IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 10, 2020, accepted December 23, 2020, date of publication December 29, 2020,
date of current version January 7, 2021.

Digital Object Identifier 10.1109/ACCESS.2020.3048053

Robot Vision for Manipulation: A Trip to Real-World Applications

ESTER MARTINEZ-MARTIN[1], (Senior Member, IEEE), AND ANGEL P. DEL POBIL[2,3], (Member, IEEE)

[1]RoViT, Department of Computer Science and Artificial Intelligence, University of Alicante, E-03690 San Vicent del Raspeig, Spain
[2]Robotic Intelligence Lab (RobInLab), Department of Engineering and Computer Science, Universitat Jaume I, E-12071 Castellón de la Plana, Spain
[3]Department of Interaction Science, Sungkyunkwan University, Seoul 110-745, South Korea

Corresponding author: Ester Martinez-Martin (ester@ua.es)

This work was supported in part by the Ministerio de Economía y Competitividad under Grant DPI2015-69041-R, in part by Universitat Jaume I under Grant UJI-B2018-74, and in part by Generalitat Valenciana under Grant PROMETEO/2020/034 and GV/2020/051.

ABSTRACT Along the last decades, Robotics research has taken a major turn from laboratories to factories and ordinary real-world environments. Consequently, new issues to be overcome have arisen, specially when autonomous, dexterous robots are in place. In this paper, we present this evolution in the case of robot vision for manipulation through several robot developments, by analysing their challenges and proposed solutions. This overview highlights the need of using different techniques depending on the task at hand and the scenario to work in.

INDEX TERMS Robotics, computer vision, applications.

I. INTRODUCTION

Robotics research has evolved from industry to everyday scenarios. This evolution has required to adapt the robot developments from restricted, controlled and well-known settings to dynamic, unknown and populated environments. Therefore, robots must be endowed with different abilities to be able to autonomously perform meaningful tasks such as navigating in populated environments, localisation, recognition of different targets, manipulation, human–robot interaction, reasoning, co-working with people, etc. This requires the development of the necessary sensory-motor skills to engage and integrate all the aspects of intelligent processing from perception to action, but without culminating in time-consumption processes.

For that, it is necessary to overcome several issues arising for each task or scenario type such that the robot is capable of operating in a flexible manner, without constraining the environment, and in a reasonable time.

This paper presents an overview of our trip from robots sealed in workstations in factories to those working in people's living spaces. This progressive way is described by means of several developments, that pertain to robot vision for manipulation, with their respective challenges and provided solutions. Our purpose is not to carry out a global comparative analysis of vision techniques for robot tasks, but rather to present a number of different application scenarios and specific robot tasks, putting in perspective the challenges to be tackled, our proposed solutions, and the lessons learnt. So, Section II describes a vision system for industrial robots aimed to guarantee the safety of all the surrounding elements whether it is performing its own tasks or carrying out a collaborative task with a human, including a comparative analysis with other approaches. Going a step further, Section III shows two robot applications in semi-structured environments. In particular, a librarian robot and a robot for a warehouse are presented. Then, Section IV analyses two main aspects for autonomous robots in real scenarios: detection and recovery from manipulation errors; and object detection and recognition. Next, Section V discusses the scientific advances of the proposed solutions, while pointing out their limitations, lessons learnt, and possible improvements. Finally, some conclusions are presented in Section VI.

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

II. HUMAN-ROBOT COLLABORATION IN INDUSTRIAL SETTINGS

Although the industrial robots were initially isolated, the human collaboration was soon required. However, the used safety systems like cages or laser fencing, became inappropriate for those collaborative tasks since they stop the robot activity when a person is close. This fact resulted in the need to develop new technologies guaranteeing the

FIGURE 1. Difference between a catadioptric image (left) and a fisheye image (right).

performance of robot tasks in a safety way (i.e. avoiding any collision with people and other robots).

Keeping in mind the robot autonomy and flexibility, vision systems could fit given the amount of data they can provide. Nevertheless, this challenge poses several issues to be overcome:

- how to efficiently cover the whole robot work-space
- how to overlook minor dynamic factors such as the blinking of computer screens, mirror images on glasses, sensor noise or non-uniform attenuation
- how to deal with changes in illumination due to both shadows or other events like switching on/off a light or opening/closing a window
- how to detect people when they stop for a moment
- how to accurately locate all the surrounding elements with respect to the robot
- how to properly track all the surrounding elements

With the purpose of covering all the embracing robot space, some devices were studied. So, traditional cameras were discarded because they have a limited field of view and the computational cost of the feature correspondence algorithms is too high when the images from several cameras or those generated from a rotating camera, are combined and processed. Something similar happens with range images and they also required mechanisms to deal with missing depth information and the adjustment of several parameters to properly establish the correspondence between several visual sensors [1]. Alternatively, a camera combined with mirrors (i.e. catadioptric cameras) could be considered [2], [3]. However, their images exhibit a dead area in the centre, what results in an important loss of valuable information (see Figure 1). Thus, dioptric (fisheye) devices are employed [4]. Unlike catadioptric cameras, a fisheye lens is used together with a traditional camera in these devices and this fact avoids dead areas in the captured images as illustrated in Figure 1. Given that a 185-degree field of view is provided, two fisheye cameras placed at both sides of the robot, pointing upwards, are enough to cover the whole robot work-space.

The next issue is how to properly detect and track both the human collaborators and the moving elements around the robot. For that, motion is considered as a primary cue since it provides a stimulus to perceive the surrounding elements within images, just as the primate brain does [5]. In addition, motion may lead to some extra information meaningful for detection and recognition such as the element's shape, speed or trajectory.

A wide research has been done along this line (see [6]–[10] for a deep analysis). However, taking into account the above-mentioned factors and the lack of constraints about the *targets*, we have designed a two-stage adaptive background model allowing the robot to monitor all the activity around it and adapting to that activity while safely performing its tasks. So, the first stage builds an initial statistical background model with no constraints on the environment and activity (i.e. there is no need of a background free of target elements). For that, it takes advantage of difference techniques such that a combination of them is in charge of controlling these factors and removing this information from the background model under construction. Thus, as illustrated in Figure 2, firstly, a simple combination of difference techniques classifies the pixels as background or targets based on their motion between frames and, then, two consecutive morphological operations are used to erase isolated points or lines. This data is the input to the module to build the statistical background model only updating the information of the pixels classified as background. In addition, global changes in illumination are also detected from the output of the *difference* approach such that a change in two thirds of the image is considered a change in illumination. This fact results in a re-initialization of the process and, consequently, the background model. Note that the applied thresholds in the *difference* approach are automatically set for each pixel from pixel neighbourhood information based on the first taken frame. After several experiments, this stage has set to last 100 frames as maximum (4 seconds at a rate of 25 fps).

Once the statistical background model (i.e. a simple Gaussian model) has been built, the second stage starts. In this case, two different tasks are performed. On the one hand, it continuously monitors the activity around the robot such that it can perform its manipulation tasks without causing any damage. On the other hand, an identification and tracking process takes place when a collaborative task is expected.

Regarding the work-space monitoring task, as previously, it implies an image analysis at two levels. So, the frame level allows the robot to detect when a global change in illumination takes place. On the contrary, the pixel level is used to properly detect the presence of any person or other moving robot within the scene. For that, the first step is to apply the *difference* approach considered in the previous stage. This raw classification is now refined by means of the background model as follows:

$$\begin{cases} 255 \text{ if } |i_t(x) - \mu_t(x)| > (k_x * \sigma_t(x)) \\ 0 \text{ otherwise} \end{cases} \quad (1)$$

where $i_t(x)$ corresponds to the value of pixel x at time t; $\mu_t(x)$ and $\sigma_t(x)$ respectively refer to the mean and standard deviation of the background Gaussian model for each pixel x at time t, while k_t is a factor between 0.0 and 3.0

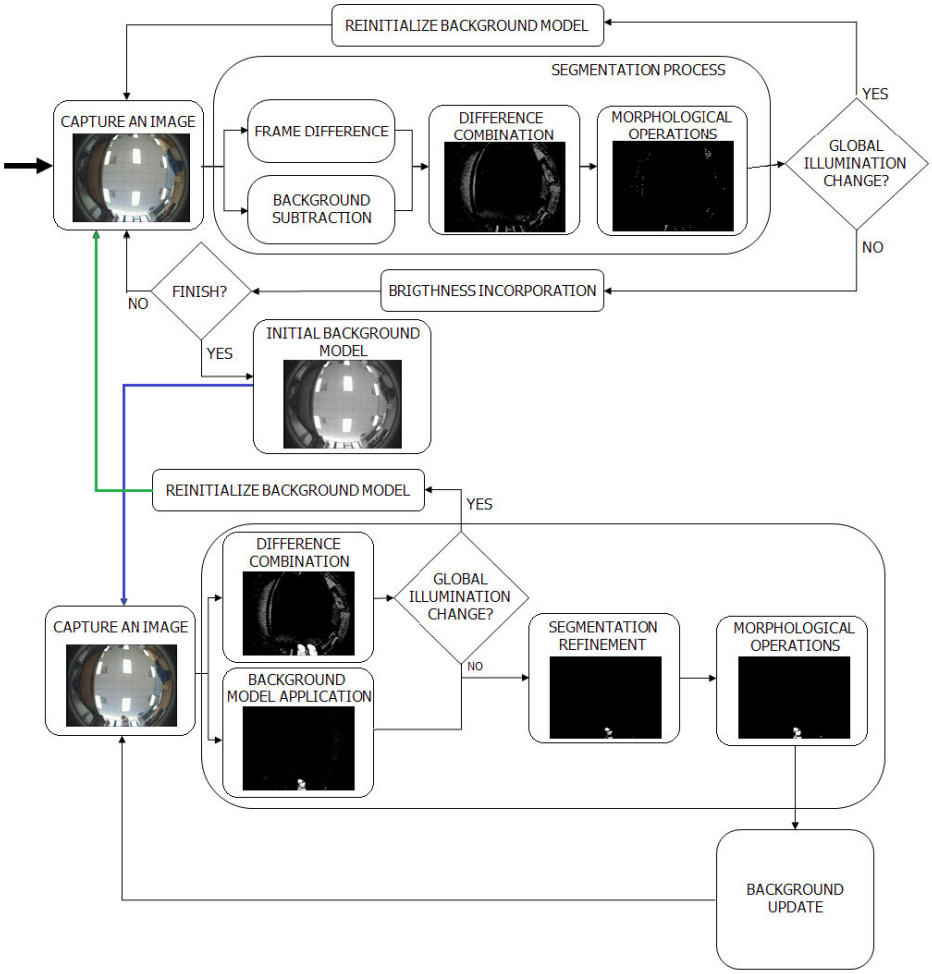

FIGURE 2. **Workflow of our approach to detect the robot surrounding elements.**

experimentally set for each pixel since it represents the pixel fluctuation being higher at the image borders. Note that both k_t and $\sigma_t(x)$ require an initial non-zero value for a proper performance. So, when no information is available for a pixel, a predetermined value is set. These variables are updated together with the background model from those pixels classified as background as follows:

$$\begin{cases} \mu_t(x) = \begin{cases} (1-\alpha)\,\mu_{t-1}(x) + \alpha i_t\,(x) & \text{if background} \\ \mu_{t-1}(x) & \text{otherwise} \end{cases} \\ \sigma_t(x) = \begin{cases} (1-\alpha)\,\sigma_t(x) + \alpha\,(i_t\,(x) - \mu_t)^2 & \text{if background} \\ \sigma_t(x) & \text{otherwise} \end{cases} \end{cases} \tag{2}$$

The α corresponds to the model learning rate and its value is updated in each frame as $\alpha = 1.0/N$ being N the number of frames used so far. Given that the higher N is, the lower α is, a high amount of frames can lead to a bad adaptation of the background model. As a solution, a new background model is built after 200 frames.

Different experiments were carried out in order to evaluate the approach performance. Firstly, the performance was deeply analysed by using Wallflower [11], a well-known dataset for video surveillance systems. This open source dataset is composed of seven image sequences covering possible critical situations for motion detection: **bootstrapping**, where all the frames contain foreground elements; **camouflage**, where a person is walking in front of a monitor that has rolling interference bars (similar to the person's clothing) on the screen; **time of day**, where a scene is observed along a day suffering from gradual illumination changes; **light switch**, where the lights of a room are continuously switched on and off; **waving trees**, where a person is walking in an outdoor scene with a swaying tree; **foreground aperture**, where a person with uniformly colour shirt wakes up and begins to move slowly; and **moved object**, where a person enters into a room, makes a phone call and leaves, so that the phone and the chair are left in a different position.

Each of these image sequences is provided together with a hand-segmented ground truth image for evaluation. This fact allows a quantitative performance comparison between approaches. In particular, in this case, *recall*, *precision* and *accuracy* measurements have been used. So, *recall* refers to the ratio of the number of foreground pixels correctly

TABLE 1. Quantitative comparison of several state-of-the-art approaches for motion detection by using the Wallflower dataset [11].

Approach	Recall	Precision	Accuracy
GMM [12]	70.28	52.75	79.87
GMM + EIC [13]	61.77	87.17	89.72
Normalised block correlation [14]	48.11	68.25	83.87
Temporal derivative [15]	62.91	34.2	65.65
Bayesian decision [16]	65.00	47.23	76.62
Eigenbackground [17]	79.58	66.48	86.85
Wallflower [11]	76.23	78.43	90.28
Tracey LAB LP [18]	68.14	87.92	91.05
RGT [19]	76.24	59.03	83.34
RGT-Euc [19]	83.84	52.85	80.24
Joint difference [20]	80.77	81.76	91.90
Our approach	**92.39**	**94.05**	**97.01**

identified to the number of foreground pixels in the ground truth; *precision* represents the ratio of the number of foreground pixels properly identified with respect to the number of foreground pixels detected; and *accuracy* indicates how well the segmentation process identifies or excludes the foreground pixels. Table 1 presents a quantitative comparison with state-of-the-art approaches that have provided results using this dataset. Those approaches can be briefly summarised as follows:

- **Mixture of Gaussians (GMM)** [12]. A pixel-wise mixture of three Gaussians models the background such that each Gaussian is weighted according to the frequency with which it explains the observed background
- **GMM + Effective Intensity Change (EIC)** [13]. This approach is based on GMM, although their learning rate is dynamically set during the video analysis. For that, a new parameter called Effective Intensity Change occupancy (EIC), is introduced. This parameter extracts background dynamics for each frame and is used to estimate the new value of the learning rate at any time
- **Normalised block correlation** [14]. In this approach, images are split into blocks such that each block is represented as its median and the standard deviation of the block-wise normalised correlation over the training images. For each incoming block, normalised correlation values that deviate too much from the expected deviations cause the block to be considered foreground
- **Temporal Derivative** [15]. In this case, the minimum and maximum inter-frame change in intensity is obtained for each pixel during the training phase. So, any pixel that deviates from its minimum or maximum by more than the maximum inter-frame change is considered foreground. They additionally enforced a minimum inter-frame difference of 10 pixels after the regular training phase
- **Bayesian Decision** [16]. This approach is based on pixel value probability densities accumulated over time represented as normalised histograms. Thus, backgrounds are determined by a straightforward maximum *a posteriori* criterion
- **Eigenbackground** [17]. In this approach, the first step is to collect images of motionless backgrounds. Then, Principle Component Analysis (PCA) is used to determine means and variances over the entire sequence (whole images represented as vectors). So, the incoming images are projected onto the PCA subspace and the differences between the projection and all the current images greater than a threshold are considered foreground
- **Wallflower** [11]. In this approach, the input images are processed at three different spatial scales: pixel level, which makes the preliminary classification foreground-background and the adaptation to changing backgrounds; region level, that refines the raw classification of the pixel level based on inter-pixel relationships; and frame level, designed for dealing with the light switch problem
- **Tracey LAB LP** [18]. In this approach, the background is represented by a set of codebook vectors locally modelling the background intensities in the spatial-range domain such that the image pixels not fitting that model are classified as foreground. In addition, as in the Wallflower algorithm, a frame-level analysis is used to discriminate between global light changes, noise, and objects of interest. Moreover, the foreground is also represented by a set of codebook vectors in order to obtain a more accurate foreground segmentation
- **RECTGAUSS-Tex (RGT)** [19]. In this approach, the image processing is carried out at region level, where the background is modelled at different scales by using the colour histogram and a texture measurement. Thus, motion is detected by comparing the corresponding rectangular regions from the coarsest scale to the finest one such that the comparisons are done at a finer scale only if some motion was detected at a coarser scale. Furthermore, a Gaussian mixture background subtraction in combination with Minimum Difference of Pair Assignments (MDPA) distance is used at the finest scale
- **RECTGAUSS-Tex-Euclidean (RGT-Euc)** [19]. This is a modification of the previous approach such that Euclidean distance is used instead of MDPA distance
- **Joint Difference** [20]. In this approach, motion is detected by means of a hybrid technique that uses both frame-by-frame difference and background subtraction. This technique integrates a selective updating method of the background model to tune background adaptation. In addition, a shadow filter in the *HSV* colour space is used to improve the motion detection

As shown in Table 1, the proposed approach overcomes the performance of the state-of-the-art approaches, getting results greater than 90% in all the considered measurements. In particular, the obtained accuracy is over 97%, the highest value, which means that the proposed approach is the one with the lowest classification error. In a similar way, the good results for recall and precision highlight its rigorous capacity for foreground pixel classification.

In addition, two fisheye cameras were mounted on a mobile platform that was located at different positions in

our laboratory. This set-up was used to detect and track a group of individuals (going from 1 to 5 depending on the timestamp) moving around the robotic system. With regard to the obtained experimental results, all the individuals were successfully detected wherever they were located with respect to the robotic system (up to a distance of 20 metres).

The second task consists of identifying and tracking the moving elements around the robot. However, it could be difficult especially when they meet, form groups or cross-over. So, it is necessary to define a target representation that makes the data association robust and accurate. So, the first step is to obtain that proper representation. In this sense, research has taken a number of forms focusing on human representation (e.g. [21], [22]). However, despite its good results under some conditions, they present some drawbacks such as the need to know a model *a priori*, only working on a type of device, failing in case of elements with a rich variability, being computationally intensive, or requiring several constraints on the environment like being uniform or static, or on the elements like presenting similar colour histograms. Therefore, the representation to develop should:

- identify an element among a broad range of elements as well as when they leave and re-enter the scene
- be robust to partial occlusions
- be open to learn new elements
- be obtained from a minimum number of training images
- provide a response time insensitive to the number of elements to be tracked and identified

Keeping in mind these goals, a new representation has been designed. Basically, each element is represented by means of a data structure composed of:

- an image pattern
- a feature array whose elements contain different kind of information (detailed below) used to properly match images of the same object in two consecutive frames

In regard to the image pattern, given that a fisheye camera pointing upward is used, the first issue to be overcome is the different orientation of the elements due to their position within the image. Due to the impossibility of comparing two consecutive images of a target in all the possible orientations, a panoramic image is obtained. For that, a correspondence between the fisheye image and the panoramic one is established [23]. As shown in Figure 3, the fisheye image is considered as a torus region that can be *cut* through Y-axis in order to stretch to a rectangle such that each fisheye pixel within a bounding rectangle is converted by using the centre coordinates of the fisheye image (x_0, y_0), and the minimum and maximum radius determined from the corners of the bounding rectangle (R_{min}, R_{max}). Note that only the bounding boxes of the detected elements are converted to panoramic images due to computational reasons.

With respect to the feature array, it describes each element as an identifier descriptor (ID), the coordinates of its gravity centre to generate the followed trajectory, and the number of frames that it has not been seen in the scene.

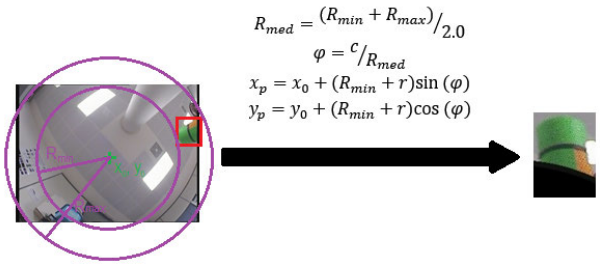

FIGURE 3. Graphical description of the transformation from a fisheye camera to its corresponding panoramic image.

FIGURE 4. Some results of the tracking process with the robot head cameras.

This representation is used to properly track each element by following a modified nearest neighbour approach. In particular, spatial information is combined with the element representation such that an element history together with a feature- and pixel-similarity likelihoods have been defined. This definition allows the robot to properly track the several elements within the scene even if they leave and re-enter the scene, meet or cross-over as shown in Figure 4.

The last task is to properly locate the target elements (i.e. the person or robot to collaborate with) for both safety reasons and collaboration. Since the overlapping area of the two fisheye cameras is too tight, the stereo system of the robot head is used to locate the surrounding elements. That is, the robot head is oriented in the direction of each detected element to accurately estimate the distance to the robot. Note that a traditional RGB stereo camera is used instead of a Microsoft Kinect camera or a similar device since they are imprecise in distances lower than 50 cm, what could be critical when a collaborative task is performed. As deeper explained in [24], the distance is estimated by means of a biologically-inspired approach. Basically, this approach can be summarised as follows (see Figure 5: the early vision area (V1/V2) is modelled as a set of complex Gabor filters with a cosine-based real part and a sine-based imaginary part. This processing at multiple scales and orientations, results in a quantitative

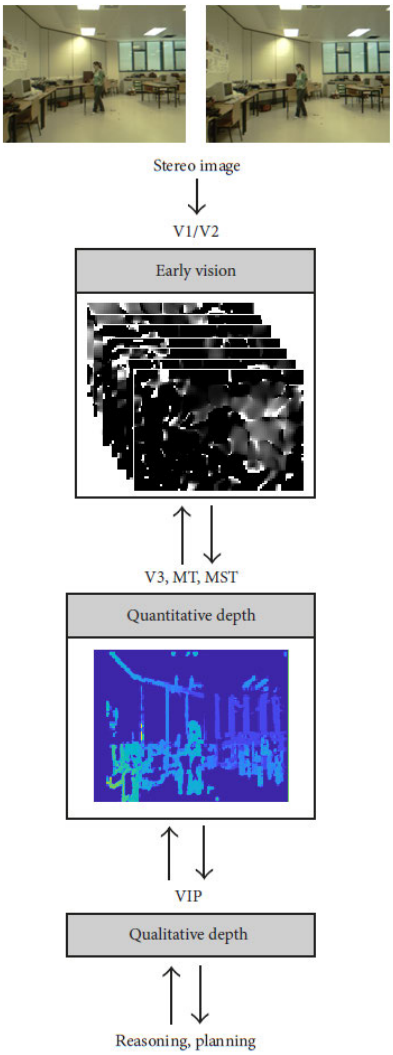

FIGURE 5. Workflow of our approach for estimating the surrounding element depth.

stereo disparity estimation that leads to a quantitative depth map. From this knowledge, an egocentric representation of the element localisation within the scene is obtained by following a qualitative approach. Finally, a qualitative reasoning method allows the system to infer new information and make decisions more accurately.

III. ROBOTS IN SEMI-STRUCTURED ENVIRONMENTS

Going a step further, semi-structured environments provide a perfect starting point for autonomous service robots. In particular, two semi-structured environments have been considered: a library and a warehouse. As illustrated in Figure 6, both environments present a strong topological structure composed of strictly arranged shelves, while coexisting with people and other robots. In addition, in both cases, the tasks to be performed are similar: require an item, navigate to the proper shelf, identify it, grasp it and deliver to the person or robot who requests it. Nevertheless, the main difference between them lies in the items to be recovered. So, while these

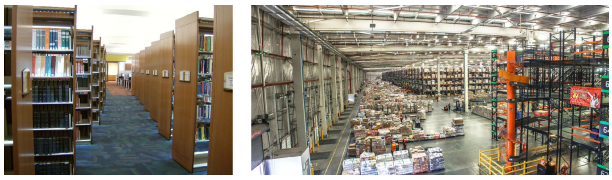

FIGURE 6. Semi-structured environments: a library (left) and a warehouse (right).

are books in the case of the library, they are market products in the warehouse. This fact results in an analysis of the visual features to be detected to properly identify the required item in each scenario.

A book can vary in size, thickness, colour, and title style on the spine. These visual features make its recognition difficult given the great amount of books in a library and recent techniques like deep learning should learn each and every one of them, what is high time-consuming. As an alternative, Radio Frequency Identification (RFID) systems could be used as in [25], [26]. However, all the books in the range of the RFID reader will be recognised what makes it inappropriate to identify just one book within the bookcase. As a solution, a new vision approach has been designed.

In our particular case, the books are classified according to the Library of Congress Classification (LCC). So, each book is tagged with a book code or signature. This code is composed of the class number and the book number, what sets its arrangement within the library and can help the robot locate a book in a bookcase. Thus, the arisen issues in this case are:

- separate book by book
- extract each white tag containing the book code
- recognise the book code in a proper way

With the purpose of overcoming these issues, we have developed and implemented a new vision approach based on traditional computer vision techniques. This approach, integrated in the UJI librarian robot [27], can be summarised as follows (see Figure 7): the robot arm is located in front of a shelf to take an image. Note that the position is determined based on the detection of the shelf basis. Then, the image is cropped keeping the bottom part since the tags are always located at the bottom part of the book spine. Then, a threshold together with Canny detector are used to detect edges. The Hough transform allows the robot to properly detect the vertical lines separating each book from the next one. The last two steps are repeated in search of the horizontal lines delimiting the book tags. The final step consists in combining the horizontal and vertical lines to properly extract the book tags.

Once the book tags are extracted, they are sent to an Optical Character Recogniser (OCR) to read the book codes. As the correct identification of the book code is crucial for the proper book identification, an image from another point of view can be required when the OCR fails in the recognition. If a positive match is provided, the robot manipulator is accurately located to be able to safely extract the requested book.

FIGURE 7. Our vision approach to separate book by book and extract each tag with the book code.

For its part, the warehouse problem requires the learning of more visual features since market products not only can vary in size, thickness and colour, but also in shape, material, and opacity. Given the challenging nature of this task, Amazon launched the *Amazon Picking Challenge* in 2015. The goal of this robot research project (2015-2017) was to automate the pick-and-place task in a warehouse. For that, two tests were designed:

1) **Pick test**: 32 items are placed in a storage system and the designed robot system must be able to pick 10 of those items and put in their correct box in 15 minutes
2) **Stow test**: the designed robot system must pick items (one by one) from a tote with 20 items, and place them in their corresponding location in the storage system in less than 15 minutes

Note that the participating teams were provided with 40 items to train and test their robotic designs. However, during the competition, they were given with new products to be learnt in a short period of time. This fact makes the task more complex and results in a need to reinforce some approaches to speed up the learning process, although this additional step could be avoided in real settings.

Keeping in mind all the visual features to be considered, deep learning techniques could fit. In particular, Convolutional Neural Networks (CNNs) have been proven a good performance in object classification tasks. Thus, this approach, inspired in the human visual cortex, uses multiple layers to provide a classification tag from an input image. One of the most popular approaches is the Residual Network (ResNet) [28]. This architecture is mainly based on the VGG network [29] by reformulating the layers as learning residual functions. Given its good generalization performance, it was part of our implementation for the *Amazon Picking Challenge* in 2017 [30] (see Figure 8). However, the learning of a new object is too time consuming. Therefore, in the context of the competition (and when it is required a quick learning stage), an alternative technique should be used for new objects, as previously pointed out. In our particular case, as most of the objects were textured, the Scale-Invariant Feature Transform (SIFT) was used [31]. This approach shares many features with neuron responses in primate vision such that a 4-stage filtering process provides a feature object description. This keypoint descriptor is then used to properly match two images of the same object by identifying its nearest neighbour.

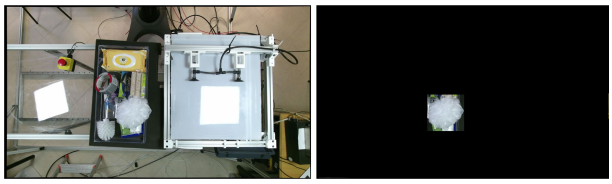

FIGURE 8. Sample of ResNet performance in a sponge detection.

IV. ROBOTS IN THE REAL WORLD

The last level of this evolution corresponds to assistive robots such that they are able to autonomously assist humans in their daily lives. In this case, real-life scenarios are considered. As a consequence, there are no constraints about the environment. This fact raises new challenges in the tasks to be accomplished by the robots. In this context, the ability to autonomously manipulate objects is of critical importance. For that reason, the detect-approach-grasp loop for object manipulation requires a robust recovery stage, especially when the held object slides. Although some devices have been developed for that purpose (e.g. tactile sensors, contact switches, or proprioception sensors), the robot gripper's dexterity and functionality can be considerably limited. In addition, neural network approaches were discarded due to the need for prior learning of both the robot gripper and the grasped object, what considerably restricts the robot tasks and the objects to interact with.

As a solution, a novel vision approach was presented in [32]. As illustrated in Figure 9, simple visual features such as colour, depth and edges are combined for object manipulation supervision such that a contact between a robot gripper and any grasped object could be detected. More precisely, the process starts with the capture of an RGB-D image. Then, the RGB image is converted into an Lab image that is colour-segmented based on the Lab coordinates corresponding to the robot gripper. On the other hand, the depth information is in charge of detecting the contact points between the robot gripper and the held object. For that, an edge detection based on depth difference is performed. Under the assumption that the robot gripper always emerges from the bottom of the image, an edge refinement takes place. The resulting edge image is combined with the colour-segmented one such that the contact points can be accurately extracted.

On the other hand, the proper object detection and recognition also plays a main role. Nevertheless, unlike the warehouse case, there is no knowledge about all the objects to interact with in the real world. In fact, a learning process must be added to successfully deal with new objects. This learning process used to involve a human-robot interaction where the user must provide the robot with the identification tag for an object. The next step is to obtain the corresponding object representation for its further recognition. In this context, deep

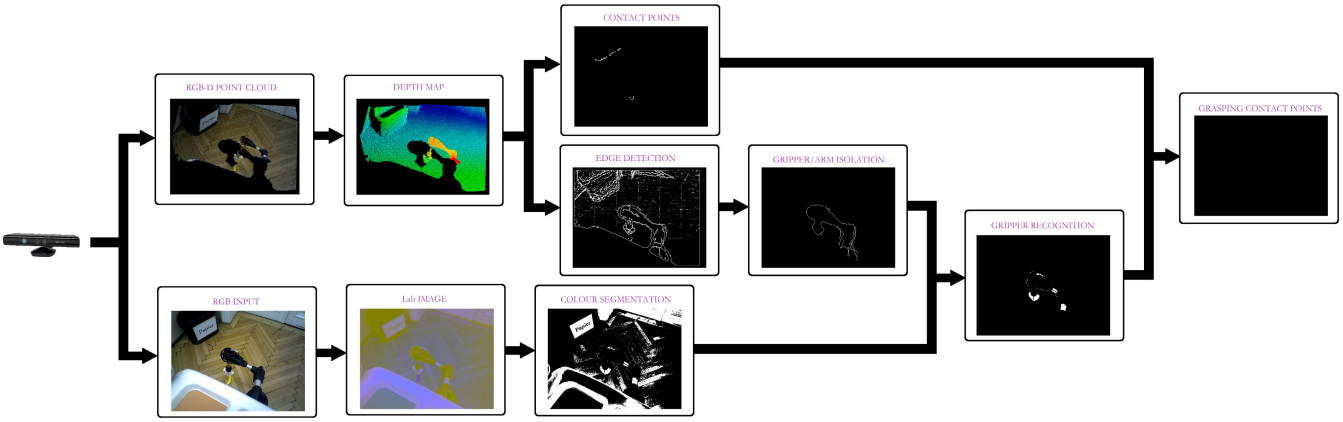

FIGURE 9. Workflow of our approach for error detection in manipulation tasks.

learning techniques are discarded due to its time-consuming learning stage. Furthermore, the necessary data for the proper object recognition in a real scenario could not be always available by requiring more than one learning process for the same object.

As a solution and based on neurological findings, a novel approach was presented in [33]. For that, three primary cues are considered: colour, motion and shape. Aimed to invariability in front of illumination changes and surface orientation, the $l_1l_2l_3$ colour space was chosen. The same approach for motion detection described in Section II is used to properly perceive moving elements. Regarding the shape, a biological approach based on Gabor filters is considered. As object shape changes depending on the point of view, different representations are required. Nevertheless, after an extensive experimental analysis, it was concluded that four views of an object could be enough: top, bottom, sideways and perspective. Therefore, an *8-Gabor* shape representation for each view is generated. With this data, an object can be accurately recognised in cluttered scenarios. In addition, the time to learn a new object is short due to the simplicity of its representation.

So, the whole object recognition approach, illustrated in Figure 10, can be described as follows: when a robot is looking for an object, a visual scrutiny is performed. So, for each taken image, two processes take place. On the one hand, the RGB image is converted to the $l_1l_2l_3$ colour space and, based on the $l_1l_2l_3$ coordinates of the target object, an image segmentation is performed. On the other hand, the gray-scale version of the taken image is obtained. From that, two images are obtained: the result of applying the motion detection approach and the shape-based segmentation. Note that the background model could not cover the whole image since it is built during the visual scrutiny where the robot cameras are continuously moving. The three segmented images are combined based on the object likelihoods for each considered cue. It is worth mentioning that this probabilistic combination varies for each object since the type of object determines what visual features are more distinctive.

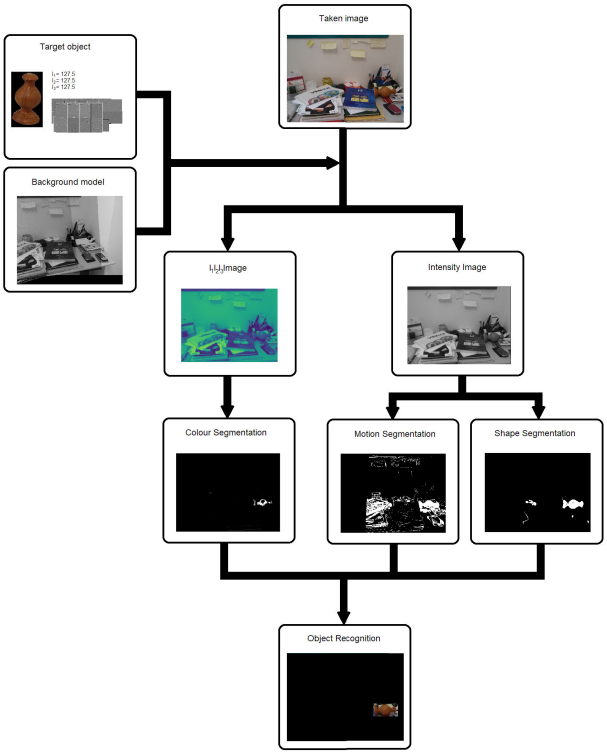

FIGURE 10. Workflow of our approach for object recognition in real scenarios.

V. DISCUSSION

The trip along robotics research and applications described in this paper highlights the fact that different scenarios result in different requirements for robot tasks, especially in terms of constraints about the targets and/or the environment. So, the issues to be overcome for each proposed scenario have been increased in number and difficulty.

Thus, in the case of industrial settings (a structured environment), most of the environmental conditions are under control. This fact leads to a problem simplification focused on the targets to work with. In particular, in this paper,

a human-robot collaboration task has been analysed. For that, the main goal was to design a system able to recognise and localise people within the robot work-space so that the person to interact with is identified and tracked, while the surveillance of the space around the robot guarantees the safety of all the remaining surrounding elements such as other persons or robots. With that purpose, a vision system has been presented. This system was designed taken into account the following requirements: the coverage of all the surrounding robot space; a robust detection of all the surrounding elements (i.e. people and other robots) based on motion detection by considering that those elements can stop, group, leave and re-enter the scene at any time; the avoidance of false positives due to minor dynamic factors such as illumination changes or blinking screens, so that the industrial activity does not stop unnecessarily; and the tracking of those elements, as well as their 3D localisation, with the aim to guarantee the safety of all of them during the robot task performance. Despite its good experimental results, the proposed approach has some limitations to be addressed. This approach has been designed for *static* robots. So, several stops would be required to apply this vision system to mobile robot platforms. In addition, qualitative depth estimation provides a region-based location what could lead to needlessly stop the activity when a person or robot is in the limit of the safe area.

Regarding semi-structured environments, two different applications have been presented. First, a librarian robot aimed to identify a book within a bookcase. In this case, based on the book cataloguing in the libraries, the issues to deal with are: the difference in visual appearance of the books; the identification of the vertical edges of all the books on the considered bookshelf, allowing the vision system to properly separate each book; the recognition of the tags attached to each book for its cataloguing; and the reading of the alphanumeric code in order to identify the searched book. In addition, other factors like changes in illumination, shadows and other environmental conditions were taken into account to get a purposeful application. From this starting point, a series of traditional Computer Vision techniques were combined to achieve the final application goal. In this case, the main limitation lies in the text recogniser (i.e. OCR) since an open source software was used and the error rate could be too high under certain conditions.

Second, a warehouse application has been presented. Although the basic problem is similar, the challenges to be tackled differ considerably since, for instance, the market products to be distinguished vary in their visual features much more than books in a library. In addition, illumination conditions can be poor due to the shelf design which, together with other environmental conditions, can considerably influence the recognition process. In this case, deep learning techniques were used to tackle these issues; ResNet, in particular. Note that, as the system knows *a priori* the market products stored in the warehouse, the data required for its proper learning can be generated. Nevertheless, this learning process is the bottleneck of the approach, since the data required for the learning must be generated for each new market product, and a training phase takes place each time a new product is introduced, a process that can be very time consuming.

When talking about the real world, a wide range of robot applications can be proposed. In this paper, two of them have been analysed. The first one refers to an application for robot grasping supervision, aimed to error recovery when the grasped object is slipped or lost. It is important to highlight that, given that no constraints about the robot gripper or grasped object are established, the complexity of the problem has been considerably increased. In addition, deformable objects and robot grippers were also considered. This fact is critical for the resulting application, since state-of-the-art grippers may undergo different deformations to better adapt to a particular object. In consequence, deep learning techniques were discarded because all the possible deformations cannot be modelled *a priori* for its learning. From this starting point, colour, edges and depth cues are combined to successfully detect grasping errors. A number of experiments (described in more detail in [32]) with different robot grippers interacting with a variety of objects demonstrate the good performance of the proposed approach by obtaining an accuracy of 97.5 %. Nevertheless, small or thin objects can lead to a failure. Therefore, the approach should be improved to properly deal with this kind of objects without constraining the robot's autonomy.

The last presented application is object detection for robot manipulation. As in the previous application, no constraints about the object and robot manipulator are specified. This fact led us to discard deep learning techniques; the main reason is the time-consuming learning stage, since each new object to interact with would result in another training stage that, in turn, would require a large amount of object data in different orientations, positions, and scales to be properly learnt. In fact, this is one of the issues to be addressed: the implementation of an object representation that is easy and quick to learn. Another important issue is the recovery of the object's shape since this is an essential information for properly grasping it. In this sense, deep learning techniques like YOLO (You Only Look Once) or ResNet (Residual Network), are not adequate, since they only provide a bounding box enclosing each detected object; this bounding box may include a lot of redundant space, as shown in [34]. Moreover, partial occlusions are represented by overlapping the bounding boxes and, consequently, another computer vision technique is required to detect the occluded part (if it is the case) and recover the object's shape. Again, environmental factors must be also taken into account to be successful in the goal task. Keeping all these issues in mind, a biologically-inspired approach was proposed. In particular, colour, motion and shape cues are combined. As extensively described in [33], three different experiments were carried out for semi-structured scenes, real scenarios, and image repository. In all these experiments, the object detection was successful, achieving an accuracy of 96.1 % when the image repository was used. This result

substantially improves the performance of other state-of-the-art classification approaches.

In a nutshell, the lessons learnt after this trip could be summarised as follows:

- The type of scenario and the task to be performed considerably affect the required solution due to the restrictions on the environmental conditions and/or the elements to interact with
- There is no a unique solution for all the robotic tasks since each robot task sets out different issues and requirements
- Deep learning techniques are not always applicable to object recognition despite its good experimental results. The main reasons are:
  - the lack of enough data for properly training the neural network
  - the huge amount of existing objects when real-world scenarios are considered
  - the continuous changes in the shape of deformable objects
  - the need of accurately identify a specific object within an object class, since neural networks are aimed to generalise the object visual features to properly classify it into its corresponding class
- Illumination together with other dynamical environmental factors can considerably influence the recognition tasks, making them fail miserably
- Neurobiological findings may help in the design of purposeful solutions

VI. CONCLUSION

The Robotics evolution to autonomous service robots from the industry has required a continuous research to properly deal with the multiple arisen issues. In fact, each new scenario or task to be performed set out a series of challenges to be faced up with.

This paper presents an overview of these requirements through several developments going from the industry to the real world. So, this trip starts with industrial settings where a robot is endowed with a vision system to guarantee the safety of all the surrounding elements while robot tasks are performed, specially when they are collaborative. It also includes a human tracking and recognition module. In this case, issues such as how to cover the whole robot work-space, how to detect people even when they stop, or how to deal with background minor dynamic factors, were overcome.

Moving to semi-structured environments, the conditions are less restricted and controlled. This fact led to deal with other challenging situations. In particular, we have studied two different scenarios: a library and a warehouse. Given the similarity between them in terms of topological distribution, we have focused on the proper recognition of the items to be recovered and reinstated in the shelves. Regarding the library, the requested item will be always a book, although it can vary in colour, size and title text style. Thus, a vision system to distinguish book by book by looking for vertical divisions is designed. Then, the correct book is recognised based on its LCC code, attached to each book, used for book classification in the library under study. A richer recognition system is provided when a robot works in a warehouse since the market products can vary in a wide range of visual features. In this case, deep learning techniques have been used, although traditional computer vision methods must be integrated when an item needs to be learnt in a short period of time.

Finally, when real scenarios are considered, robots must be endowed with those sensory motor skills to work in an autonomous and reliable way. So, two main tasks have been considered: the detection of an error during object manipulation for a quick and safe recovery; and object recognition in cluttered scenes. In both cases, traditional approaches together with sensor fusion have been combined to successfully achieve the goal.

This overview brings to light the increase in complexity in the designed approaches whenever the environmental restrictions are less limited and controlled. In addition, these approaches also change depending on the goal task or the considered type of scenario, as it is the case of object recognition.

ACKNOWLEDGMENT

The research conducted at UJI Robotic Intelligence Laboratory.

REFERENCES

[1] G. Rodríguez-Canosa, J. del Cerro Giner, and A. Barrientos, "Detection and tracking of dynamic objects by using a multirobot system: Application to critical infrastructures surveillance," *Sensors*, vol. 14, no. 2, pp. 2911–2943, Feb. 2014.

[2] S. Baker and S. K. Nayar, "A theory of catadioptric image formation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*. New Delhi, India: Narosa Publishing House, Jan. 1998, pp. 1–8.

[3] S. Baker and S. K. Nayar, "A theory of single-viewpoint catadioptric image formation," *Int. J. Comput. Vis.*, vol. 35, no. 2, pp. 175–196, 1999.

[4] R. W. Wood, "Fish-eye views, and vision under water," *London, Edinburgh, Dublin Philos. Mag. J. Sci.*, vol. 12, no. 68, pp. 159–162, 1906.

[5] N. Kruger, P. Janssen, S. Kalkan, M. Lappe, A. Leonardis, J. Piater, J. A. Rodriguez-Sanchez, and L. Wiskott, "Deep hierarchies in the primate visual cortex: What can we learn for computer vision?" *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1847–1871, Aug. 2013.

[6] T. Mahalingam and M. Subramoniam, "A robust single and multiple moving object detection, tracking and classification," *Appl. Comput. Informat.*, Jul. 2020.

[7] K. Sehairi, F. Chouireb, and J. Meunier, "Comparative study of motion detection methods for video surveillance systems," *J. Electron. Imag.*, vol. 26, no. 2, Apr. 2017, Art. no. 023025.

[8] S. Manchanda and S. Sharma, "Analysis of computer vision based techniques for motion detection," in *Proc. 6th Int. Conf.-Cloud Syst. Big Data Eng. (Confluence)*, Jan. 2016, pp. 445–450.

[9] I. Markovic, F. Chaumette, and I. Petrovic, "Moving object detection, tracking and following using an omnidirectional camera on a mobile robot," in *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, May 2014, pp. 5630–5635.

[10] E. Martinez-Martin and P. A. del Pobil, *Robust Motion Detection in Real-Life Scenarios*. London, U.K.: Springer, 2012.

[11] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers, "Wallflower: Principles and practice of background maintenance," in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, Sep. 1999, pp. 255–261.

[12] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 1999, pp. 246–252.

[13] L. Alandkar and S. Gengaje, "Novel adaptive learning scheme for GMM," in *Proc. Int. Conf. Wireless Commun., Signal Process. Netw. (WiSPNET)*, Mar. 2017, pp. 850–855.
[14] T. Matsuyama, T. Ohya, and H. Habe, "Background subtraction for non-stationary scenes," in *Proc. 4th Asian Conf. Comput. Vis.*, Singapore, Jan. 2000, pp. 662–667.
[15] I. Haritaoglu, D. Harwood, and L. S. Davis, "W^4: Real-time surveillance of people and their activities," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 809–830, Aug. 2000.
[16] H. Nakai, "Non-parameterized Bayes decision method for moving object detection," in *Proc. Asian Conf. Comput. Vis. (ACCV)*, 1995, pp. 447–451.
[17] N. M. Oliver, B. Rosario, and A. P. Pentland, "A Bayesian computer vision system for modeling human interactions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 831–843, Aug. 2000.
[18] D. Kottow, M. Köppen, and J. R. D. Solar, "A background maintenance model in the spatial-range domain," in *Statistical Methods in Video Processing* (Lecture Notes in Computer Science). Berlin, Germany: Springer, 2004, pp. 141–152.
[19] P. D. Z. Varcheie, M. Sills-Lavoie, and G.-A. Bilodeau, "An efficient region-based background subtraction technique," in *Proc. Can. Conf. Comput. Robot Vis.*, May 2008, pp. 71–78.
[20] D. A. Migliore, M. Matteucci, and M. Naccari, "A revaluation of frame difference in fast and robust motion detection," in *Proc. 4th ACM Int. Workshop Video Surveill. Sensor Netw. (VSSN)*, 2006, pp. 215–218.
[21] G. Zhong, L.-N. Wang, X. Ling, and J. Dong, "An overview on data representation learning: From traditional feature learning to recent deep learning," *J. Finance Data Sci.*, vol. 2, no. 4, pp. 265–278, Dec. 2016.
[22] F. Han, B. Reily, W. Hoff, and H. Zhang, "Space-time representation of people based on 3D skeletal data: A review," *Comput. Vis. Image Understand.*, vol. 158, pp. 85–105, May 2017.
[23] H. Liu, W. Pi, and H. Zha, "Motion detection for multiple moving targets by using an omnidirectional camera," in *Proc. IEEE Int. Conf. Robot., Intell. Syst. Signal Process.*, Oct. 2003, pp. 422–426.
[24] E. Martinez-Martin and A. P. del Pobil, "A biologically inspired approach for robot depth estimation," *Comput. Intell. Neurosci.*, vol. 2018, pp. 1–16, Aug. 2018.
[25] M. Morenza-Cinos, V. Casamayor-Pujol, J. Soler-Busquets, J. L. Sanz, R. Guzmán, and R. Pous, "Development of an RFID inventory robot (AdvanRobot)," in *Robot Operating System (ROS)* (Studies in Computational Intelligence). Cham, Switzerland: Springer, 2017, pp. 387–417.
[26] R. Li, Z. Huang, E. Kurniawan, and C. K. Ho, "AuRoSS: An autonomous robotic shelf scanning system," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2015, pp. 6100–6105.
[27] E. Martinez-Martin, G. Recatala, and P. A. del Pobil, "Transforming library operation with robotics," in *Proc. IFLA WLIC-Satell. Meeting Robots Libraries, Challenge Opportunity*. Berlin, Germany: Technical Univ. Applied Sciences Wildau, Aug. 2019, pp. 1–5.
[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015, *arXiv:1512.03385*. [Online]. Available: http://arxiv.org/abs/1512.03385
[29] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–14.
[30] A. P. del Pobil, M. Kassawat, A. J. Duran, M. A. Arias, N. Nechyporenko, A. Mallick, E. Cervera, D. Subedi, I. Vasilev, D. Cardin, E. Sansebastiano, E. Martinez-Martin, A. Morales, G. A. Casan, A. Arenal, B. Goriatcheff, C. Rubert, and G. Recatala, "UJI RobInLab's approach to the Amazon robotics challenge 2017," in *Proc. IEEE Int. Conf. Multisensor Fusion Integr. Intell. Syst. (MFI)*, Nov. 2017, pp. 318–323.
[31] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.
[32] E. Martinez-Martin and A. del Pobil, "Vision for robust robot manipulation," *Sensors*, vol. 19, no. 7, p. 1648, Apr. 2019.
[33] E. Martinez-Martin and A. P. del Pobil, "Object detection and recognition for assistive robots: Experimentation and implementation," *IEEE Robot. Autom. Mag.*, vol. 24, no. 3, pp. 123–138, Sep. 2017.
[34] A. Mallick, A. P. del Pobil, and E. Cervera, "Deep learning based object recognition for robot picking task," in *Proc. 12th Int. Conf. Ubiquitous Inf. Manage. Commun. (IMCOM)*, Jan. 2018, pp. 1–9.

ESTER MARTINEZ-MARTIN (Senior Member, IEEE) received the Ph.D. degree in engineering (robotics), the master's degree (Secondary Education, Vocational Training and Language Teaching), and the master's degree in mobile and video games programming from Jaume-I University, in 2011, 2013, and 2017, respectively. She has been a Computer Science Engineer, since 2004. She is currently an Associate Professor with the University of Alicante. She has published 12 JCR-indexed articles and several articles in Scopus-indexed journals, a research book (Springer), several book chapters, three research books (as a co-editor), and more than 25 articles in congresses, both national and international. She has participated in several research projects of community, national, and European nature. She has also done four research stays in prestigious foreign centers, including the Università degli Studi di Genova (Prof. Silvio Sabatini), Sungkyunkwan University (Prof. Suthan Lee), the Universidade do Minho (Prof. Paulo Novais), and the Technische Universitat (TU) Wien (Prof. Markus Vincze), all of which are financed in competitive public calls. Her research interest includes the use of vision in robotic tasks, such as object detection and action recognition. She is a member of AERFAI. She has been the Organization Chair of the IEEE-RAS Summer School on Experimental Methodology, Performance Evaluation and Benchmarking in Robotics, in 2015, the 13th International Conference on the Simulation of Adaptive Behavior (SAB 2014), and the 12th International AERFAI/UJI Robotics School on Perceptual Robotics for Humanoids (IURS 2012). She was also a co-organizer of some tutorials in international conferences, including HRI, in 2017, ICINCO, in 2014, IAS-13, in 2014, and the IEEE RO-MAN, in 2013. She has been a member of the programme committee and organization committee in several national and international conferences. She is a regular reviewer of JCR journals and international congresses.

ANGEL P. DEL POBIL (Member, IEEE) received the B.Sc. degree in physics and the Ph.D. degree in engineering from the University of Navarra, Pamplona, Spain, in 1986 and 1991, respectively. He is currently a Professor with Jaume I University (UJI), Castellón de la Plana, Spain, where he is also the Founding Director of the UJI Robotic Intelligence Laboratory. He is also a Visiting Professor with Sungkyunkwan University, Seoul, South Korea. He has been a principal investigator of 38 research projects. He has over 300 publications, including four authored and ten edited books. He is the Co-Chair of the IEEE RAS Technical Committee on Performance Evaluation and Benchmarking of Robotic Systems and a member of the Governing Board of the Intelligent Autonomous Systems (IAS) Society. He was a member of the European Robotics Research Network of Excellence (EURON), from 2001 to 2009. He was a Co-Organizer of over 50 workshops and tutorials at ICRA, IROS, RSS, ROMAN, IJCNN, and HRI. He has been the Program Chair or the General Chair of international conferences, such as Adaptive Behaviour (SAB 2014) or Artificial Intelligence and Soft Computing. He serves regularly as an Associate Editor for ICRA and IROS. He serves on the program committee of over 200 international conferences, such as IJCAI, ICPR, IAS, SAB, ICDL-EPIROB, and ROMAN. He has been involved in intelligent robotics research for the last 30 years. He has been an invited speaker of 77 tutorials, plenary talks, and seminars in 15 countries. He serves as an associate or a guest editor for 12 journals. He has supervised 16 Ph.D. thesis, including winner and finalists of the Georges Giralt PhD Award and the Robotdalen Scientific Award.