IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received November 28, 2019, accepted December 15, 2019, date of publication December 19, 2019,
date of current version December 31, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2960798

Refining the Fusion of Pepper Robot and Estimated Depth Maps Method for Improved 3D Perception

ZURIA BAUER, FELIX ESCALONA, EDMANUEL CRUZ, MIGUEL CAZORLA, AND FRANCISCO GOMEZ-DONOSO
Institute for Computer Research, University of Alicante, 03080 Alicante, Spain

Corresponding author: Francisco Gomez-Donoso (fgomez@ua.es)

This work was supported in part by the Spanish Government, through Feder funds under Grant TIN2016-76515R, and in part by the Spanish Grant for Ph.D. studies under Grant ACIF/2017/243 and Grant FPU16/00887. The work of E. Cruz was supported by the Panamenian Grant for Ph.D. studies IFARHU & SENACYT under Grant 270-2016-207.

ABSTRACT As it is well known, some versions of the Pepper robot provide poor depth perception due to the lenses it has in front of the tridimensional sensor. In this paper, we present a method to improving that faulty 3D perception. Our proposal is based on a combination of the actual depth readings of Pepper and a deep learning-based monocular depth estimation. As shown, the combination of both of them provides a better 3D representation of the scene. In previous works we made an initial approximation of this fusion technique, but it had some drawbacks. In this paper we analyze the pros and cons of the Pepper readings, the monocular depth estimation method and our previous fusion method. Finally, we demonstrate that the proposed fusion method outperforms them all.

INDEX TERMS Computer vision, image denoising, image registration, object recognition.

I. INTRODUCTION

In recent years, the interest for humanoid and social robotics has grown steadily. This expectancy has been fueled by the recent advances in materials, devices and artificial intelligence. In fact, a lot of manufacturers have already developed its own humanoid robots. However, the most famous of them is the Pepper robot, which has been created by Softbank Robotics. Unlike the others, the Pepper robot has not been developed to carry out heavy duty tasks or grasping objects precisely. Instead of that, the Pepper robot is intended to be deployed at indoor environments and has a clear social appeal. Nonetheless, this robot is equipped with a range of different sensors including color cameras, laser, ultrasonic sensor, touch surfaces and bumper switches. In addition it also features a depth camera. In the 1.8a version of the robot, which is the top seller, an ASUS Xtion device is in charge of the depth perception. This camera is widely used by the research community and has been praised for its quality despite being a low-cost device. However, the Xtion featured by the Pepper Robot v1.8a is not working properly. The point clouds that it provides are noisy and distorted, and full of artifacts. The scientific community believes that this is due to the lenses that the robot wears right in front of the sensor, but it has not been confirmed or denied by the manufacturer.

Anyway, the robot is used for a variety of research purposes that includes object recognition using 3D data, such as [1], or SLAM, such as [2], [3] or [4]. So, to implement these methods on it would be a hard task because of the faulty depth camera.

In addition to that specific issue, there are also a variety of other problems that can affect all time of flight and structured light cameras. For instance, these sensors provide low density point clouds, or fail on specular surfaces.

In this paper, we propose a method to improve the point clouds provided by the Pepper robot v1.8a but it can be used to enhance all point cloud-based cameras. Our proposal is able to provide higher density point clouds and fill the depth information of specular surfaces by involving a deep learning approach for depth estimation. A set of benchmarks validate the improvement of our approach over the depth maps provided by a Pepper robot v1.8a.

The paper is organized as follows: First, in Section II the state-of the art in the field is presented. After that, Section III

The associate editor coordinating the review of this manuscript and approving it for publication was Sudhakar Radhakrishnan.

describes the issues of Pepper's 3D camera and the issues of another depth prediction techniques. Next, Section IV details a description of the proposal. This is followed by the Section V, where the procedures for testing the proposed approach are described and the results of the experiments are presented. Finally, Section VI includes the discussion and conclusions of the work.

II. RELATED WORKS

This paper aims to overcome several existing three-dimensional sensor weaknesses by merging point clouds provided by a real depth camera and an estimated depth map from a deep learning approach. Therefore, this section is divided into the three relevant subsections. In addition, it is worth noting that this work is an extension of one of our previous work [5].

A. DEPTH CAMERAS ISSUES

Researchers in various fields, in particular, computer vision, human-computer interaction (HCI), augmented reality (AR), and robotics, rely on the use of depth sensors for their research. However, when it comes to the three-dimensional representation of the environment, these sensors face several problems. Factors related to deep noise maps were approached from multiple perspectives.

In [6], the noise of the Kinect depth images is characterized based on several elements. Their works also introduce a standard nomenclature for these noise types. They observed that there are four variables, namely, *imaging geometry*, *surface/medium property*, *sensor technology* control spatial noise, and *object distance*. These four control parameters and the noise source were used to characterize the noise behavior into *axial*, *out-of-range*, *specular surface*, *shadow*, *lateral*, *non-specular surface*, *band*, *residual noise* and *structural* classes. Every class summarizes the behavior of the noise either from the reports of others or from their tests. Motion, object distance, surface properties, and frame rate control the behavior of the temporal noise. Some of the problems with the depth camera using this categorization are described below.

The out-of-range noise [7] is produced by objects too near or too far. The characteristics of this noise are periphery usually has a high adjacency with far or near depth, Shares border with the image frame, and zero-depth (fails to estimate depth). The control parameter of this noise is Object distance.

The *lateral noise* [8] is produced by shadow-like non-uniformity at edges. The characteristics of this noise are pronounced for vertical edges and straight edges. Error along edges, varies linearly with depth, zero-depth, decrease with nearby background, and occurs at both: non-shadow edges and shadow.

In the case of the axial noise [8], it is produced by speckles per unit area drops quadratically with increasing distance. The characteristic of this noise is the accuracy decreases quadratically with increasing depth [9] and Wrong Depth (WD) value that occurs when the sensor reports a non-zero-depth value, its accuracy depends on the depth itself. The control parameter of this noise is Object distance.

The *shadow noise* [10] is caused by shadow-like non-uniformity on the edges. The properties of this noise are pronounced for vertical and straight edges. The error along the edges changes linearly with depth, zero-depth, decreases with nearby background, and occurs in both: non-shadow edges and shadow.

The *specular surface noise* [6] is caused by surfaces that are highly reflective to IR, so it fails to diffuse the speckle pattern. These noise properties are consistent across the frames; the zero-depth could be large irregular patches of zero-depth not adjacent to the edge of the image.

The Band Noise [11] is produced probably due to the effect of the windows of block correlation. The characteristics of this noise are zero-depth, Spreads as a vertical band at the left end, Has 8-pixels width. This occurs in all depth images. The control parameter of this noise is Sensor Technology.

The Structural Noise [12] is produced by Low spatial resolution of the sensor, wrapping of IR image by the lens, or disparity to depth transform. The characteristics of this noise are the depth of a plane varies at different points; Variations appear as waves, or circular ripples, Variations increase with distance. The control parameter of this noise is Sensor Technology.

In the case of the Residual Noise [11], its source is unknown - observed even after careful calibration with stereo-rig. The characteristics of this noise are Positive in the center and negative at the periphery, Vertical periodic stripes in-depth error, Independent of depth, and Dependent on the presence of other objects. The control parameter of this noise is Sensor Technology.

The Vibrating Noise [12] is produced by Speckle-based triangulation, the presence of depth-edges and specular surfaces, and motion. The characteristics of this noise are the depth of a stationary object vibrates with time, Vibrations increase with distance, Error increases with X but not with Y and Unstable Depth (UD) Value entailing reports a non-zero-depth value. However, that value changes over time even when there is no change of depth in the scene. The control parameter of this noise is Object, Distance, Motion, Surface, and Frame Rate.

B. SENSOR FUSION TECHNIQUES

In nature, it is widespread to identify species that can combine different signals to recognize their environment. Humans are capable of identifying their environment combining the information provided by a variety of sensors such as the auditory system, the speech system, the cutaneous system, and the visual system. Now the utilization of fusion concepts in technical areas is a discipline encompassing many fields of science [13]. Sensor fusion provides a substantial possibility to be able to overcome the physical limitations of the detection systems. In the state-of-the-art are many different definitions for data fusion.

Joint Directors of Laboratories (JDL) [14] defines data fusion as "A multi-level process dealing with the association, correlation, combination of data and information from single and multiple sources to achieve the refined position, identify estimates and complete and timely assessments of situations, threats, and their significance.”

Elmenreich W. [13] defines sensor Fusion as the combination of data from the sensors or data that derives from sensory data in such a manner that the output information is in some sense better than it would be possible when these sources were used individually. According to [15] the data fusion techniques can be categorized into three non-exclusive classes: (i) data association, (ii) state estimation, and (iii) decision fusion.

Data Association Techniques [15]. The purpose of data association methods is to define the set of observations or measurements produced by the same target through time. The most commonly used techniques to solve the data association problem are *Probabilistic Data Association*, *Nearest Neighbors* and *K-Means*, *Multiple Hypothesis Test*, *Joint Probabilistic Data Association*, *Distributed Joint Probabilistic Data Association*, *Graphical Models* and *Distributed Multiple Hypothesis Test*.

State Estimation Methods [15]. State estimation methods determine the state of the moving target (typically the position) given the observation or measurements. The most common estimation methods are the *Kalman filter*, *maximum likelihood* and *maximum posterior*, *particle filter*, the *distributed Kalman filter*, *covariance consistency methods* and *distributed particle filter*.

Decision Fusion Methods [15]. Usually, a choice is made based on the knowledge of the perceived situation, which is given by several sources in this domain. The most common Fusion Methods are the the *Dempster-Shafer Inference*, *bayesian methods*, *Semantic methods* and *Abductive Reasoning*.

Choosing the appropriate technique relies on the nature of the problem and the assumptions that have been made.

C. DEPTH ESTIMATION FROM MONOCULAR FRAMES

The estimation of depth based on image data has been under study for some time, based on stereo vision [16]. However, in this section, only the methods based on the monocular depth predictions will be reviewed (they are chronologically listed).

The first essay on this topic was published in 2005 by A.Saxena [17]. The approach used a supervised learning algorithm to estimate depth from a single monocular image. The method used discriminative-trained Markov Random Field (MRF) that included local and global features from the images.

Filling the gap between 2005 and 2010, appeared publications such as [18]–[22] which also contribute to improve monocular depth estimation task.

Liu et al. published [23] in 2010; they applied semantic segmentation of the scene as a guide to the three-dimensional reconstruction. The method also used MRF to enforce neighboring constraints.

Then was [24] published, it presented a depth transfer approach in three stages. The first stage was the search for the candidate images. The second stage applied to warp to the candidate and depth images. At the last stage, they interpolated and smoothed the warped candidate depth values; these outputs where the inferred depth.

Another approach was released the same year, [25], it was based on the observation that of all the image + depth pairs that were available online, there were probably many pairs whose three-dimensional content matched a 2D input (query).

David Eigen wrote in 2014 one of the most exceptional articles, [26]. Their approach relied on a coarse-scale network to predict the depth of the scene on a global scale. This was subsequently refined within the local regions using a large-scale network. The two stacks were applied to the original input, but also, the output from the coarse network was passed to the fine network as an added to the first layer image feature.

In 2014 was also published [27]. The authors present an approach to estimate the depth of a scene from a single image. They used the superpixel terms to address this issue, making the common assumption that each superpixel is planar.

In 2015, three different articles were released in this field:

[28] presented a deep convolutional neural field model for estimating depths from a single image, aiming to explore the capacity of deep CNN and continuous CRF jointly.

[29] have presented a new common framework for estimating the depth and normality of the surface of single monocular images, consisting of regression through deep CNNs and refinement through a hierarchical CRF.

[30] applied a trained Convolutional Neural Network (CNN) to jointly predict a global layout composed of pixel-wise depth values and semantic labels.

In 2016, Iro Laina et al. published [31], this work has been used as our baseline in this article. Their methods was based on the application of a fully convolutional architecture to predict depth, endowed with novel up-sampling blocks, that allowed for dense output maps of higher resolution and at the same time required fewer parameters.

Also the same year was published [32]. It used as input an image sized randomly and the outputs was a dense score map. Then, they applied fully connected CRFs to get the final depth estimation.

Another paper was [33], which captured scene details by considering information contained in depth gradients. It postulated that local structure encoded with first-order derivative terms.

In 2017 was published [34]. It predicted pixel-wise depth from a single color image. They proposed a simple and effective dilated deep residual CNN architecture, which converged with much fewer training examples and model parameters.

[35] proposed a supervised scheme for depth estimation employing unlabeled video clips with synthesis of the view, this is implemented using a depth CNN and a Pose CNN.

In 2018, [36] proposed three architectural and loss innovations that lead to large improvements in monocular depth. The innovations where: the introduction of a matching loss to solve the problem of occluded pixels. A automasking approach to ignore pixels where no relative camera motion is observed in monocular training. And a multi-scale appearance matching loss that performs all image sampling at the input resolution.

III. THE PEPPER ROBOT AND DEPTH ESTIMATION PROBLEMS

This paper introduces a resolution to the problem of the 3D camera for the Pepper Robot. First, we will briefly present the specifications of the Pepper Robot. The Pepper robot was developed by Aldebaran robotics, which is owned by Softbank. It was designed to interact with humans. The dimension of the Robot are: height 1.21 meters, width 0.48 meters, and depth 0.425 meters. It has a weighs of 28 kilograms (62 lb), and it contains a lithium-ion battery which allows an operation time of approximately 12hrs. In the Head, are two RGB cameras, four Mic, a 3D sensor, and three touch sensors. In the hands are two touch sensors and in the legs: two sonar sensor, six laser sensor, three bumper sensor, and a gyro sensor. Figure 1 shows some of the sensor listings.

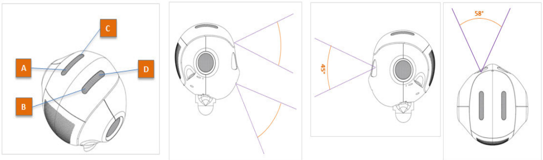

FIGURE 1. Sensors listed on the Pepper robot. The first image represents the location of the microphone set. In the second image, the position of two cameras are shown in color and also indicates the field of view. Third and fourth images feature the 3D camera.

Currently, the Robot has three different hardware versions. The latest version 1.8 includes a brand-new three-dimensional sensor. In the newer robot version, the depth sensor was replaced per two 4 MPx color cameras, and using a stereo algorithm provided depth perception. Outcome depth maps are generated at 15 fps and have a resolution of 1280 × 720. However, versions 1.6 and 1.8a, are equipped both with the defective depth sensor, Asus Xtion. The Asus Xtion depth sensor can provide depth maps with a resolution of 320 × 240 at 20 frames per second.

A. THE ISSUE WITH THE 3D SENSOR OF THE PEPPER ROBOT

As already mentioned, the Pepper Robot has an Asus Xtion sensor as a depth camera. Theoretically, this camera can provide accurate depth maps of the scenes. However, the Xtion camera mounted on this Robot appears to provide erroneous depth maps, which also results in incorrect point clouds.

As shown in Figure 2, this issue lies in a distorted representation of tridimensional space. The resulting point clouds reveal a wave-like pattern throughout the scene. This issue becomes more evident when it represents plane artifacts, like walls or floors, but it occurs throughout the whole scene. Also, we noticed that the distortion worsens with increasing depth, namely, objects near to the sensor were less affected than those further away.

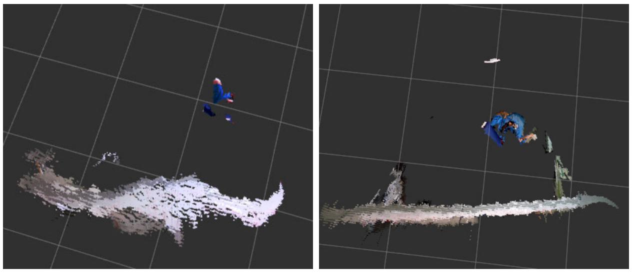

FIGURE 2. In these images, it is shows an external standalone Xtion point cloud and the Pepper Robot's Xtion camera. Both images are taken from the same point of view and represent the same scene. The white artifact is a planar surface that, in this case, is a wall.

In order to quantitatively evaluate the distortion of the depth camera of the Pepper Robot, a simple experiment was performed. We took a flat object and placed it at different distances from the 3D sensor of the Pepper Robot. The distances varied between 1 and 3 meters, with an increase in distance of 0.5 meters. For each case, the corresponding depth map was captured and projected into 3D space. Then, all the points on the flat object were selected manually. Depth maps and colored frames are recorded, making this step simple. Then, we fitted a plane using RANSAC [37], setting the inlier threshold to infinite. This process is done to ensure that all points are unstable values. 60000 different plane were tested, but only the one with the lowest RMSE is returned, this was done due to the random nature of RANSAC. The results of these experiments can be found in Figure 3. The mean Euclidean distance from each point to the estimated plane is reported for each experiment in the Figure.

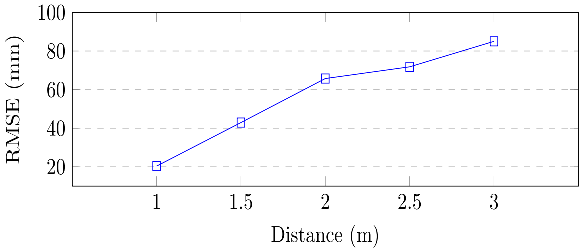

FIGURE 3. This Figure shows the mean error per distance to the plane reached by the Robot.

With the results obtained, we can affirm that the mean error grows as the distance to the object increases. In other words, the quality of the resulting point clouds gradually worsens. In absolute terms, the 3D representation of the planar object is reduced even when the object is as close as 1 meter. In this case, the mean error is 20.36 mm, but the furthest point is 79.15*mm* from the plane. The furthest point in the 3-meter experiment is 279.12*mm*. Methodological details related to Pepper and the version of NaoQi used in this experiment can be found in Section V. Figure 4 shows the point clouds and

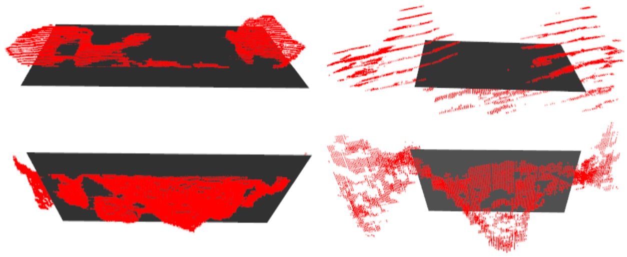

FIGURE 4. Estimated planes (in grey) for the point cloud representations (in red) of a planar object. The first column is the same planar object located at 1 meter from the sensor and the second column is the same planar object located at 3.5 Meters. It is worth mentioning that the representation is distant from accurate in both cases, and the problem becomes worse as the distance increases.

the corresponding planes estimated for 1 and 3 meters for qualitative assessment purposes.

It should be noted that this is not an isolated problem or a flaw in our unit. To make sure of them, we have personally contacted different researchers from different laboratories and universities who have reported the same issue on the Robot with the same version.

B. OTHER ISSUES RELATED TO 3D CAMERAS

There are other problems that commonly affect all sensors based on time-of-flight and structured light. These technologies are used by most commercial depth cameras, such as Microsoft Kinect and Asus Xtion. The first issue we are discussing is the impossibility of calculating the depth in shadow that foreground objects project onto background objects. This case produces some areas without depth enforcing around the boundaries of the objects in the scene. The Figure 5 Shown, this case. In there is no in the shadow of a conical form that projects the chair in the foreground.

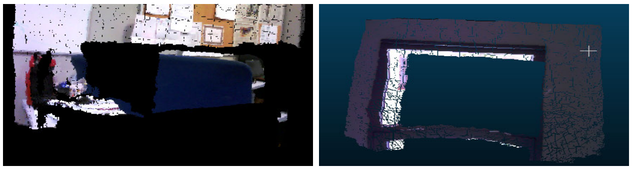

FIGURE 5. The leftmost image depicts the shadow effect issue of the structured light and time-of-flight sensors. The rightmost image shows another problem with these sensors, which is the incompatibility with specular surfaces.

The incompatibility with specular surfaces is another issue. Objects with specular surfaces are not properly sensed by these devices. They produce faulty depth values or areas with no depth information. This case is shown in Figure 5. As can see, the depicted point cloud shows a hole in the place of the specular object.

C. ISSUES RELATED TO MONOCULAR DEPTH ESTIMATION

The monocular depth estimation obtained by the Iro Laina network has two main advantages. First, it provides estimations for every pixel on the image, so it generates a very dense point cloud, and it also usually keeps the geometry of the entities present in the scene.

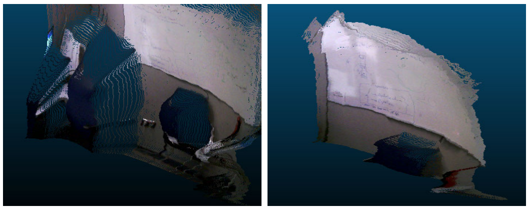

FIGURE 6. Iro Laina's approach produces trailing artifacts from the edges of the objects prointing to the background.

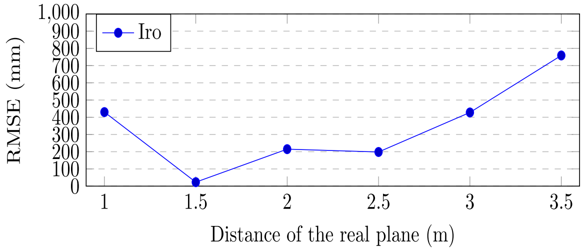

FIGURE 7. RMSE of the distance to the real plane of the points inferred by Iro Laina.

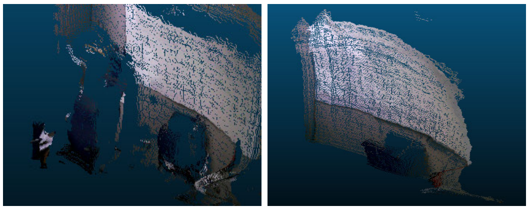

FIGURE 8. FusionV1 provided poor density point clouds as a result of a strict filtering process.

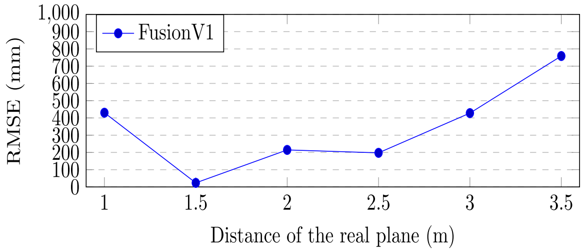

FIGURE 9. RMSE of the distance to the real plane of the points fused by FusionV1.

Nevertheless, it produces several artifacts on the resulting estimation. As shown in Figure 6, a trailing artifact is generated near the edges of the objects in the foreground pointing the background.

We tested the accuracy of the scale and depth values by carrying out the following experiment. We took images of a planar object with the sensor located at certain distances from it. Then, we estimated the depth maths with the Iro Laina's approach and computed the root mean square error (RMSE) error between the actual distance and the obtained by the sensor. Results of this experiment are shown in Figure 7.

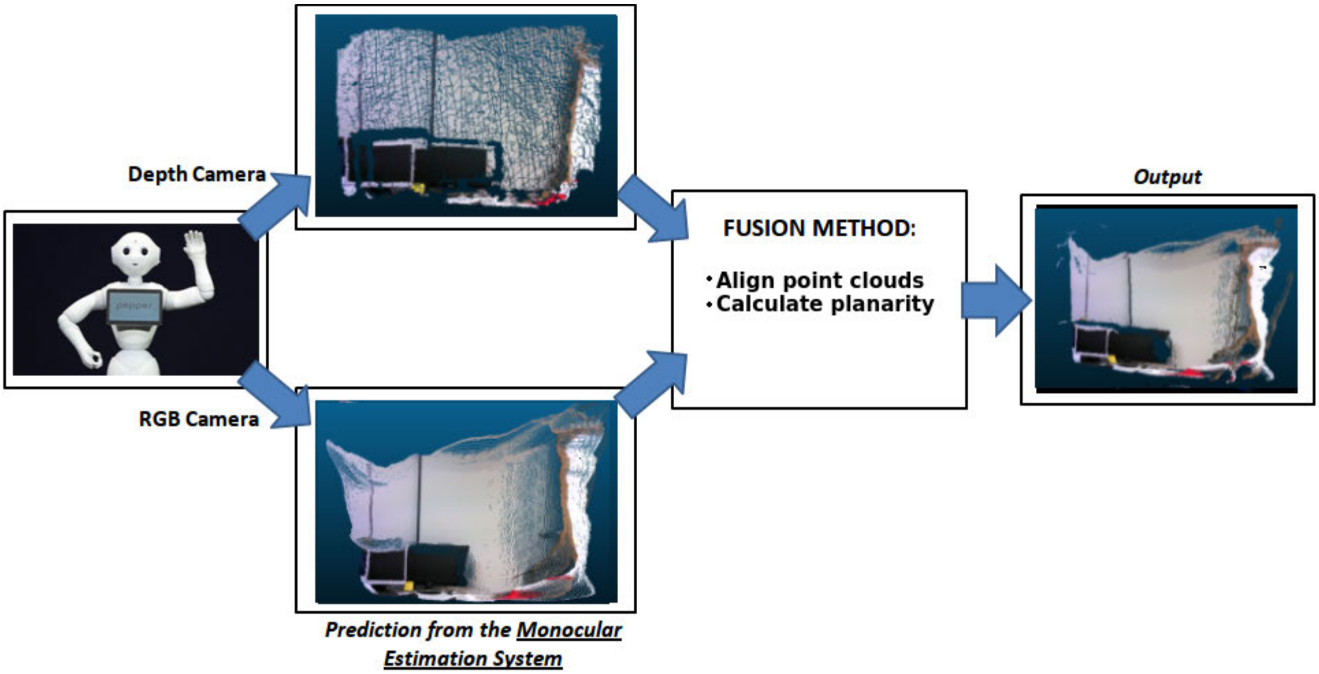

FIGURE 10. Scheme of the new fusion method FusionV2.

D. ISSUES RELATED TO THE PREVIOUS FUSION METHOD

In our previous work [38], presented at the Workshop of Physical Agents 2018 (WAF 2018), we proposed a method to fusion the depth information from Pepper and Iro Laina's approach. Despite the fact that it solved the trailing artifacts contributed by the monocular depth estimation, our method, hereinafter referred to as *FusionV1*, establishes a very restrictive filter and rejected a great quantity of points, as depicted in Figure 8. As a result, the method provided low-density point clouds. In addition, FusionV1 preserves the poorly accurate depth predictions of Iro Laina's approach as shown in Figure 9.

IV. FUSIONV2: REFINING THE FUSION OF MONOCULAR AND PEPPER DEPTH MAPS

In order to solve the main issues of FusionV1, we propose changes to this previous method, that takes the noisy and incorrect point clouds from Pepper and the point clouds as provided by the Iro Laina's method, and returns a new corrected point cloud.

First, we obtain the raw point cloud from Pepper and make no preprocessing.

Then, Iro Laina's approach [31] estimates depth map from a monocular frame using a fully convolutional neural network. Then, both point clouds are summarized. The fusion process consists in mixing the point cloud from Pepper's camera and the one estimated from the monocular color image. This process is carry out as follows:

1) **Align the monocular depth estimation with Pepper's point cloud.** First, we want to align both point clouds. This step would be computationally expensive if we use the whole points, so we perform uniform sampling on the Pepper's point cloud and obtain the corresponding points from the monocular cloud (this step is straightforward because both are registered and obtained from a depth map of the same resolution). Then, we align the point cloud estimated with Iro Laina's approach with Pepper's using a variant of the Iterative Closest Point algorithm that employs the Single Value Decomposition technique. This method not only calculates translation and rotation between both point clouds but the scale component is inferred too [39]. As a result, the estimated point cloud with Iro Laina's method is set in the correct scale.
2) **Calculate the planarity of every point in monocular estimation.** We calculate a planarity measure for each point in the Iro Laina's method estimated point cloud. First, we search its neighboring points in the same point cloud using a radial search within a range threshold r. Then, we use RANSAC to fit a plane using these points. This method is robust against outliers and calculates the plane that better represents the data. Then, the ratio of neighboring points whose distance to the plane is less than a certain threshold of d is calculated. This ratio is called *planarity measure*. If this ratio is higher than a threshold of Tp, then the monocular estimation point is inserted in the output points Cloud fused. Contrarily, the output information is taken from the Pepper point cloud.

The resultant point cloud has the density and geometric shapes of the point clouds estimated by Iro Laina's method and preserves the correct depth values and scale provided by Pepper. This approach is depicted in Figure 10.

V. EXPERIMENTATION

The hardware setup we used is described next.

The experiments are carried out using version 1.8a of the Pepper Robot, which includes an Asus Xtion as a 3D sensor. The manufacturer provides a set of parameters for the camera, so it is not necessary to estimate the intrinsic parameters. The center of the image in the X-axis is 319.5, and the center of the image in the Y-axis is 239.5. Its focal length is 525. With these parameters, we can project the provided depth maps into 3D point clouds.

For the computation of the fusion method and the monocular depth estimation, we used an external computer with 8 GiB HyperX DDR3 RAM (Kingston) 1600 MHz on an Asus P8H77-M PRO. Also, include a processor Intel Core i5-3570 and an NVIDIA Quadro P6000 GPU. To execute the DL tasks, we used TensorFlow 1.8 as the core with Keras 1.2.0. The OS used was Ubuntu 16.04. To accelerate the computations, we used CUDA 9.0 and cuDNN v7.5.

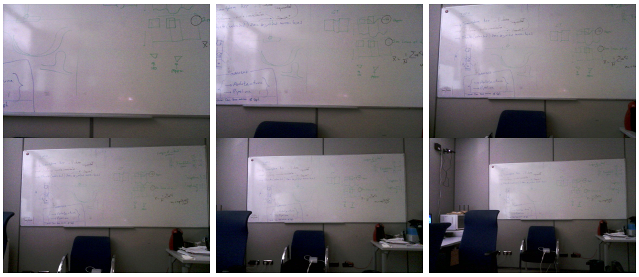

FIGURE 11. Monocular captures of the scenes we used in the experiments.

To ensure the reliability of our proposal, we have performed a set of experiments. We took several captures of a whiteboard of our workplace at different distances with Pepper's color camera and depth sensor, as seen in Figure 11.

We put to test the point clouds generated by Pepper, Iro Laina's method, our former fusion method *FusionV1*, and our new fusion method *FusionV2*. As the generated point clouds depicted more objects in addition to the planar object, we manually selected the whiteboard planes and removed the remaining points. These resultant point clouds, which are depicted in Figure 12, are used to the experiments.

The first metric we tested is *Planarity*. This metric consists in fitting the best possible plane for every cloud using RANSAC, and calculate the RMSE of all the points. The results are shown in Table 1 and Figure 13. As can be seen, the planarity RMSE of the Pepper point clouds worsen as the distance is increased. On the other hand, the planarity RMSE of Iro Laina's approach and both fusion methods are more dependent on the visual features than on the distance ob the objects to the sensos. Nonetheless these methods provide better planar objects.

The second metric is *Distance precision*. This metric consists in determine the difference between the estimated distance and the real one (we have measured the exact distance to the planes with a laser), and calculate the RMSE of all the points. The results are shown in Table 2 and Figure 14. As depicted in previous figures, the distance RMSE of the Pepper planes is far better than the Iro's, that suffers much more problems with the scale of the cloud. However, our method *FusionV2* preserves the good distance estimation of Pepper.

The last metric is *Density*. This metric measures the proportion of good points (not NaNs) in respect of the totality of

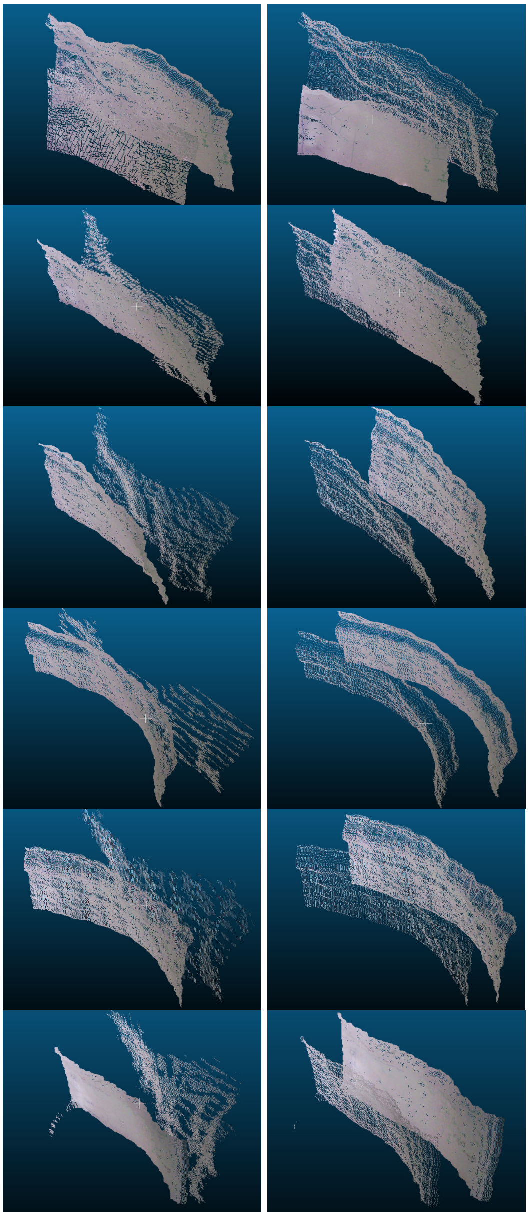

FIGURE 12. Comparative of the generated planes from Pepper, Iro, FusionV1, FusionV2 point clouds. On left images, Pepper and Iro clouds are compared. On right images, FusionV1 and FusionV2 are compared. This split has been done because Pepper and FusionV2 overlap in distance, and Iro and FusionV1 are quite similar in distance and geometry. On the left image, the less dense cloud corresponds to Pepper. On the right image, the less dense cloud corresponds to FusionV1.

TABLE 1. Planarity results obtained for the point clouds provided by the different methods at different distances. Distance is in meters whilst the results are in millimeters.

Distance	Pepper	Iro Laina	FusionV1	FusionV2
1	12.422	22.0738	22.2634	15.1578
1.5	23.2205	11.9214	12.3538	12.079
2	37.6301	19.7721	21.2261	22.2253
2.5	56.4277	69.0246	67.4973	73.1071
3	70.0697	62.7262	61.3264	71.8479
3.5	112.202	47.8647	38.4259	53.9838

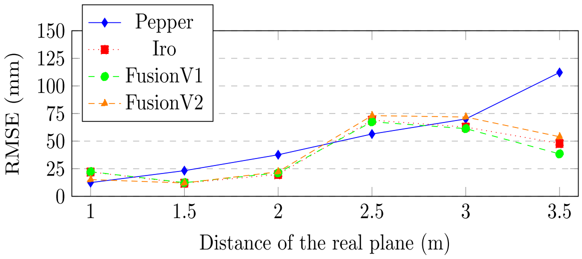

FIGURE 13. RMSE distances of the point clouds to the fitted plane.

TABLE 2. RMSE distances of the point clouds to the real plane. Distance is in meters whilst the results are in millimeters.

Distance	Pepper	Iro Laina	FusionV1	FusionV2
1	93.7632	430.007	430.114	91.2979
1.5	81.9094	23.0732	23.5829	79.1551
2	55.7695	214.786	214.566	129.712
2.5	70.1055	198.306	197.639	120.525
3	119.266	427.858	427.958	106.923
3.5	285.817	758.839	758.784	345.566

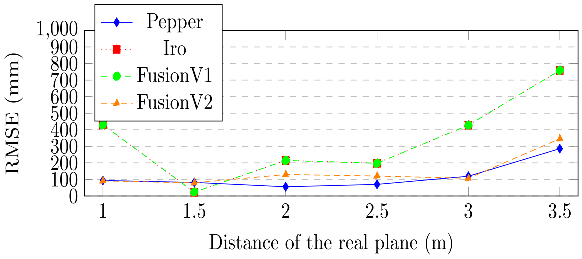

FIGURE 14. RMSE distances to the real plane for this experiment.

TABLE 3. Density of the point clouds.

Distance (m)	Pepper (%)	Iro (%)	FusionV1 (%)	FusionV2 (%)
1	23.9551	100	23.9551	100
1.5	23.5073	100	39.8207	100
2	23.4122	100	23.4122	100
2.5	23.2482	100	23.2482	100
3	23.2825	100	23.2825	100
3.5	22.9414	100	22.9414	100

points in the cloud. The results are shown in Table 3. These results show the lack of resolution that suffers Pepper's point cloud and the good quality of Iro's. Our previous method *FusionV1* inherits these bad density in some situations but our new method, *FusionV2*, preserves the density quality of Iro's.

VI. CONCLUSION

We present a refinement of our previously proposed method for improving the quality of the Pepper 1.8a depth map.

The camera of Pepper suffers some kind of radial distortion, arguably produced by its lenses, that produces several artifacts on planes. In order to correct this geometrical issue, we make use of the monocular depth estimation of *Iro Laina*'s architecture, that performs better with planar representation.

We make a fusion between the Pepper and monocular point cloud, aligning the second with the first one and looking for planar areas. Thus, we provide a point cloud of better quality.

In our previous paper, we suggested the improve of the distance estimation provided by the monocular method. As shown in the Experimentation section, our new fused output point cloud represents far better the distance of the scene than our previous one, and preserves the density of points of the monocular depth estimation.

ACKNOWLEDGMENT

Thanks to Nvidia for the generous donation of a Quadro P6000 and a Titan Xp.

REFERENCES

[1] A. Garcia-Garcia, F. Gomez-Donoso, J. Garcia-Rodriguez, S. Orts-Escolano, M. Cazorla, and J. Azorin-Lopez, "PointNet: A 3D convolutional neural network for real-time object class recognition," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2016, pp. 1578–1584.

[2] N. Engelhard, F. Endres, J. Hess, J. Sturm, and W. Burgard, "Real-time 3D visual SLAM with a hand-held RGB-D camera," in *Proc. RGB-D Workshop 3D Perception Robot. Eur. Robot. Forum*, Vasteras, Sweden, Apr. 2011.

[3] P. Kim, J. Chen, and Y. K. Cho, "SLAM-driven robotic mapping and registration of 3D point clouds," *Autom. Construct.*, vol. 89, pp. 38–48, May 2018.

[4] P. Kim, J. Chen, J. Kim, and Y. K. Cho, "SLAM-driven intelligent autonomous mobile robot navigation for construction applications," in *Advanced Computing Strategies for Engineering*, F. C. Smith and B. Domer, Eds. Cham, Switzerland: Springer, 2018, pp. 254–269.

[5] Z. Bauer, F. Escalona, E. Cruz, M. Cazorla, and F. Gomez-Donoso, "Improving the 3D perception of the pepper robot using depth prediction from monocular frames," in *Advances in Physical Agents*, R. F. Pizán, Á. G. Olaya, M. P. S. Lorente, J. A. I. Martínez, and A. L. Espino, Eds. Cham, Switzerland: Springer, 2019, pp. 132–146.

[6] T. Mallick, P. P. Das, and A. K. Majumdar, "Characterizations of noise in Kinect depth images: A review," *IEEE Sensors*, vol. 14, no. 6, pp. 1731–1740, Jun. 2014.

[7] *Depth Camera*, Microsoft, Redmond, WA, USA, 2012.

[8] C. V. Nguyen, S. Izadi, and D. Lovell, "Modeling kinect sensor noise for improved 3D reconstruction and tracking," in *Proc. 2nd Int. Conf. 3D Imag., Modeling, Process., Vis. Transmiss.*, Oct. 2012, pp. 524–530.

[9] K. Khoshelham and E. O. Elberink, "Accuracy and resolution of kinect depth data for indoor mapping applications," *Sensors*, vol. 12, no. 2, pp. 1437–1454, 2013.

[10] Y. Yu, Y. Song, Y. Zhang, and S. Wen, "A shadow repair approach for kinect depth maps," in *Computer Vision—ACCV*, K. M. Lee, Y. Matsushita, J. M. Rehg, and Z. Hu, Eds. Berlin, Germany: Springer, 2013, pp. 615–626.

[11] J. Smisek, M. Jancosek, and T. Pajdla, "3D with kinect," in *Consumer Depth Cameras for Computer Vision*. London, U.K.: Springer, 2013, pp. 3–25.

[12] M. R. Andersen, T. Jensen, P. Lisouski, A. K. Mortensen, M. K. Hansen, T. Gregersen, and P. Ahrendt, "Kinect depth sensor evaluation for computer vision applications," Dept. Eng. Elect. Comput. Eng., Aarhus Univ., Aarhus, Denmark, 2012.

[13] W. Elmenreich, "An introduction to sensor fusion," Inst. Technische Informatik, Technische Univ. Wien, Vienna, Austria, Res. Rep. 47/2001, 2001.

[14] F. E. White, "Data Fusion Lexicon," Joint Directors Lab., Tech. Panel C3, Data Fusion Sub-Panel, Nav. Ocean Syst. Center, San Diego, CA, USA, Tech. Rep. 144275, 1991.

[15] F. Castanedo, "A review of data fusion techniques," *Sci. World J.*, vol. 2013, p. 19, Aug. 2013.

[16] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Proc. 12th Eur. Conf. Comput. Vis. (ECCV)*. Berlin, Germany: Springer-Verlag, 2012, pp. 746–760.

[17] A. Saxena, H. S. Chung, and Y. A. Ng, "Learning depth from single monocular images," in *Advances in Neural Information Processing Systems*, Y. Weiss, B. Schölkopf, and J. C. Platt, Eds. Cambridge, MA, USA: MIT Press, 2006, pp. 1161–1168.

[18] E. Delage, H. Lee, and Y. A. Ng, "A dynamic Bayesian network model for autonomous 3D reconstruction from a single indoor image," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jun. 2006, pp. 2418–2428.

[19] A. Saxena, J. Schulte, and Y. A. Ng, "Depth estimation using monocular and stereo cues," in *Proc. 20th Int. Joint Conf. Artif. Intell. (IJCAI)*. San Francisco, CA, USA: Morgan Kaufmann Publishers, 2007, pp. 2197–2203.

[20] A. Saxena, H. S. Chung, and Y. A. Ng, "3-D depth reconstruction from a single still image," *Int. J. Comput. Vis.*, vol. 76, no. 1, pp. 53–69, 2008.

[21] A. Saxena, M. Sun, and A. Y. Ng, "Make3D: Learning 3D scene structure from a single still image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 5, pp. 824–840, May 2009.

[22] H. Trinh and A. D. McAllester, "Unsupervised learning of stereo vision with monocular depth cues," in *Proc. BMVC*, 2009.

[23] B. Liu, S. Gould, and D. Koller, "Single image depth estimation from predicted semantic labels," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 1253–1260.

[24] K. Karsch, C. Liu, and S. B. Kang, "Depth extraction from video using non-parametric sampling," in *Proc. 12th Eur. Conf. Comput. Vis. (ECCV)*. Berlin, Germany: Springer-Verlag, 2012, pp. 775–788.

[25] J. Konrad, M. Wang, and P. Ishwar, "2D-to-3D image conversion by learning depth from examples," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2012, pp. 16–22.

[26] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Proc. 27th Int. Conf. Neural Inf. Process. Syst. (NIPS)*. Cambridge, MA, USA: MIT Press, vol. 2, 2014, pp. 2366–2374.

[27] M. Liu, M. Salzmann, and X. He, "Discrete-continuous depth estimation from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Washington, DC, USA, Jun. 2014, pp. 716–723.

[28] F. Liu, C. Shen, and G. Lin, "Deep convolutional neural fields for depth estimation from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 5162–5170.

[29] B. Li, C. Shen, Y. Dai, A. van den Hengel, and M. He, "Depth and surface normal estimation from monocular images using regression on deep features and hierarchical CRFs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1119–1127.

[30] P. Wang, X. Shen, Z. Lin, S. Cohen, B. Price, and A. Yuille, "Towards unified depth and semantic prediction from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2800–2809.

[31] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," *CoRR*, vol. abs/1606.00373, Sep. 2016.

[32] Y. Cao, Z. Wu, and C. Shen, "Estimating depth from monocular images as classification using deep fully convolutional residual networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 11, pp. 3174–3182, Nov. 2018.

[33] J. Li, R. Klein, and A. Yao, "Learning fine-scaled depth maps from single RGB images," *CoRR*, vol. abs/1607.00730, Aug. 2016.

[34] B. Li, Y. Dai, H. Chen, and M. He, "Single image depth estimation by dilated deep residual convolutional neural network and soft-weight-sum inference," *CoRR*, vol. abs/1705.00534, May 2017.

[35] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proc. CVPR*, Jul. 2017, pp. 1851–1858.

[36] C. Godard, O. M. Aodha, M. Firman, and G. Brostow, "Digging into self-supervised monocular depth estimation," 2018, *arXiv:1806.01260*. [Online]. Available: https://arxiv.org/abs/1806.01260

[37] A. M. Fischler and C. R. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, Jun. 1981.

[38] Z. Bauer, F. Escalona, E. Cruz, M. Cazorla, and F. Gomez-Donoso, "Improving the 3D perception of the pepper robot using depth prediction from monocular frames," in *Proc. Workshop Phys. Agents*. Cham, Switzerland: Springer, 2018, pp. 132–146.

[39] S. Du, N. Zheng, S. Ying, Q. You, and Y. Wu, "An extension of the ICP algorithm considering scale factor," in *Proc. IEEE Int. Conf. Image Process.*, vol. 5, Sep./Oct. 2007, pp. V-193–V-196.

ZURIA BAUER received the bachelor's degree in chemical engineering and the master's degree in automation and robotics from the University of Alicante, where she is currently pursuing the Ph.D. degree in computer science with the RoViT lab, under the supervision of M. Cazorla and S. Orts. Her thesis is focused on the development of a system to help people with visual impairment to be able to move autonomously in outdoor environments (covering depth prediction, and obstacle detection). Her publications include four JCR, two Workshops, and one Core A Congress.

FELIX ESCALONA was born in Alicante, Spain, in 1994. He received the B.S. degree in computer science, earning the extraordinary end-of-course prize, and the M.S. degree in robotics from the University of Alicante, Spain, in 2016 and 2017, respectively, where he is currently pursuing the Ph.D. degree in computer science, with an FPU scholarship.

His research interests include 3D object recognition, segmentation, scene understanding, and domestic and social robotics.

EDMANUEL CRUZ received the B.S. degree in software development from the Technological University of Panama, Panama, in 2010, the master's degree in higher education from the University of Panama, in 2013, and the double M.S. degrees in communication technology from the University of Wales and Centro Universitario CESTE, Zaragoza, Spain. He is currently pursuing the Ph.D. degree in computer science with the University of Alicante. He has published six JCR and four international conferences. The main interests of his research are human–computer interaction, deep learning, and machine learning.

MIGUEL CAZORLA received the degree in computer science from the University of Alicante, in 1995, and the Ph.D. degree in computer science from the University of Alicante, in 2000. He is currently a Full Professor with the University of Alicante.

His research has focused on lines related to the use of computer vision as the main sensor in certain robotics tasks, such as mapping and location. In recent years, he has focused on 3D data processing (mapping, object recognition). In his latest publications he focuses on deep learning techniques for traditional and 3D vision. More specifically, his latest projects have focused on welfare robotics, creating applications and solutions for people with disabilities. He has focused mainly on people with acquired brain damage. This group has the particularity of needing products that adapt to their disability, as it is usually very variable. This is what these projects pursue: to find systems that adapt to the patient's disability. Recently, his research has been opened to elderly people in their homes, as they share some of the solutions sought. In this case it is oriented to the monitoring, attention, and identification of people.

As a result of these projects, the large number of publications (in the last two projects, 6 years, more than 40 JCRs), thesis direction (8) and a patent stand out. We are in the process of creating a technology-based company, with results from the projects mentioned above.

FRANCISCO GOMEZ-DONOSO received the B.S. degree in computer science and the master's degree in robotics and automatic from the University of Alicante, Spain, in 2014 and 2015, respectively, where he is currently purusing the Ph.D. degree in computer science programme. His main interests are human–computer interaction, deep learning and machine learning, and tridimensional data processing. Regarding his experience as a scientific, he has published more than 35 articles in high-impact journals and conferences.