MULTIMODAL IMAGING SYSTEM BASED ON SOLID-STATE LIDAR FOR ADVANCED PERCEPTION APPLICATIONS

PALAIS DES CONGRES, VERSAILLES, FRANCE | 01 – 03 FEBRUARY 2022

Pablo GARCÍA-GÓMEZ [(1)], Noel RODRIGO [(1)], Jordi RIU [(2)], Josep R. CASAS [(3)], Santiago ROYO [(1)]

[(1)] Centre de Desenvolupament de Sensors, Instrumentació i Sistemes, Universitat Politècnica de Catalunya (UPC-CD6), Rambla Sant Nebridi 10, 08222, Terrassa, Spain.
[(2)] Beamagine S.L., Carrer de Bellesguard 16, 08755 Castellbisbal, Spain.
[(3)] Grup de Processat d'Imatge, Departament TSC, Universitat Politècnica de Catalunya (UPC), Carrer de Jordi Girona 1-3, 08034 Barcelona, Spain.
Correspondence: pablo.garcia.gomez@upc.edu

KEYWORDS: sensor fusion, enhanced perception, solid-state LiDAR, autonomous navigation, robotics

ABSTRACT:

Perception of the environment is a crucial requirement for cutting-edge advanced applications. Light Detection And Ranging (LiDAR) systems are a clear example of this market-pull disruptive innovation. LiDAR has evolved towards imaging systems of great interest due to their 3D sensing capability with higher spatial resolution than radars.

Nonetheless, these applications claim large amounts of complementary data from the environment, even redundant, for making reliable decisions about the most adequate response according to the environment's perception. Consequently, multiple sensors of different natures, working principles and failure modes must be wisely combined. This is known as sensor fusion.

Hence, we present a multimodal imaging system, consisting of a pulsed solid-state LiDAR combined with three other additional imaging sensors, that provides multimodal information with low parallax fusion error thanks to the accurate high-density measurements from the LiDAR system.

Thus, multimodal information can feed perception algorithms for enhancing their performances as well as accurately combining their outputs.

1. INTRODUCTION

The development of both Artificial Intelligence (AI) and Computer Vision (CV) plus the wide variety of available sensors have bestowed artificial systems like vehicles and robots on advanced perception about the environment, mimicking human perception and pursuing their autonomy [1].

The larger the information about its environment, the more robust and reliable decisions an Autonomous Vehicle (AV) may make.

In response, data fusion has the aim of increasing the amount of data through adequately and efficiently combining the information from different sensors. The variety of sensors must include different working principles and failure modes [2].

Concerning imaging sensors, perception algorithms have evolved towards combining images (2D) and depth information (3D) for enhancing their performance [3-5]. Consequently, Light Detection And Ranging (LiDAR) systems have arisen great interest due to their precise depth information [6].

LiDAR devices calculate the elapsed time between two events on light, known as Time of Flight (TOF), and directly relates to the distance to a target. Although based on the same principle as radars, using shorter wavelengths than radio ones provides LiDAR systems with higher spatial resolutions. Consequently, imaging LiDARs offer 3D representations of the environment known as Point Clouds which are sets of points given by 3D coordinates.

Fig. 1 shows a Point Cloud of a scenario whose perception is enhanced using different colourmaps and finally data fusion. From left to right, firstly all points present the same arbitrary colour. Secondly, each point is coloured according to its depth value. Thirdly, the colour corresponds to the measured light intensity at the laser's wavelength. Finally, the colour from a conventional colour camera is combined on the Point Cloud. Thanks to the last case, we can easily interpret the environment, which is the main aim of data fusion.

In this paper, we are presenting and describing a multimodal fusion system. Moreover, we will compare its fusion accuracy to other state-of-the-art fusion systems like [7].

This paper is organized into different sections, starting with a theoretical background in Section 2. Then, we explain the materials and methods in Section 3 and the main results in Section 4. Finally, a discussion follows in Section 5.

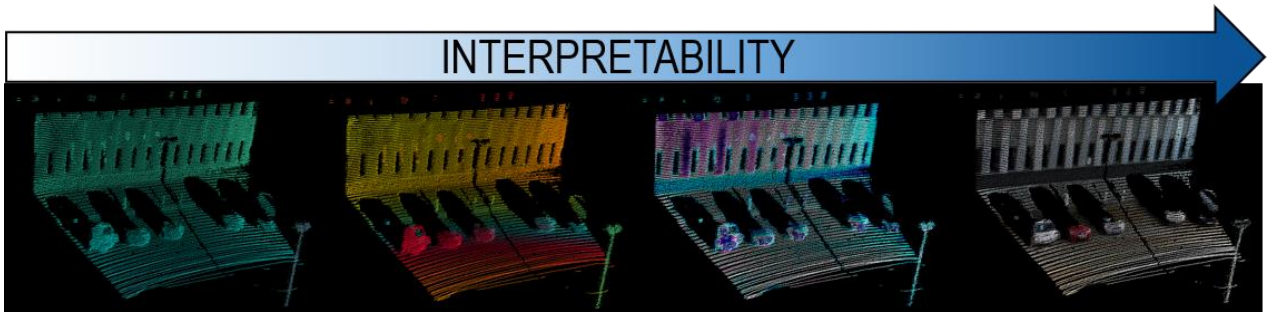

Figure 1. Perception enhancement of a Point Cloud thanks to data fusion. From left to right, the colour assigned to the Point Cloud is based on an arbitrary colour, depth, laser's intensity and a colour camera.

2. THEORETICAL BACKGROUND

Data fusion properly combines several sensors with different working principles for describing the whole environment from each particular view. Hence, each sensor's information must be adequately managed to compose the global framework which is the 3D space.

This data management is divided into two steps. Firstly, the sensor's information must be translated into their particular view of the 3D space. This step is described by the Intrinsics of the sensor. Afterwards, the Extrinsics of the sensor, which describe its position and orientation from the environment, place the subset of the sensor within the global space. Consequently, we can build a more complex and complete vision of the environment. Notice that the analogue procedure enables estimating a sensor's output.

Let us first describe the Extrinsics of a sensor and later on the different Intrinsics present in the system.

2.1. Extrinsics of a sensor

Algebraically, the previous data management in 3D coordinates is described using rigid transformations that are geometric transformations, isometries, of a Euclidean space that preserve the Euclidean distance between points. Moreover, proper rigid transformations not only keep distance but also the handedness (left- or right-hand system), size and shape. Generally, they are described as a rotation followed by a translation. For example, the special Euclidean group in the 3D Euclidean space denoted as $SE(3)$ describe rigid body displacements in kinematics.

Let us consider a single sensor from a system of N sensors observing the environment E which is the 3D Euclidean space. Hence, the sensor perceives information from a subset S^n defined by its Field-of-View (FOV) and of equal dimensions of E such that $S^n \subset E$, where the superscript $n = 1, \dots, N$ represents the particular sensor. Thus, the sensor is represented as the basis for the subspace and describes its particular coordinate system.

From now on, we use the ISO convention describing scalars with lowercase roman letters a and vectors with bold ones $\boldsymbol{a}$, subset spaces with uppercase roman letters A and matrices with bold ones $\boldsymbol{A}$, and angular magnitudes with Greek letters. In addition, we use row vectors.

Eq.1 $$S^n \stackrel{\text{def}}{\equiv} \{ {}^n\widehat{s_1}, {}^n\widehat{s_2}, {}^n\widehat{s_3} \}$$

Where ${}^n\widehat{s_k}$ is a unitary vector that represents the *kth*-axis.

Consequently, a sensor is a 6 Degrees-of-Freedom (DOF) coordinate system, defined by its position or location (with coordinates *x*, *y* and *z* – 3 DOF) plus its orientation (with rotation angles for each axis – 3 DOF). Hence, any 3D point $\boldsymbol{p}$ in E (${}^E\boldsymbol{p} \coloneqq {}^{\boldsymbol{E}}[a, b, c] \in E$) can be described in S^n using the above-mentioned rigid transformations as follows:

Eq.2 $$ {}^{S^n}\boldsymbol{p} = {}^E\boldsymbol{p} \cdot \boldsymbol{R}_{ES^n} + \boldsymbol{t}_{ES^n} = {}^{\boldsymbol{S^n}}[u, v, w] \in S^n$$

From the above formula, $\boldsymbol{R}_{ES^n}$ and $\boldsymbol{t}_{ES^n}$ are the rotation matrix and the translation vector from E to S^n respectively.

On the one hand, each row of $\boldsymbol{R}_{ES^n}$ is each axis of E expressed in S^n so this matrix defines the orientation of E from S^n. This is the reason why they are also known as orientation matrices.

Eq.3 $$\boldsymbol{O}_{E|S^n} \equiv \boldsymbol{R}_{ES^n} = \begin{bmatrix} \boldsymbol{S^n}\left({}^E\widehat{s_1}\right) \\ \boldsymbol{S^n}\left({}^E\widehat{s_2}\right) \\ \boldsymbol{S^n}\left({}^E\widehat{s_3}\right) \end{bmatrix}$$

On the other hand, and analogous to the orientation's definition, the translation vector $\boldsymbol{t}_{ES^n}$ defines the location of E in S^n because it relates both systems' origins $\boldsymbol{o}_E$ and $\boldsymbol{o}_{S^n}$ in S^n.

Eq.4 $$\mathbf{loc}_{E|S^n} \equiv \boldsymbol{t}_{ES^n} = {}^{S^n}\boldsymbol{o}_E - {}^{S^n}\boldsymbol{o}_{S^n} = {}^{S^n}\boldsymbol{o}_E$$

Notice that ${}^{S^n}\boldsymbol{o}_{S^n}$ is the null vector because it is the origin of S^n expressed in the same S^n.

The above Eq. 2 can be reduced to a simple matrix operation after extending $\boldsymbol{p}$ to a 4-dimensional vector using homogeneous coordinates, yielding the definition of the matrix $\boldsymbol{H}_{ES^n}$.

Eq.5 $$ {}^{S^n}\boldsymbol{p} = \left[{}^{E}\boldsymbol{p}, 1\right]\begin{bmatrix}\boldsymbol{O}_{E|S^n} \\ \mathbf{loc}_{E|S^n}\end{bmatrix} \coloneqq \left[{}^{E}\boldsymbol{p}, 1\right]\cdot\boldsymbol{H}_{ES^n} $$

This matrix is usually expanded and called the homogeneous matrix of the transformation $\widehat{\boldsymbol{H}}_{ES^n}$:

Eq.6 $$ \widehat{\boldsymbol{H}}_{ES^n} \overset{\text{def}}{=} \begin{bmatrix}\boldsymbol{H}_{ES^n} & \begin{matrix}0\\0\\0\end{matrix} \\ \begin{matrix}1\end{matrix}\end{bmatrix} = \begin{bmatrix}\boldsymbol{O}_{E|S^n} & 0 \\ \mathbf{loc}_{E|S^n} & 0 \\ & 1\end{bmatrix} $$

Provided that rotation matrices are orthonormal, thus $\boldsymbol{R}^T\boldsymbol{R} = \boldsymbol{R}^{-1}\boldsymbol{R} = \boldsymbol{I}$, the inverse transformation $\widehat{\boldsymbol{H}}_{S^nE}$ is easily obtained transposing the orientation matrix and applying Eq. 2 on Eq. 4:

Eq.7 $$ \begin{aligned}\boldsymbol{O}_{S^n|E} &= \left(\boldsymbol{O}_{E|S^n}\right)^{\boldsymbol{T}} \\ \mathbf{loc}_{S^n|E} &= -\mathbf{loc}_{E|S^n}\left(\boldsymbol{O}_{E|S^n}\right)^{\boldsymbol{T}}\end{aligned} $$

To summarize, the Extrinsics of a sensor relate the location and orientation of the sensor with respect to the environment through its location vector $\mathbf{loc}_{S^n|E}$ and its orientation matrix $\boldsymbol{O}_{S^n|E}$. Nevertheless, these expressions and definitions are also valid for any pair of sensors, enabling the data transformation between them as well.

Hence, let a pair of sensors S^n and S^j be a data fusion system observing a common environment. Then, Fig. 2 schematizes their data fusion procedure explained in the introduction of this section using the above definitions of the Extrinsics.

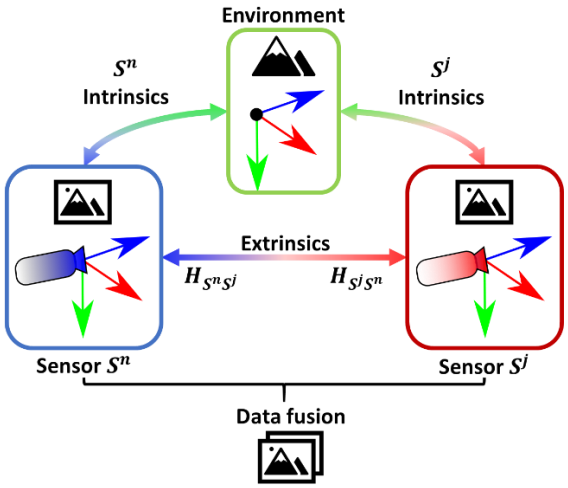

Figure 2. Scheme of data fusion between two sensors.

2.2. Intrinsics of the sensors

Once we have defined the Extrinsics that relate sensors with the environment they sense and between them, let us understand how they measure in order to define their Intrinsic parameters.

Given that our data fusion system consists of a LiDAR for providing reliable 3D information of high spatial resolution combined with different cameras, we are going to first derive the Intrinsics for the LiDAR system and, secondly, for a general camera.

2.2.1. LiDAR

As previously commented during the introduction, imaging LiDARs are based on measuring the TOF for a number of points within their FOV, yielding Point Cloud as their output. Generally, any point in the LiDAR's Point Cloud ${}^{L}\boldsymbol{p}$ can be expressed as the combination of the depth measurement resolved from the TOF, t_{TOF}, and a direction in the space, ${}^{L}\hat{\boldsymbol{s}}$.

Eq.8 $$ {}^{L}_{k}\boldsymbol{p} = {}^{L}_{k}[x, y, z] = \frac{c}{2}\,{}_{k}t_{TOF}\cdot{}^{L}_{k}\hat{\boldsymbol{s}} $$

Where c is the speed of light in air and the subscript k applies for the *kth*-point in the Point Cloud. Regarding that t_{TOF} is a radial measurement, spherical coordinates are useful for describing the direction vector ${}^{L}\hat{\boldsymbol{s}}$.

This subdivision of the 3D measurement brings researchers to split imaging LiDARs into the receiving subsystem, responsible for measuring ${}_{k}t_{TOF}$ for each point, and the imaging system that scans the FOV sectioning it in different directions ${}^{L}_{k}\hat{\boldsymbol{s}}$ and forms the Point Cloud [6].

Notice that the precision and accuracy of the points depend on properly determining both measurements. In this work, we are going to focus on the imaging subsystem which determines the spatial resolution of the system.

Ideally, consecutive points must be evenly spaced meaning that directions ${}^{L}_{k}\hat{\boldsymbol{s}}$ and ${}_{k+1}{}^{L}\hat{\boldsymbol{s}}$ should present a constant spacing across the FOV of the system. Nonetheless, the spatial resolution may vary yielding distortion in the Point Cloud similar to the camera's distortion [8,9] regardless of the imaging technique.

Up to now, researchers have mainly used mechanical scanning LiDARs so the available models and calibrations can only describe them [10,11], lacking generality with respect to other imaging techniques such as solid-state ones. Moreover, the novelty of LiDARs compels many manufacturers to protect their Intellectual Property (IP) and do not offer raw measurements so calibrations usually correct the Point Cloud through post-processing.

Consequently, we formulated and published an imaging model and a calibration procedure in [12] valid for Micro-Electro-Mechanical Systems (MEMS) mirror scanning techniques, which is the imaging technique of our LiDAR devices, but also

feasible for other imaging techniques capable of precisely and accurately predicting the variable angular resolution across the FOV from raw measurements. Hence, using this model yields accurate and reliable Point Clouds free of distortion.

The model consists of using non-linear 2D maps as proposed in [13-15]. Provided that the MEMS scan offers an ordered arrangement of the measurements, the mapping functions $\boldsymbol{f}$ relate each acquisition coordinate or position in a matrix ${}_k(u,v)$ with its corresponding scanning direction ${}^L_k\hat{\boldsymbol{s}}$ defined by two spherical angles ${}_k(\theta_H, \theta_V)$.

Eq.9
$$\boldsymbol{f}: (u,v) \xrightarrow{\vec{\varphi}} (\theta_H, \theta_V)$$

The set of parameters $\vec{\varphi}$ of the mapping function defines the Intrinsics of the LiDAR. Using spherical coordinates together with the defined scanning angles, Eq. 2 can be expressed as follows:

Eq.10
$${}^L_k\boldsymbol{p} = \frac{c}{2}\,{}_kt_{TOF}\begin{bmatrix}\sin({}_k\theta)\cos({}_k\phi)\\ \sin({}_k\theta)\sin({}_k\phi)\\ \cos({}_k\theta)\end{bmatrix}^T$$

Where the spherical angles ${}_k\theta$ and ${}_k\phi$ relate to the scanning angles ${}_k\theta_H$ and ${}_k\theta_V$ as follows:

Eq.11
$${}_k\theta = \arctan\left(\sqrt{\tan({}_k\theta_H)^2 + \tan({}_k\theta_V)^2}\right)$$
$${}_k\phi = \arctan\left(\frac{\tan({}_k\theta_V)}{\tan({}_k\theta_H)}\right)$$

Thanks to these expressions and the calibration procedure of [12], the variation of the angular resolution can be corrected yielding lateral errors in the Point Cloud of less than 5cm at a range of 100m, so errors below 15 millidegrees.

2.2.2. Cameras

Contrary to LiDAR systems, cameras have been studied since the beginnings of CV [16]. They project the 3D space onto a 2D image plane. This projection is described using the pinhole camera model [17] and generalized with the thin lens approximation. It must be noticed that CV's convention places the image plane in front of the focal point, as shown in Fig. 3 contrarily to the optic's convention.

In such a way, a 3D point in the FOV of the camera ${}^C\boldsymbol{p}$ projects onto a pixel position on the sensor $[u,v]$ through a matrix $\boldsymbol{K}_C$ known as the Intrinsic matrix.

Eq.12
$$\lambda\begin{bmatrix}u\\ v\\ 1\end{bmatrix}^T = {}^C\boldsymbol{p}\cdot\begin{bmatrix}f_x & 0 & 0\\ s & f_y & 0\\ c_x & c_y & 1\end{bmatrix} = {}^C\boldsymbol{p}\cdot\boldsymbol{K_C}$$

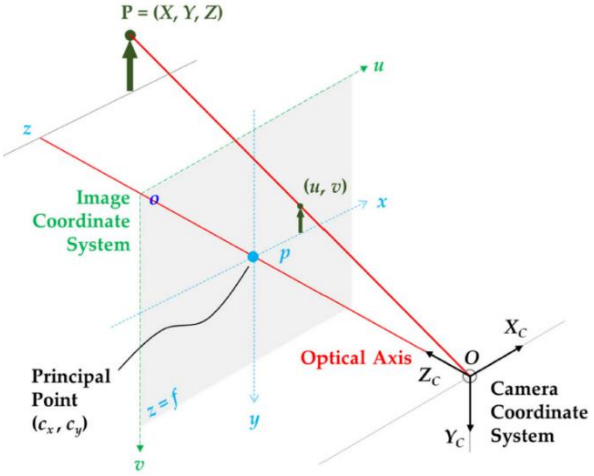

Figure 3. Diagram of the pinhole camera model [5].

Where λ is an arbitrary constant that expresses the depth information loss due to the projection. Then, $\boldsymbol{K}_C$ contains the optical information of the camera. In particular, f_x and f_y are the focal length in pixel units in both horizontal and vertical directions of the sensor respectively, s is the skew parameter related to the perpendicular condition of the two dimensions and c_x and c_y correspond to the projection of the optical axis on the image, known as the principal point.

As one may notice, Eq. 12 is a linear operation that assumes no-distortion effects. Optical distortions are included afterwards as the combination of radial and tangential non-linear terms [8,18] that are defined with up to 5 parameters.

These parameters represent the Intrinsics of a camera and are generally determined using the calibration method known after Bouguet [19], who implemented and published the theoretical developments of [20] and [21].

3. MATERIALS AND METHODS

After presenting the two working models for the sensors in our data fusion system, let us now present the sensors used in this work.

As previously mentioned, we base our data fusion system on a pulsed MEMS solid-state LiDAR (L). We used a pre-commercial unit courtesy of Beamagine SL with a high-density of points (>45k points) per frame within a 30ºx20º FOV, similar to one from a camera, achieving designed angular resolutions of 0.1ºx0.13º. Its operating range is up to 150 m at 10 frames per second (FPS).

Complementarily, we used three different camera sensors for acquiring information from either different parts of the spectrum (visible and long-wave infrared – LWIR) or properties of light (intensity or polarization state).

Thus, we use a conventional colour camera (R, 400 – 700 nm), a linear polarization one (P, 320 – 1000 nm) and an LWIR thermal one (T, 8 – 14 μm).

Below, Tab. 1 and Fig. 4 present the multimodal data fusion system, specifying the different sensors combined and which type of information provides to the global system. Notice that the largest dimension of the system is 20 cm (a hand span).

Table 1. List of sensors in the multimodal fusion system.

Sensor		FOV (º) and Size (pix)	Measure
L	L3CAM	30º x 20º 300 x 150	3D (1064 nm)
R	PHX050S-CC	30.9º x 23.2º 1224 x 1024	Colour
P	PHX050S-PC	30.9º x 23.2º 1224 x 1024	Linear Polarization
T	OPTRIS Xi 400	29º x 22º 382 x 288	Temperature

Figure 4. Multimodal fusion system based on LiDAR courtesy of Beamagine SL.

Notice that all sensors present similar FOVs of around 30ºx20º for sharing as much as possible the information from a common FOV. Moreover, their apertures, thus their physical locations, are situated close to each other, which helps reduce the dimensions of the enclosure and the parallax error, since occlusions are produced by the same source objects.

Following [7], we decide to connect all sensors between them according to the Fully Connected Pose Estimation (FCPE) because this configuration adds the loop-closure constraint. However, we always consider the LiDAR as the principal sensor for improving the data fusion between any pair of sensors because it provides reliable and precise 3D geometry information: accurate depth information and lateral measurements thanks to the above-discussed Intrinsics model.

We correlated the sensors of the system, thus calibrated their Extrinsics, using [18] with a multimodal planar checkerboard pattern and 14 captures from different locations and orientations.

4. RESULTS

In this section, we are presenting some multimodal fusion examples obtained with the presented data fusion system.

As briefly discussed in the previous section, the parallax error of the data fusion, defined as the mismatch between known 3D data points after applying Eq. 2 between pairs of sensors, is effectively improved compared to the literature thanks to the reduced dimensions of the system and the precise and accurate 3D information provided by the LiDAR sensor.

Tab. 2 below presents the parallax error in cm for all possible pairs of sensors in the multimodal data fusion system after calibrating their Extrinsics. The parallax error is calculated as the average error for all the known 3D points back-projected onto the calibration's pattern plane (which is precisely and accurately provided by the LiDAR system) for all the 14 calibration captures. Hence, we are directly measuring the lateral error of the fusion in the 3D space. Additionally, we compute the average parallax error of the system considering all the sensor pairs.

Table 2. Parallax error for all possible pairs of sensors in the multimodal data fusion system in mm.

L-R	L-P	L-T	R-P	R-T	P-T
3.6	4.3	5.8	1.5	4.5	4.8
Average error			4.1		

For instance, comparing our results with the ones reported in [7], we achieve an average parallax error of 4.1 mm with only 14 captures, which is one order of magnitude below the cm range reported in the literature. We must emphasize that this is mainly thanks to the precise and accurate 3D information that LiDAR provides to the system, that can be used for reliably estimating the pose of the calibration target and, as a consequence, the back-projection of cameras onto it.

Let us now present some fused outputs, Point Clouds and images, from our proposed multimodal data fusion system.

4.1. Colour fusion

The following examples of data fusion between the LiDAR and the colour camera are coloured Point Clouds from urban scenarios ranging from few tenths of meters up to hundred meters.

Thanks to the fused colour information, Point Clouds are more interpretable, improving human perception for instance. This can be easily observed in Fig. 5, where the FOV is divided into colour (left) and intensity at the LiDAR's laser wavelength (right).

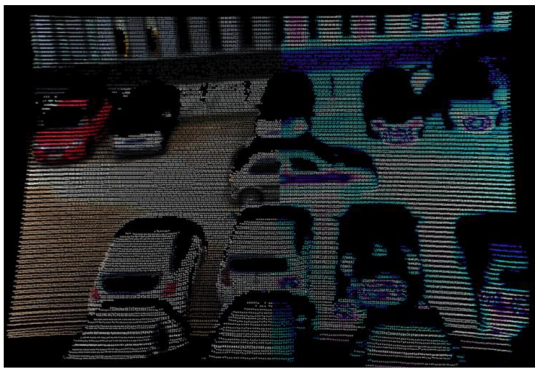

Figure 5. Fused Point Cloud up to 30 m range where the left-hand side contains colour information from the colour camera and the right-hand side, the intensity at the LiDAR's laser wavelength (1064 nm).

Clearly, colour information helps to understand the car park scenario and even assists in identifying the different cars per their colour.

The following Fig. 6 and Fig. 7 demonstrate the low parallax fusion error because it can be observed the accuracy of the fusion up to a hundred meters and with multiple objects within the FOV.

Figure 6. Fused Point Cloud with colour information up to 40 m range and its corresponding image.

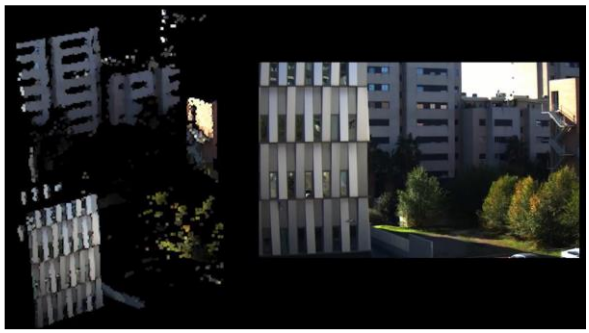

Figure 7. Fused Point Cloud with colour information up to 120 m range and its corresponding image.

4.2. Polarimetric fusion

Secondly, we present some Point Clouds fused with the polarimetric information which does not measure light's intensity but its polarization state, which is the vector property of light. The polarimetric images here presented show the Degree of Linear Polarization (DoLP) that indicates how much linearly polarized is the incoming light from objects. Hence, DoLP images differentiate between those objects that keep or polarize light into linear polarization states (metals with flat surfaces) and those that depolarize (diffusive rough objects).

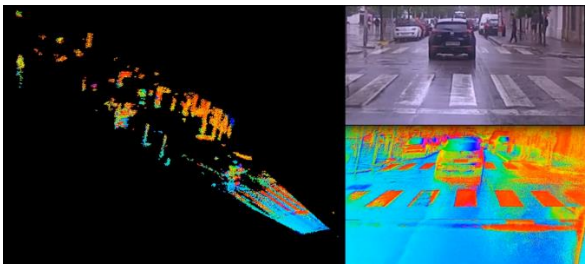

Figure 8. Fused Point Cloud with polarimetric information of an urban scene in rainy conditions.

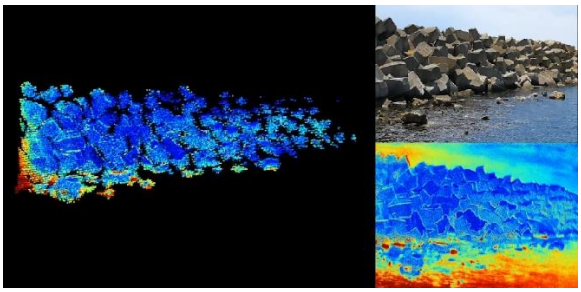

Figure 9. Fused Point Cloud with polarimetric information of a harbour scene.

For instance, Fig. 8 represents a Point Cloud from an urban scene in rainy conditions and Fig. 9, one from a maritime scene inside a harbour. Thanks to Brewster's angle (angle of incidence at which incident unpolarized light is perfectly polarized after reflecting from a surface), both wet painted pedestrian crossings and seawater polarize light into a linear polarization state.

Consequently, they present reddish colours in the polarimetric image representing high DoLP values whereas the pavement and salient objects from the sea like breakwater's rocks depolarize light. Thus, DoLP offers additional contrast that can be used for improving perception in maritime and under adverse weather conditions.

Not only that, we must consider a key advantage of using LiDAR devices in this kind of scenarios. Since colour and polarimetric cameras are passive sensors, they need illuminators or ambient light for acquiring images. However, LiDAR performs active measurements so, it performs better with no background illumination in fact. Hence, under poor illumination conditions, both colour and polarimetric cameras might improperly work whereas LiDAR will still offer reliable 3D information of the environment. This enhances the importance of having multiple sensors of different natures for avoiding unique common failure modes.

4.3. Thermal fusion

Lastly, we present some results of combining the LWIR thermal camera with the LiDAR and the other sensors.

Both Fig. 10 and Fig. 11 present Point Clouds from both indoor and outdoor scenes, demonstrating the accuracy of the fusion for all sensors.

*Figure 10. Multimodal fusion between a LiDAR and different cameras of an outdoor scene. (**a**) Original Point Cloud coloured with depth. (**b**) Colour and (**c**) thermal image with their corresponding fused Point Clouds.*

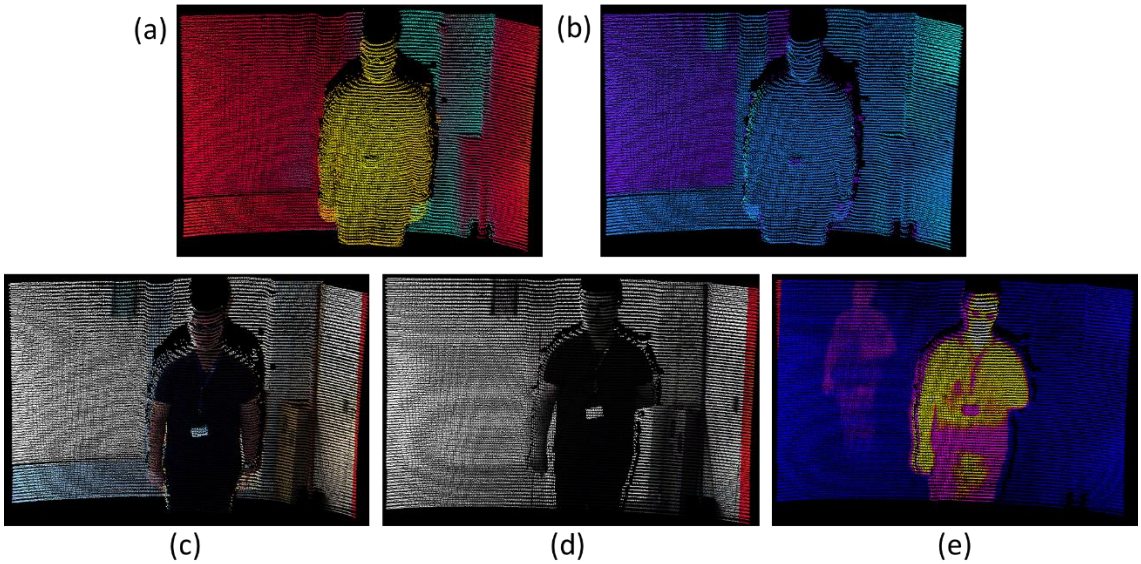

*Figure 11. Multimodal fusion Point Clouds of an indoor scene with a person. The displayed colour-maps are (**a**) depth, (**b**) intensity at the LiDAR's laser wavelength, (**c**) colour, (**d**) mono-polarimetric and (**e**) temperature.*

Including thermal cameras in our multimodal data fusion system enables us to distinguish objects from the environment depending on their temperature. Moreover, notice that thermal cameras, even though being passive sensors, offer more robustness to background illumination since temperature changes are slow and LWIR wavelengths radiate from hot objects such as people. Consequently, LWIR cameras can properly work under adverse weather conditions and even at night like LiDARs, as we previously discussed.

Moreover, thanks to the complementarity of the information, we can fuse multiple images in a single Point Cloud and even directly register (fuse 2D images) two images based on the 3D information provided by the LiDAR as Fig. 12 presents.

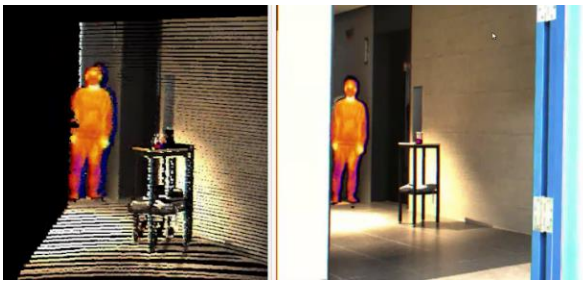

Figure 12. Fused Point Cloud with colour and thermal information and the corresponding colour image registered with the thermal image for detecting hot objects like a pedestrian and a tea cup.

Image registration based on the 3D information from the LiDAR is useful for reducing perception's computational cost since 2D processing in CV is more robust and efficient than 3D processing currently. This suggests we can combine 2D perceptions from an image onto another one. Moreover, this also permits registering images from unshared FOVs for creating panoramic views with high accuracy as long as the LiDAR information is used as a nexus.

5. DISCUSSION AND CONCLUSIONS

Our findings suggest that combining data from different sensors, either 2D or 3D, based on the depth information from a high-density LiDAR system improves the accuracy of the data fusion.

We find that having a high-density Point Cloud together with a reduced mechanical embodiment yields shorter parallax fusion errors than the reported ones in the existing literature. Moreover, this work also suggests that multimodal fusion devices enhance the perception of the environment by means of providing complementary data from different modes, either prior to processing or at later stages, that avoid unique failure modes.

The main limitation of this work is the lack of a well-established corpus about 3D perception techniques, which is currently not as mature as the 2D one. Consequently, 3D data is just used for properly doing data fusion although future developments in the field of 3D perception might provide more robustness and features extraction to the global perception. This is a current challenge in Deep Learning.

To conclude, we have presented a multimodal fusion system with 3D information combined with colour, polarimetric, Near-Infrared and thermal LWIR information. Furthermore, this work has demonstrated that the obtained accuracy of the data fusion is greater than most of the reported in the literature nowadays thanks to the high-density Point Clouds and the reduced enclosure of the system.

Additionally, we have presented several examples of multimodal data fusion of different imaging modes with unshared failure modes together with some potential applications like AV and security. Moreover, we have demonstrated that the LiDAR 3D information is essential for combining 3D+2D and even 2D+2D data with accuracy and robustness.

6. REFERENCES

1. Szeliski, R. (2010). Computer Vision: Algorithms and Applications. *Springer Science & Business Media.*

2. Wang, Z., Wu, Y. & Niu, Q. (2020). Multi-Sensor Fusion in Automated Driving: A Survey. *IEEE Access.* **8**, 2847-2868.

3. Rosique, F., Navarro, P.J., Fernández, C. & Padilla, A. (2019) A Systematic Review of Perception System and Simulators for Autonomous Vehicles Research. *Sensors*, **19**(3), 648.

4. Shahian Jahromi, B., Tulabandhula, T. & Cetin, S. (2019) Real-Time Hybrid Multi-Sensor Fusion Framework for Perception in Autonomous Vehicles. *Sensors*, **19**(20), 4537.

5. Yeong, D.J., Velasco-Hernandez, G., Barry, J. & Walsh, J. (2021) Sensor and Sensor Fusion Technology in Autonomous Vehicles: A Review. *Sensors*, **21**(6), 2140.

6. Royo, S. & Ballesta, M. (2019) An Overview of Lidar Imaging Systems for Autonomous Vehicles. *Applied Sciences*, **9**(19)

7. Domhof, J., Kooij, J.F.P. & Gavrila, D.M. (2019) An Extrinsic Calibration Tool for Radar, Camera and Lidar. In Proc. 2019 ICRA, IEEE, Montreal, Canada.

8. Brown, D.C. (1996) Decentering Distortion of Lenses. *Photogrammetric Engineering*, **32**, 442-462.

9. Ricolfe-Viala, C. & Sanchez-Salmeron, A.-J. (2010) Lens Distortion Models Evaluation. *Applied Optics*, **49**(30), 5914-5928.

10. Glennie, C. & Derek, D.L. (2010) Static Calibration and Analysis of the Velodyne HDL-64E S2 for High Accuracy Mobile Scanning. *Remote Sensing*, **2**(6), 1610-1624.

11. Yu, C., Chen, X. & Xi, J. (2017) Modeling and Calibration of a Novel One-Mirror Galvanometric Laser Scanner. *Sensors*, **17**(1), 164.

12. García-Gómez, P., Royo, S., Rodrigo, N. & Casas, J.R. (2020) Geometric Model and Calibration Method for a Solid-State LiDAR. *Sensors*, **20**(10), 2898.

13. Bauer, A., Vo, S., Parkins, K. Rodriguez, F., Cakmakci, O. & Rolland, J.P. (2012) Computational optical distortion correction using a radial basis function-based mapping method. *Optics Express*, **20**(14), 14906-14920.

14. An, L., Wu, Y., Xinxing, X. Huang, Y., Feng, C. & Zheng, Z. (2015) Computational method for correcting complex optical distortion based on FOV division. *Applied Optics*, **54**(9), 2441-2449.

15. Su, C., Guo, X., Wang, P. & Zhang, B. (2017) Computational optical distortion correction based on local polynomial by inverse model. *Optik*, **132**, 388-400.

16. Thomas, B., & Granum, E. (2001) A Survey of Computer Vision-Based Human Motion Capture. *Computer Vision and Image Understanding*, **81**(3), 231-268.

17. Young, M. (1971) Pinhole Optics. *Applied Optics*, **10**(12), 2763-2767.

18. Devernay, F. & Faugueras, O. (2001) Straight lines have to be straight. *Machine Vision and Applications*, **13**(1), 14-24.

19. Bouguet, J.-Y. & Perona, P. (1998) Camera Calibration from Points and Lines in Dual-Space Geometry. pp. 18.

20. Heikkila, J., & Silven, O. (1997) A four-step camera calibration procedure with implicit image correction. In Proc. CVPR., IEEE Computer Society, San Juan, Puerto Rico.

21. Zhang, Z. (1999) Flexible camera calibration by viewing a plane from unknown orientations. In Proc. 7th ICCV., IEEE, Kerkyra, Greece.