for depth densification. The experimental results show the advantages of our new low level data representation over using colour only.

Our framework achieves, from a camera and sparse/cheap range sensors, recognition results that are equivalent to those obtained from a camera and dense/expensive velodyne.

Currently, our method assigns the same depth value to all pixels inside a particular super-pixel. It also neglects the dependency among neighbouring range measurements and their potential correlation with changes in colour. A possible future direction is to interpolate and extrapolate depth estimates using colour gradients. Additionally, the current energy function used in the grap-cut segmentation

stage includes colour statistics only as part of future work, we intend considering depth statistics as well.

## ACKNOWLEDGMENT

This work has been supported by Toyota Motor Europe and the French Institute for Research in Computer Science and Automation - Inria. The authors would like to thank Gabriel Othmezouri, David Sierra-Gonzalez, Jean-Alix David and Jerome Lusserau for their feedback and helpful discussions.

### *Author's note*

The first author undertook this work while he was part of INRIA.

## REFERENCES

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 34(11):2274–2281, 2012.

[2] G. Agamennoni, P. Furgale, and R. Siegwart. Self-tuning M-estimators. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 4628–4635, 2015.

[3] A. Blake, P. Kohli, and C. Rother. *Markov Random Fields for Vision and Image Processing*. The MIT Press Cambridge, 2011.

[4] G. Bradski. opencv library. *Dr. Dobb's Journal of Software Tools*, 2000.

[5] C. Cadena and J. Kosecka. Semantic Segmentation with Heterogeneous Sensor Coverages. *IEEE International Conference on Robotics and Automation (ICRA))*, pages 2639–2645, 2014.

[6] F. Castanedo. A review of data fusion techniques. *The Scientific World Journal (SWJ)*, 2013:19, 2013.

[7] H. Cho and Y.-w. Seo. A Multi-Sensor Fusion System for Moving Object Detection and Tracking in Urban Driving Environments. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 1836–1843, 2014.

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B: Statistical Methodology (JRSS)*, 39(1):1–38, 1977.

[9] A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller, and W. Burgard. Multimodal Deep Learning for Robust RGB-D Object Recognition. In *IEEE International Conference on Intelligent Robots and Systems (IROS)*, 2015.

[10] E. P. Fotiadis, M. Garz, and A. Barrientos. Human Detection from a Mobile Robot Using Fusion of Laser and Vision Information. *sensors*, pages 11603–11635, 2013.

[11] M. P. Gerardo-Castro, T. Peynot, F. Ramos, and R. Fitch. Non-Parametric Consistency Test for Multiple-Sensing-Modality Data Fusion. In *IEEE International Conference on Information Fusion (FUSION)*, pages 443–451, 2015.

[12] R. Kaestner, J. Maye, Y. Pilat, and R. Siegwart. Generative object detection and tracking in 3D range data. *IEEE International Conference on Robotics and Automation (ICRA)*, pages 3075–3081, may 2012.

[13] C. Kerl, J. Sturm, and D. Cremers. Robust Odometry Estimation for RGB-D Cameras. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 3748–3754, 2013.

[14] B. Khaleghi, A. Khamis, F. O. Karray, and S. N. Razavi. Multisensor data fusion: A review of the state-of-the-art. *Information Fusion*, 14(1):28–44, 2013.

[15] R. C. Luo, C.-C. Yih, and K. L. Su. Multisensor fusion and integration: approaches, applications, and future research directions. *IEEE Sensors Journal*, 2(2):107–119, 2002.

[16] A. Negre, L. Rummelhard, and C. Laugier. Hybrid Sampling Bayesian Occupancy Filter. *IEEE Intelligent Vehicles Symposium, Proceedings*, pages 1307–1312, 2014.

[17] P. Pinies, L. M. Paz, and P. Newman. Too Much TV is Bad: Dense Reconstruction from Sparse Laser with Non-convex Regularisation. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2015.

[18] C. Premebida and U. Nunes. Fusing LIDAR, camera and semantic information: A context-based approach for pedestrian detection. *The International Journal of Robotics Research (IJRR)*, 32(3):371–384, jan 2013.

[19] N. D. Reddy, P. Singhal, V. Chari, and K. M. Krishna. Dynamic body VSLAM with semantic constraints. *IEEE International Conference on Intelligent Robots and Systems (IROS)*, pages 1897–1904, 2015.

[20] C. Y. Ren, V. A. Prisacariu, and I. D. Reid. gSLICr: SLIC superpixels at over 250Hz. *arXiv:1509.04232*, pages 1–6, 2015.

[21] V. Romero-Cano, G. Agamennoni, and J. Nieto. A Variational Approach to Simultaneous Tracking and Classification of Multiple Objects. In *International Conference on Information Fusion (FUSION)*, pages 1 – 8, 2014.

[22] V. Romero-Cano, G. Agamennoni, and J. Nieto. A variational approach to simultaneous multi-object tracking and classification. *The International Journal of Robotics Research (IJRR)*, 35(6):654–671, 2015.

[23] V. Romero-Cano and J. I. Nieto. Stereo-based Motion Detection and Tracking from a Moving Platform. In *Intelligent Vehicles Symposium*, pages 499–504, 2013.

[24] C. Rother, V. Kolmogorov, Y. Boykov, and A. Blake. Interactive Foreground Extraction using graph cut. Technical report, Microsoft, 2011.

[25] J. Schlosser, C. K. Chow, and Z. Kira. Fusing LIDAR and Images for Pedestrian Detection using Convolutional Neural Networks*. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 2198–2205, 2016.

[26] L. Spinello and K. O. Arras. Leveraging RGB-D Data: Adaptive Fusion and Domain Adaptation for Object Detection. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 4469–4474, 2012.

[27] L. Spinello, R. Triebel, and R. Siegwart. Multiclass Multimodal Detection and Tracking in Urban Environments. *The International Journal of Robotics Research (IJRR)*, 29(2):1498–1515, 2010.

[28] M. Tanner, P. Piniés, L. Paz, and P. Newman. What Lies Behind: Recovering Hidden Shape in Dense Mapping. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 979–986, 2016.

[29] Z. Taylor and J. Nieto. Motion-Based Calibration of Multimodal Sensor Arrays. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 4843–4850, 2015.

[30] X. Wang, M. Yang, S. Zhu, and Y. Lin. Regionlets for generic object detection. *International Conference on Computer Vision (ICCV)*, 2013.

[31] R. Zhang, S. A. Candra, K. Vetter, and A. Zakhor. Sensor Fusion for Semantic Segmentation of Urban Scenes. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 1850–1857, 2015.