Object Detection and Recognition for Assistive Robots

Article *in* IEEE Robotics & Automation Magazine · March 2017

DOI: 10.1109/MRA.2016.2615329

2 authors:

Ester Martínez
University of Alicante

Angel P. del Pobil
Universitat Jaume I

Object Detection and Recognition for Assistive Robots

Ester Martinez-Martin, *Member, IEEE,* and Angel P. del Pobil, *Member, IEEE*

***Abstract*—Technological advances are currently being directed to assist the human population in performing ordinary tasks in everyday settings. In this context, a key issue is the interaction with objects of varying size, shape and degree of mobility. Consequently, autonomous assistive robots must be provided with the ability to process visual data in real time so that they can react adequately for quickly adapting to changes in the environment. Reliable object detection and recognition is usually a necessary early step to achieve this goal. In spite of significant research achievements, this issue still remains a challenge when real-life scenarios are considered. In this paper, we present a vision system for assistive robots that is able to detect and recognise objects from a visual input in ordinary environments in real time. The system computes colour, motion and shape cues combining them in a probabilistic manner to accurately achieve object detection and recognition, taking some inspiration from vision science. In addition, with the purpose of processing the input visual data in real-time, a *Graphical Processing Unit* (GPU) has been employed. The presented approach has been implemented and evaluated on a humanoid robot torso located at realistic scenarios. For further experimental validation, a public image repository for object recognition has been used, allowing a quantitative comparison with respect to other state-of-the-art techniques when real-world scenes are considered. Finally, a temporal analysis of the performance is provided with respect to image resolution and number of target objects in the scene.**

***Index Terms*—Object detection, Object recognition, Robot vision systems, Service robots**

I. Introduction

NOWADAYS, robots have found their way from sealed working stations in factories to people's living and working spaces, where they should be able to autonomously perform different services useful to the well-being of humans, such as domestic tasks, healthcare services, entertainment, and education. In particular, with the purpose of improving people's quality of life, especially for the elderly, the field of assistive robotics is becoming increasingly popular. Research is progressing from special-purpose service robots such as autonomous cleaning or transport systems, to multi-functional assistive robots able to integrate diverse abilities such as person detection and tracking, human-robot interaction, reasoning, localization, navigation, object detection and recognition, planning and manipulation. In addition, these assistive robots are expected to operate in a flexible manner, without constraining the environment, and in a reasonable time, while guaranteeing the safety of all their surrounding elements, especially when they are human beings [1] [2].

However, despite the wide research in this area (e.g. Johnny [3], HOBBIT [4], KSERA [5], Cogniron [6], Care-O-Bot [7], HERB [8], Accompany [9], AAL4ALL [10] and many others), the progress in assistive robotics has been relatively slow to date. This is mainly due to the fact that the environments to cope with are dynamic, unpredictable and human-oriented. In addition, depending on the application, long human-robot interactions could miserably fail because of the limited system's autonomy and abilities, as broadly analysed in [11]. Thus, an assistive robot should be provided with a vast set of perception and action capabilities to efficiently perform its goal tasks in real scenarios, while properly interacting with its users along its life. Among all these capabilities, this paper is focused on perception for object detection and recognition, a key task for a meaningful assistance.

In this context, vision is considered a primary cue because of the information it can provide. Actually, vision has been used in numerous robotic applications to successfully achieve a task (e.g. obstacle avoidance for navigation [12], [13], [14], [15], human recognition for Human-Robot Interaction [16], [17], activity recognition for cooperative behaviour [18], [19], [20] and object identification for manipulation [21], [22], [23], to name only a few). However, despite significant achievements, the problem of detecting and recognising objects efficiently and accurately still remains a scientific challenge when real scenes are considered. Apart from a great number of objects in the images, the reasons for this difficulty are to be found in issues such as their interactions and occlusions, along with photometric and geometric variations in pose, size, etc. Furthermore, noise in images, the nature of objects themselves, complex object shapes and illumination changes, make it a hard task. This is becoming still harder with the advent of digital cameras with resolutions of megapixels and frame rates exceeding 100 frames per second, since considerably more data needs to be processed in less time. Therefore, given that a practical assistive robot requires real-time performance, optimized implementations and novel insights are necessary.

Many efforts have been made to overcome these problems. The most habitual way to recognise shapes and objects is by means of model-based approaches [24], [25], [26]. These techniques start by taking a large set of images in different poses and from different viewpoints. From them, an object model is built and learnt in advance. Then, the features extracted from the objects in a scene are matched against features of the previously stored object models. It is important

E. Martinez-Martin and A.P. del Pobil were with the Robotic Intelligence Lab (RobInLab), Universitat Jaume-I, Avda. Sos Baynat, s/n, 12071-Castellón, Spain e-mail: {emartine, pobil}@uji.es.

A.P. del Pobil was also with the Department of Interaction Science, Sungkyunkwan University, Seoul 110-745, Korea

Manuscript received September 16, 2016; revised September 16, 2016.

to highlight that the considered features must be invariant with respect to various transformations (such as view direction, scale and changes in illumination) and also need to be robustly extracted; conditions that can hardly be met in unconstrained environments. Despite being a good procedure for some kind of objects, it is difficult to learn models of objects with a high dimensionality or with a rich variability in their motion, such as human beings. In addition, autonomy is a requirement in assistive robotics and, consequently, no constraints about the object appearance or motion can be established. On the other hand, there exist methods based on local features. In this case, objects are represented via their edges, colour or corner cues [27], [28], [29]; steerable filters [30]; haar-like features [31]; or scale-invariant descriptors (e.g. SIFT, SURF) [32], [33]. These approaches are commonly used for their computational simplicity, efficiency, and robustness to affine transformations. Nevertheless, their accuracy is tightly coupled to the number of features used for describing an object. Also, a trustworthy segmentation for obtaining object features is especially complex when real scenarios are considered. In addition, object features are only relatively robust to small affine transformations, a condition that, again, can hardly be fulfilled when unconstrained scenarios are considered.

Alternatively, the concept of *Object Action Complexes* (OACs) could be used. In this case, objects and actions are assumed to be inseparably intertwined. Thus, OACs are proposed as a framework for representing actions, objects, and the learning process that constructs such representations at all levels, from the high-level planning and reasoning processes to the sensorimotor low-level. Therefore, OACs can act as an interface between the Artificial Intelligence planning and the diverse representation languages for robot control [34]. Moreover, a connection between robot actions and the visual and haptic perception is defined for the interaction objects [35] [36].

The same idea underlies in approaches in which a process to segment interest objects and to extract their shape is based on active visual exploration [37]. Even though the exploration system is completely autonomous, the system still requires a significant amount of *prior knowledge* about the world (in terms of a sophisticated visual feature extraction process in an early cognitive vision system), *knowledge* about its body schema and *knowledge* about geometric relationships such as rigid body motion. That is, it is necessary to know the system's visuomotor map in order to be successful.

The perception-action relationship was also studied from a cognitive point of view [38] [39]. In this case, perception and action are linked through a memory component. Basically, perception allows the system to sense its surroundings with three sensor modalities: audio, vision, and touch. This data is fed into the memory module to produce motor-control signals, that are translated into robot responses by the action unit. In this way, the intermediate mechanism acts as the *robot's brain* by making the recognition task easier. However, despite the vast analysis of existing perceptual systems, the conclusion is that semantic and emotion understanding still remains an open problem. Consequently, in a similar way, robust object recognition still requires much efforts, especially when real scenarios are used. Palomino et al. [40] presented an attention-based cognitive architecture in which reasoning is the bond between perception and action. In this case, the core idea is to select the tasks that will be active at each time based on the context data and the state of achievement of each action. So, depending on the perceived elements, a task can be executed or not since the accomplishment of a task is closely linked to the presence of specific elements in the scene. This system has a high success rate (85%) when only one type of object is used (balls) and the distinctive feature is colour; considerable additional efforts are still required for an object-based visual attention system to accurately detect and categorize a wider range of objects.

New approaches are called for to achieve our goal. In principle, we would like the required knowledge for object detection and recognition to be only obtained from the visual input. From a biological point of view, psychophysics experiments have shown that humans perform some pre-segmentation using boundaries and regions as a previous step prior to actual image understanding [41]. This early segmentation is then tuned by using a huge object database stored in our brains. Thanks to this process, real-life objects can be perfectly recognised even with intense shadows, large occlusions or geometric distortions.

From the same underlying idea and with the purpose of overcoming these problems, a combination of several visual object features can be a promising approach. In this way, colour-based invariant gradients have been combined with Histogram of Oriented Gradient (HoG) local features [42] for object detection in outdoor scenes (such as urban scenes) under cast shadows. The approach is, however, limited by the constrained nature of the environments.

This work is based on our previous ideas on this topic [43]. Motivated by the challenges discussed above, we present new scientific results with a focus on working systems. Indeed, our robot system is capable of detecting and recognising objects from a visual input in realistic, truly unconstrained scenarios in real time. For that, and based on the amazing ability of the human visual system for object identification, the system computes object-specific colour, motion and shape cues and combines them in a probabilistic manner to adequately detect and recognise objects. Moreover, a *Graphical Processing Unit* (GPU) is used to achieve real-time performance in processing the visual data. Extensive experimental validation has been conducted with a humanoid torso and an image repository, as well as a temporal analysis of the performance.

The rest of this article is organised as follows: Section II describes the architecture of the designed system. Section III provides the implementation details. The obtained experimental results are presented and discussed in Section IV, and the guidelines for our future work are introduced in Section V.

II. System Description

From a biological point of view, humans are able to easily identify the objects present in their environment. Therefore, insights from human visual processing could be a starting point for developing computer models. This is the case of Al-Absi and Abdullah [44], who designed BIORecS emulating the human vision system. Concretely, BIORecS achieves accurate object recognition in complex scenarios by combining functions of some areas of the human visual cortex and the connection mechanisms between the visual areas in humans, implemented by feedforward and feedback techniques. This model consists of four stages closely intertwined: feature extraction (object shapes are obtained by combining the image edges extracted with Gabor filters); visual attention (a support vector machine is used as object shape classifier); recognition (carried out by Principal Components Analysis) and image database (containing the objects to be recognised).

However, although this architecture may allow the system to overcome some key issues in object recognition -such as changes in illumination, occlusions and high-cluttered scenes- the description of objects is not adequate since different objects can have the same visual shape. For example, a ball, a bracelet, a disk, a coin or a drum would all belong to the category of circular shape. Furthermore, some factors such as its pose, scene background or illumination conditions may modify the object's shape. Consequently, a model reformulation is necessary.

Alternatively, object detection and recognition could be considered as an attentional mechanism since it refers to the extraction of target information from the observed scene. In this sense, a *dorsal attention system* could fit. Generally speaking, this system could be defined as a top-down (goal-oriented) modulation of stimulus-driven (e.g. saliency) attentional capture by targets versus distractors. In this regard, a four-module attentional architecture has been defined by Lanillos et al. [45] in which the first module corresponds to the perception sense by building an egocentric map according to relevance encoded as saliency. This information is fed to the top-down controller which ensures that the selection of the new focus of attention will take into account the current system goals and context. Then, the action module chooses the next fixation location and translates it into the proper control signals for the actuators. Finally, the behavioural reorienting module is responsible for detecting novel and behaviourally-relevant stimuli that should result in interrupting and resetting the attentional process as an action-perception loop.

Focusing on the task at hand, the developed visual system should be provided with a *perception* module which builds a saliency map based on the most distinctive visual features, followed by a module in charge of object recognition. In this way, the system will be centred in the potential targets by reducing the sensory data to be processed and, therefore, making tractable the unmanageable amount of information received from the visual sensors. In addition, a memory that stores information about the objects to be recognised should also be integrated. Therefore, our vision system consists of three different modules (Figure 1):

- *Feature Extraction*, that generates a saliency map from image segmentation based on three object properties: colour, shape and motion
- *Memory*, which stores the models of the potential target objects
- *Recognition*, that is responsible for recognising the objects from the visual input and the data coming from the previous modules

Thus, this architecture is based on a richer object description for robustly detecting and recognising any object in real scenarios without establishing any constraint about the objects and the environment.

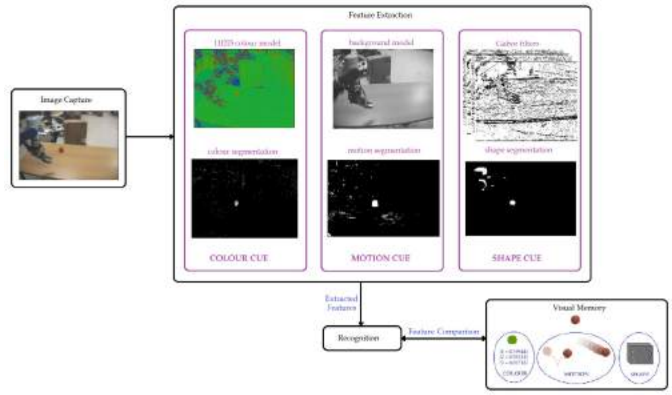

Fig. 1. Overview of the system architecture for object detection and recognition, showing its three main modules (Feature Extraction, Memory and Recognition) and the threefold object description (colour, motion and shape)

A. Feature Extraction

Visual features are a key point in any detection and recognition procedure. Deciding what features are required to properly detect and recognise a *target* object in detriment of others is not an easy task. The reason lies in the fact that a wide variety of features would result in a very time-consuming processing, while a poor feature-based object description would lead to an inefficient recognition. So, similar to human attentional mechanisms (see for example [46] for an extensive survey), a discrimination between features of incoming stimuli has to be defined to properly establish *behaviour*- and *task*-relevance. In particular, in this work three distinctive feature types are considered: colour, motion and shape. Therefore, in an early step an image is divided into semantically meaningful parts according to the values of those properties, which will be part of the robot's focus of attention for further processing.

1) Colour cues: Colour plays a main role in object detection and recognition due to the rich information it can provide. A wide range of approaches can be found in the literature. For instance, colour histograms can be used to represent and match images or objects. However, despite its simplicity and efficacy, its accuracy is significantly deteriorated when the illumination conditions change.

As an alternative, the colour gradient obtained from the addition of channel derivatives could be considered. Nevertheless, given that the colour derivatives are separately computed, differences in the colour edge directions can make this technique miserably fail.

Another possibility could be to use a different colour model. Actually, a great variety of colour spaces are normally used for different purposes such as video and television (YIQ, YUV); display and printing (RGB, CMY); perceptual uniform spaces (U*V*W*, L*a*b*, Luv); human perception (HSI); or standard primary colours (*rgb*, *xyz*). However, a large number of these colour models are combinations of RGB (e.g. CMY, XYZ and $I_1 I_2 I_3$) or normalizations of *rgb* in terms of intensity (e.g. IQ, *xyz*, UV, U*V*, a*b*, uv); others, on the contrary, are correlated to intensity I (e.g. Y, L* and W*).

Thus, keeping in mind the goal of a visual system able to accurately detect and recognise multi-coloured objects in real scenes, existing colour models have been analysed in order to determine which one is more robust to changes in illumination, object geometry and camera viewpoint. The aim is a colour model that is less sensitive to imaging conditions and has a higher discriminative ability, removing the constraints on the image process and, as a consequence, considerably improving object detection and recognition.

In this sense, Gevers and Smeulders [47] and later Villamizar et al. [42], deeply analysed diverse colour models by evaluating their robustness for object recognition under different image parameters. This comparison, summarized in Table I, concluded that the colour model to be chosen depends on the imaging conditions. Indeed, if all the imaging conditions are controlled, RGB is the most invariant colour model for object recognition. However, under the constraints of white illumination and no presence of highlights, normalized colour rgb and $c_1c_2c_3$ are the most robust colour spaces. On the contrary, in the presence of highlights, o_1o_2 is the most appropriate despite its sensitivity to all the other parameters. Finally, $l_1l_2l_3$ is the best alternative for the job at hand due to its invariance.

	shadow	geometry	material	highlights
RGB	+	+	+	+
rgb	-	-	+	+
$c_1c_2c_3$	-	-	+	+
o_1o_2	+	+	+	-
$l_1l_2l_3$	-	-	-	-

TABLE I
COLOUR MODEL SENSITIVITY TO IMAGE PARAMETERS SUCH THAT **+** DENOTES SENSITIVITY, WHILE **-** INDICATES INVARIANCE TO A PARTICULAR PARAMETER.

Given that no environmental and object constraints are established, the $l_1l_2l_3$ colour space is used in our system for object recognition due of its robustness in the presence of varying illumination across the scene (e.g. multiple light sources with different spectral power distributions), and also with changes in surface orientation of the object (i.e. its geometry), and with object occlusion and cluttering. Thus, the first step is to obtain $l_1l_2l_3-$images from the captured $RGB-$images as follows:

$$\begin{cases} l_1 = \frac{(R-G)^2}{(R-G)^2+(R-B)^2+(G-B)^2} \\ l_2 = \frac{(R-B)^2}{(R-G)^2+(R-B)^2+(G-B)^2} \\ l_3 = \frac{(G-B)^2}{(R-G)^2+(R-B)^2+(G-B)^2} \end{cases} \quad (1)$$

Nevertheless, with the aim of robustly detecting and recognising objects in realistic scenarios, other cues must also be used.

2) Motion cues: The capability of visually perceiving motion is a key issue in computer vision. Actually, this is a requirement for a wide range of applications. By way of example, Orabona et al. [48] used motion as a salient feature to focus attention on moving elements. Another alternative is to use independent motion in weakly supervised object recognition settings thanks to the priors provided on the visual target location [49]. In addition, other object characteristics that are significant for detection and recognition can be generated from motion data (e.g. trajectory, speed or shape).

Nonetheless, the motion present in a visual input could be caused by various circumstances such as the camera's movement, a flickering scene illumination, the movement of scene elements (targets or vacillating background elements), or a combination of them. As a consequence, these factors must be considered when image segmentation for motion detection is performed.

Research on this topic has taken a number of forms. The early algorithms [50] were based on temporal information by using a thresholded frame difference of temporally adjacent frames. This kind of methods have some well-known problems, such as ghosts and foreground aperture [51]. As a consequence, they were mostly replaced by methods based on spatial information in the image sequence, namely background subtraction. This technique, in its simplest form, detects moving regions in an image by taking the pixel-by-pixel difference between the current image and a reference background image. This approach is sensitive to changes in the scene background due to the lack of a reliable reference image or the effect of changing illumination, noise or periodic motion, and requires the use of a good background model [52] [53] together with a well-defined stationarity criterion to decide when a pixel deviates from the background [54]. Afterwards, most of the research focused on methods for background maintenance, that is, the construction and updating of a statistical representation of the background trying to capture the temporal evolution of the image sequence. As a representative selection of methods we can mention Pfinder [52] in which a single Gaussian distribution was used, multimodal statistical models such as a mixture of Gaussians (MoG) [55] [56] or Normal distributions [57]; adaptive background estimation based on Wiener (Wallflower [54]) or Kalman filtering [29] [58] to make predictions of the expected background; statistical models based on the minimum and maximum intensity values and the maximum inter-frame change (temporal derivative [59]). Other methods incorporate spatial region-based scene information such as Kernel Density Estimation (KDE), a Parzen-window estimate with a kernel [60]), Eigenbackground (eigenspace decomposition based on images of motionless backgrounds [61]) or Independent Component Analysis (ICA [62]). A number of alternative approaches used Hidden Markov Models [63], codebook vectors [64] [63] or explicit models of the foreground [65].

More recent approaches tend to incorporate specific knowledge of the particular application [29] [66] [67]; introduce a number of enhancements and refinements in the fundamental methods above [68]; or apply other techniques such as saliency maps [29] or regions of interest [69] prior to background subtraction.

Despite the wide research on this topic, there are still some issues to be solved such as how to arrange for a training period with foreground objects in dynamic, real environments; the adaptation to minor dynamic, uncontrolled changes such as the passage of time, blinking of screen or shadows; the adaptation to sudden, unexpected changes in illumination; or the differentiation between foreground and background objects in terms of motion and motionless situations.

With the purpose of overcoming these problems, Martinez-Martin and del Pobil proposed a hybrid algorithm based on frame differencing and background subtraction along with a single-Gaussian background model and a mechanism for its effective maintenance (which is described in depth in [70]). The underlying idea of this method is to mutually reinforce frame difference and background subtraction so that the drawbacks of both approaches are overcome while keeping their original advantages.

So, in a first stage, an initial background model is built. Unlike most background estimation algorithms, another technique for controlling the activity within the system workspace is performed. As computational and time cost are critical issues, this control is performed by means of a combination of difference techniques: frame difference with reference frame subtraction. Thus, frame difference allows the system to identify objects which have moved from one frame to the next one. However, it is important to take into account that both previous position and the current one are detected. This problem was solved by using background subtraction since the only highlighted position is the current one. Note that, as the reference frame is the first taken frame, it might be possible that it contains objects that are not part of the background. For that reason, some additional constraints have been defined in order to solve this kind of situations. Furthermore, the used thresholds for those subtraction approaches are automatically set for each pixel from pixel neighbourhood information. In a similar way, the stationary object problem has been solved with the combination of both subtraction techniques. Therefore, there is no danger of missing foreground objects while the initial model is being built. Moreover, the obtained background model does not contain information about those moving targets thanks to the use of a simple frame-difference approach that detects moving objects within the robot workspace.

In a second stage, adjacent frame difference, background subtraction and background maintenance techniques are used. So, the detection and identification of moving objects is composed of two processes:

1) the adaptive background model, built initially, is used to classify pixels as foreground or background. This is possible because each pixel belonging to the moving object has an intensity value that does not fit into the background model. That is, the used background model associates a Gaussian distribution to each pixel of the image, as defined by its mean colour value and its variance. Then, when an interest object enters or moves around the system workspace, there will be a difference between the background model values and the object's pixel values. A criterion based on stored statistical information is defined to deal with this classification and it can be expressed as follows:

$$b(r,c) = \begin{cases} 1 & \text{if } |i(r,c) - \mu_{r,c}| > k \times \sigma_{r,c} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

where $b(r,c)$ is the binary value of the pixel at row r and column c to be calculated, $i(r,c)$ represents the pixel brightness in the current frame, $\mu_{r,c}$ and $\sigma_{r,c}$ are the mean and standard deviation values estimated by the background model and k is a constant value which depends on the point distribution

2) improvement of the raw classification based on the background model as well as detection and adaptation of the background model when a global change in illumination occurs. The proper combination of subtraction techniques is used to improve the segmentation carried out at pixel level by using background subtraction. Furthermore, this difference processing allows the system to identify global illumination changes. It is assumed that a significant illumination change has taken place when there is a change in more pixels than two thirds of the image size. When an event of this type occurs, a new adaptive background model is built because, otherwise, the application would detect background pixels as targets, since the model is based on intensity values and a change in illumination produces a variation of them.

Once the whole image is processed, those pixels classified as background are incorporated into the adaptive background model. For that, the following formulas are used:

$$\begin{cases} \mu_{r,c}(t+1) = \begin{cases} (1-\alpha)\,\mu_{r,c}(t) + \alpha i_{t+1}(r,c) & \text{if background} \\ \mu_{r,c}(t) & \text{otherwise} \end{cases} \\ \sigma_{r,c}(t+1) = \begin{cases} (1-\alpha)\,\sigma_{r,c}(t) + \alpha i_{t+1}(r,c) & \text{if background} \\ \sigma_{r,c}(t) & \text{otherwise} \end{cases} \end{cases} \quad (3)$$

Here, the constant α $(0 < \alpha < 1)$ controls the adaptation rate and it is given by the number of pixels which are part of the Gaussian distribution. However, sometimes the pixel grey level might change quicker than the background model as when illumination gradually brightens. As the proposed updating process is too slow, after a certain period of time, the background model might not be suitable for foreground pixel detection. For that reason, a new updating process was designed. So, during the updating phase two different tasks are carried out:

- the background model is being updated with each new frame by using Eq. 3
- a new background model is being built from the segmentation obtained with the current background model

In this way, after some time, the background model is replaced by a new one more suitable for the current background scene.

3) Shape cues: Shape is the third characteristic describing an object in our system. Similar to the motion cue, enriched information can be obtained from shape data. However, object shape may change when the object is observed from a different point of view. For instance, a car presents different shapes depending on the location of the observer (front, bottom, sideways or in perspective). To overcome this problem, different object shapes should be represented in accordance with the distinct observable views. Obviously, the robustness obtained from a greater number of shapes will come hand in hand with a higher computational cost.

As a solution, Principal Components Analysis (PCA) has been widely used (e.g. [71] [72] [73]) as a statistical tool for finding patterns in data of high dimension, highlighting their similarities and differences. In our case, object templates are matched with their *appearance* in the current image. First the provided training data is pre-processed in some way (e.g. image normalization for contrast, optical flow computation, face alignment, etc.) and then, the dimensionality of the search space is reduced by converting a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables (i.e. *principal components*). As a consequence, invariance with respect to object contrast, rotation or scale is not provided by PCA itself. In a similar way, other problems such as occlusions, illumination variations, high object dimensionality or image noise, are not solved with this approach.

A neuroscientific viewpoint reveals that Gabor filters is the approach with a higher biological plausibility [74] [75] [76] [77]. In this way, images are represented by a sinusoidal function moved in depth and the wavelength of any sinusoidal shape pattern can be detected and recognised. What is more, phase-based methods have been shown to be robust to changes in contrast, scale and orientation [78] [79].

Therefore, a symmetrical and an anti-symmetrical filter kernels can be used to estimate the phase difference at any point x. As a result, the two obtained filter outputs for an image I would be:

$$\begin{cases} I_{sin,\sigma}(x,\omega) = \int \omega \left(\frac{x-x'}{\sigma} I(x')\, sin\left(\omega\left(x-x'\right)\right)\right) dx' \\ I_{cos,\sigma}(x,\omega) = \int \omega \left(\frac{x-x'}{\sigma} I(x')\, cos\left(\omega\left(x-x'\right)\right)\right) dx' \end{cases} \tag{4}$$

where σ corresponds to the spatial expansion of the kernel filter and ω refers to its frequency. Note that when the ratio between ω and σ is a constant and a Gaussian bell curve represents the window function, then Equation (4) describes a convolution with Gabor functions.

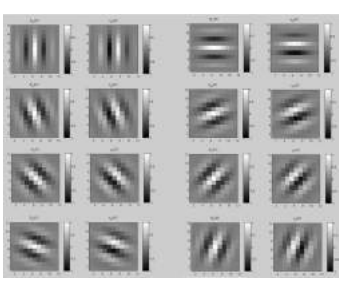

Fig. 2. Bank of oriented Gabor filters used for shape detection and recognition

In particular, the proposed method extracts the object shape using a bank of eight oriented Gabor filters (Figure 2). For that, we have constrained the number of shape representations to four at most: (1) a shape when the object is seen from the front; (2) a shape when the object is observed sideways; (3) a shape when the object is seen from the top; and, (4) one shape representation when it is seen in perspective (chosen thinking of autonomous systems performing a task). Note that the system only requires a certain number of shape representations to recognise an object. For instance, objects like balls only require one shape representation, while other objects will need two or three shapes. An example of some shape models for different objects are shown in Figure 3.

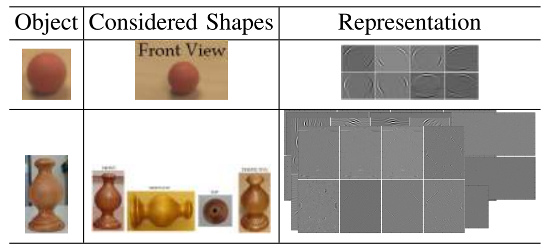

Fig. 3. Shape cue in terms of Gabor filters based on four shape representations (front, sideways, top and perspective) used in the proposed approach

B. Memory

Memory performs a fundamental role in human object recognition. Similarly, in our system, a *memory* module stores the description of all the potential targets to be recognised. It contains all the features integrating the description of each *known* object, as shown in Figure 4.

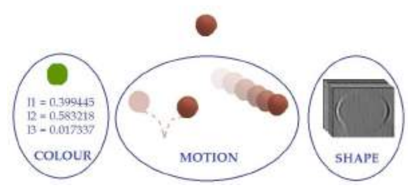

Fig. 4. Object description in terms of colour, motion and shape properties saved in the system memory for proper object detection and recognition

C. Recognition

The last stage of the process is performed by the recognition module, which is responsible for the object recognition itself. At this point, it is important to take into account that two different kinds of object recognition can be distinguished; namely, object categorization and object identification. On the one hand, the goal of object categorization is to classify an object as belonging to an abstract object class (e.g. animal, person, car, building, etc.). On the other hand, object identification is aimed at identifying an object as a unique instance within a class. In this paper object identification is addressed, since no category abstraction is intended.

Our approach is aimed at visually identifying the surrounding objects in their corresponding object classes. For that, a statistical combination of similarity likelihood is used, based on all the considered cues. Assuming independence between the three cues (colour, motion and shape), the object-based likelihood can be obtained as follows:

$$P(I|o) = P(I_c|o)|P(I_m|o)|P(I_s|o) \tag{5}$$

where $P(I_c|o)$, $P(I_m|o)$ and $P(I_s|o)$ respectively correspond to colour-based, motion-based and shape-based likelihoods for an object o.

Note that the task to be performed and the object characteristics will determine what features are more distinctive for achieving an accurate object recognition. For that reason, the cue weights have to be experimentally set. By way of example, for recognizing a ball, a greater weight is assigned to colour as compared to shape, since a circular shape is very common in real-world scenarios and its discriminative value is lower.

III. Implementation Details

Real-time processing is a critical demand when state-of-the-art robot systems are designed. This requirement calls for an efficient processing unit. A solution is to process visual input with a Graphical Processing Unit (GPU), potentially reducing time consumption in a drastic way. However, despite its highly parallel computation capabilities, writing efficient GPU programs is not evident, especially for uneven workloads (e.g. the higher the number of interest objects is, the higher the computational costs are).

In particular, our algorithms have been implemented on an NVIDIA GeForce GTX 745. It includes 384 Compute Unified Device Architecture (CUDA) cores with 4-GB memory and chip-level power enhancements. A fast access to shared and GPU's main memories characterizes these CUDA cores. Moreover, graphics API functions are not required for parallel implementations in C language; this is very convenient for properly implementing the necessary parallel algorithms that deal with irregular workloads.

The CPU-GPU system implementation is shown in Figure 5. The CPU captures an image and uploads it to the GPU, which will perform the subsequent image processing steps, namely, from feature extraction to object recognition. The GPU will return the output to the CPU for it to decide the next action to be performed by the robot. Then, the visual processing starts again.

Since object feature detection and tracking is a computationally intensive task, but highly parallelizable, a good parallel solution can be devised to the effect that all image processing is carried out by the GPU (using 1023 threads per block). As a final system output, the CPU shows on the screen the detected objects.

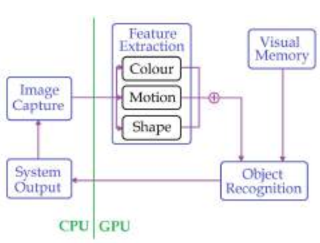

Fig. 5. Overview of our CPU-GPU implementation, meeting real-time performance by parallelly implementing on the GPU the computationally intensive task of object detection and recognition

IV. Experimental Results

The proposed approach for object detection and recognition in real scenarios has been tested in three different kinds of scenarios. First of all, a semi-structured scene was considered so that a methodical study of the efficiency based on different factors could be carried out (e.g. occlusions, light reflexes, changes in illumination, shadows, etc.). Then, the second set of experiments involved two real, cluttered environments in which the target objects were to be found amongst a set of ordinary items such as calendars, books, clocks or pens. Finally, an image dataset has been used to evaluate the performance of the system by means of object instance recognition and in comparison with other state-of-the-art approaches. To conclude, a performance analysis in terms of execution time is presented.

For the two first experiments, a humanoid torso endowed with a Robosoft TO40 pan-tilt-vergence stereo head and two multi-joint arms was used (see Figure 6). The head mounts two Imaging Source DFK 31BF03-Z2 cameras acquiring colour images at 30 Hz with a resolution of $1024x768$ pixels. The baseline between cameras is 270 mm and the motor positions are provided by high-resolution optical encoders.

Fig. 6. External view of the humanoid torso employed for the experiments (left) and a detailed view of the pan/tilt/vergence head (right)

A. Experiment 1: Semi-structured scenes

In the case of semi-structured scenes, the robot was located in front of a table on which the objects were placed. In this experimental setup, the table was initially empty and, after a little while, a human was placing and removing the different objects on the table without interacting directly with the robot system. In this way, the motion cue was instrumental in detecting both the human presence in the robot workspace as well as the new object instance on the table. Actually, in this experiment, the three visual cues have the same weight when the segmentation result is determined. Four different objects have been used as targets: a red ball, a toy car, a bottle and a money box. The object position and orientation were modified for each frame. Obviously, the number of resulting orientations varies based on the considered object; for instance, the red ball has only one orientation, while the toy car was observed in 12 different orientations (approximately every 30 degrees). As depicted in Figure 7, the implemented approach starts with capturing an image. This image is the input of two different processes: the colour cue segmentation and the segmentation of the other two considered cues (i.e. motion and shape). This distinction is for efficiency reasons. Therefore, on the one hand, the image is expressed in $L1L2L3-$coordinates and segmented by using the memory information about the different objects to be found. On the other hand, an intensity image is obtained with the purpose of speeding up motion and shape segmentation. Note that shape detection is obtained from the combination of the 8 Gabor-filtered images. Once segmentation for each cue is performed, their fusion allows the system to reduce the search area for object recognition and, despite the presence of factors such as shadows or reflexions, the red ball is properly detected in the image.

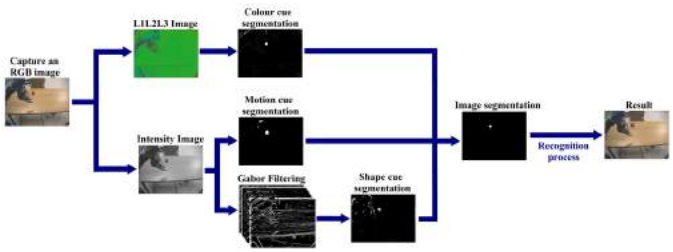

Fig. 7. Object detection and recognition process in semi-structured scenes

In a similar way, experiments with the other objects were carried out. Figure 8 shows some of the obtained results (only the final result). Note that the illustrated results correspond to a single trial since there is no randomness in the data. As it can be observed, only one object is searched each time. The reason lies in the performance analysis in the presence of different factors susceptible of making the system fail (e.g. shadows, flickering light sources, variable light reflexes, objects partially visible, etc.). As shown, all the objects were successfully recognised even when they changed their orientation or location in the scene, or the cameras changed their viewpoint.

Fig. 8. Qualitative experimental results in a semi-structured scenario in which both the camera viewpoint and the object location and orientation were continuously changed

With the purpose of validating the obtained qualitative results, a quantitative evaluation has been carried out. In this case, the *true positive rate (TPR)* and *false positive rate (FPR)* measurements are used [80]. That is, the proportion of correctly classified positives (TPR); and the proportion of incorrectly classified negatives (FPR). From their definition, a good performance is obtained when both measurements are close to 1. As shown in Figure 9, the obtained results (blue line) are above the line dividing the ROC space (grey line), which means a good performance. Consequently, it can be concluded that the system was successful in the object detection/recognition task.

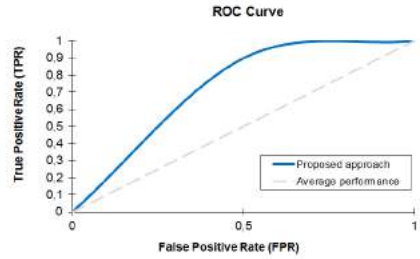

Fig. 9. ROC curve for the quantitative evaluation of the proposed approach in semi-structured scenes

B. *Experiment 2: Real scenarios*

In this experiment, the objects to be detected and recognised were placed on a desk. Two unstructured environments were used composed of everyday objects of different nature and features such as textured books, pens, clock, etc. In this context, the objects to be detected and recognised include a red ball, a toy car, a yellow ball, a green bulb, a stapler, and a wooden generalized cylinder. These objects were located at different positions and/or orientations within the considered scenario, resulting partially occluded in some cases. As in the previous case, a human is continuously interacting with the target objects, but not with the robot system, so that the motion cue triggers again a visual attention focus. However, the other two visual cues are required to distinguish between the target objects and other moving elements in the scene such as the person. For this reason, the three cues have the same weight in the object recognition process.

In the first experiment, three different objects were used: a toy car, a stapler, and a wooden generalized cylinder. Some of the one-trial results are shown in Figure 10. Note that, despite the nature of the environment and that of the objects themselves, all the targets were properly detected even in the case of the toy car, which had a great colour similarity with its background. An example of the detection of two objects in the same image is also illustrated, in which the car and the stapler have been correctly detected and recognised. In a similar way, the developed approach adequately focuses its attention on the target object (i.e. the generalized cylinder), although several objects were added to the scene (the toy car and the stapler) as shown in the rightmost example in Figure 10.

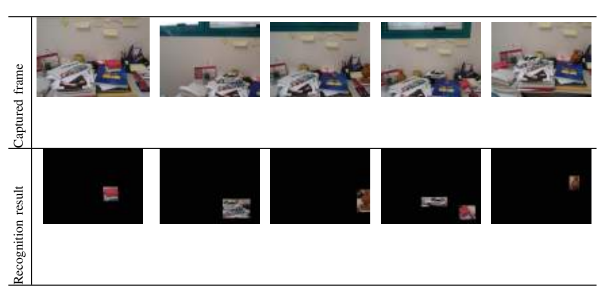

Fig. 10. Qualitative experimental results when a real scenario is considered

In the subsequent experiment, the visual system was aimed at detecting and recognising four objects (a red ball, a green bulb, a yellow ball, and a wooden generalized cylinder) while a person is interacting with the objects in the scene, changing their position on the desk. As a consequence, the motion cue again plays a main role in the object recognition process. Some of the obtained one-trial results are presented in Figure 11. In this case, unlike in previous examples, the binary image is shown, highlighting the detected objects, especially when they are partially occluded, or colour similarity with the background is considerably high.

Once more, a quantitative analysis validates the above qualitative results. In this case, as shown in Figure 12, the detection and recognition results are presented for two different objects: the stapler and the toy car. In both cases, the ROC curve are above the division line (grey dashed line), confirming the quality of the results for real-life scenarios.

C. Experiment 3: Image Repository

For the third validation experiment we compare the performance of our approach with state-of-the-art methods by using a public image repository. Actually, given that the ability to recognise objects is crucial for many applications, a wide range of public image repositories is available. These datasets allow researchers to evaluate their approaches with a large number of objects and under different conditions, as well as to compare their performance with other state-of-the-art approaches. However, these repositories could be classified based on the goal to be satisfied. That is, object recognition has multiple levels of semantics (e.g. category recognition, instance recognition, pose recognition, etc.), it can refer to different application scenarios or it could be based on certain input data. Consequently, the required evaluation dataset must correspond to the needs of a particular approach. This is why the RGB-D Object Dataset [81], publicly available at http://www.cs.washington.edu/rgbd-dataset, has been used for this validation. This dataset is composed of thousands of images of 300 objects commonly found in home and office environments, taken from multiple views by using an RGB-D camera (see Figure 13 for some examples). Objects are organized into a hierarchy of 51 categories composed of a number of instances between three and fourteen, so that each object belongs to only one category. In addition, ground truth images are provided to adequately assess the segmentation process. In consequence, this image dataset allows object recognition techniques to be evaluated at two levels:

- *Category level*. Category recognition refers to classifying previously unseen objects in a category based on objects from the same category that have been previously seen. That is, this recognition level corresponds to answering questions such as *is this an apple or a cup?*
- *Instance level*. Instance recognition, on its behalf, involves identifying if an object is physically the same object that has been previously seen. In this case, the questions to be answered take the form *is this Angel's coffee mug or Ester's?*

Despite the fact that the ability to recognise objects at both levels is a key point in the context of robotic tasks, in this work only the instance recognition is considered since no *category* abstraction was carried out. So, the task for the recognition algorithm is to detect the exact physical instance of an object that was previously presented. In our case, the previous instance (i.e. the first frame of each object sequence) based on colour and shape cues is used to build an object model that will be used for object detection and recognition. Note that, in this case, the motion cue has not been used because objects are not moving, although the camera is.

For comparison reasons, we consider the cropped RGB-D frames that tightly include the object, exactly as used in the object recognition evaluation of the paper introducing the RGB-D Object Dataset [81] (i.e. subsampled every 5th video frame). Actually, these are the images used for obtaining the different results over this image repository.

Table II compares the obtained results with those from different state-of-the-art approaches; namely, EB Local (an exemplar-based local distance function learning technique [82]), Linear Support Vector Machine (Linear SVM), Gaussian kernel Support Vector Machine, Random Forest (RF), kernel descriptors, Convolutional K-Means descriptors (CKM Desc), HMP and IDL) described in [83] [84] [85] [86]. As it can be observed from the results, our technique substantially improves upon the performance of the several considered state-of-the-art classification approaches.

Approach	Accuracy based on RGB information
EB Local	84.5
Linear SVM	90.2
Nonlinear SVM	90.6
RF	90.5
IDL	91.3
CKM Desc	92.1
The proposed approach	**96.1**

TABLE II
ACCURACY COMPARISON ON THE RGB-D OBJECT DATASET WHEN USING ALTERNATING CONTIGUOUS FRAMES

In addition, the RGB-D image dataset also includes video sequences of real-life scenarios such as office workspaces, meeting rooms, and kitchen areas, where some database objects are visible from different viewpoints and distances and may be partially or completely occluded in some frames. Thus, the proposed algorithm has been also tested in those common indoor environments. Some of the obtained results are illustrated in Figure 14. The first two images show an office and, although the scene illumination and the point of view have been changed, they correspond to the same video sequence. As it can be observed, the cellophane box has been recognised in both of them, highlighting the approach robustness to lighting changes. Furthermore, the second row refers to different scenarios with the same target object: a green bowl. As it is apparent, it was properly detected, even when it was partially occluded.

D. Experiment 4: Execution Time Analysis

The last evaluation experiment refers to the analysis of the benefits of using the GPU for parallel computing. A similar study was presented by Ferreira et al. [87] in the context of Bayesian models for multimodal perception. With that aim, we carried out a comparison between the performance using parallel and non-parallel computing depending on the image resolution and the number of potential targets.

Fig. 11. Qualitative experimental results when a real scenario is considered

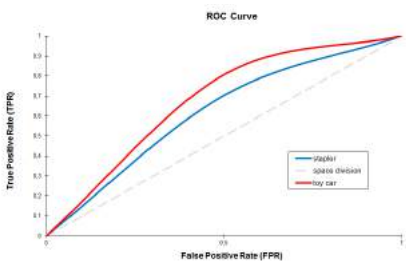

Fig. 12. ROC curve for the quantitative evaluation of the proposed approach in unstructured scenarios

Fig. 13. Some objects from the RGB-D Object Dataset belonging to different object categories

Fig. 14. Some object recognition results on the real-life scenarios provided by the RGB-D image dataset

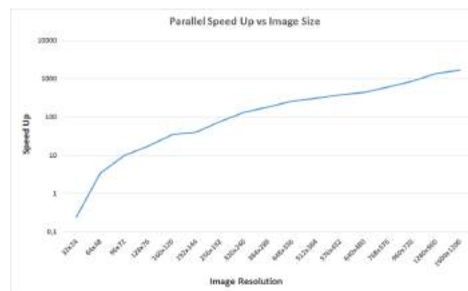

Fig. 15. Speedup versus image size for parallel (GPU) and non-parallel (CPU) computing

First, the execution time is analysed for different image resolutions. Our results show that a similar performance is obtained with the two methods when the image resolution is low. However, when the image resolution is increased, the non-parallel computing time drastically climbs, while the GPU implementation shows a gradual, much slower, growth. This is apparent in Figure 15 that plots the speedup with respect to image size. In fact, the execution time for the GPU remains virtually constant (around 0.48 seconds) for the first ten image resolutions considered because the thread loads remain similar. Given that the number of threads is limited, when the image resolution is increased, both the thread work load and, consequently, the execution time rise, resulting in 0.95 seconds for our higher resolution (1600x1200).

Another key issue in practical object recognition is that of scalability, and our last experiment analyses the execution time when the number of potential target objects is increased. With that aim, different image sequences from the RGB-D image dataset were used. The results, shown in Figure 16, illustrate the speedup evolution for an averaged image resolution of 84x85 pixels when the number of objects that could be found in the scene increases. As it can be observed, our results highlight the efficiency when parallel computing is used; computation times remain almost unchanged between one object (0.46 seconds) and 50 target objects (0.47 seconds). Keeping in mind our final goal, an autonomous assistive robot, the system should provide a similar response time regardless of the task at hand, as it is the case, and, ideally, this response time should be the same as that of human beings. As our results show, the obtained response time is similar in all the studied cases (up to 50 target objects) and below 0.5 seconds, approximately twice the average human reaction time (between 200-250 milliseconds [88] [89] [90]). In the context of human-computer interaction [91] [92] [93], a response time below 0.1 second is regarded as an instantaneous reaction, whereas a response delay between 0.1 and 1.0 second is considered as fast enough for a fluent interaction, even though the user would notice the delay. Consequently, a response time of 0.5 seconds is a real-time performance in this sense. In fact, with this implementation, real-time processing could be obtained even when hundreds of object instances are searched, taking us closer to the possibly thousands of objects that could be found in everyday life.

On the other hand, advances in image technology are leading to visual sensors with higher image quality to the effect that higher and higher image resolutions can be expected in the future. For resolutions higher that 1600x1200, execution times would be presumably beyond 1.0 second. In this case, image resolution could be decreased by using, for instance, pyramidal images, in order to obtain real-time performance.

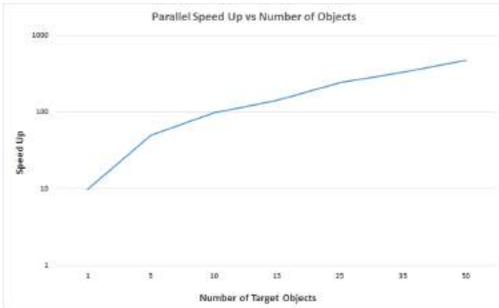

Fig. 16. Speedup versus the number of potential target objects for parallel (GPU) and non-parallel (CPU) computing

V. Conclusion and Future Work

During the last decades, robotics research moved from stationary robotic systems in constrained environments to mobile and service-oriented robots operating in realistic and unconstrained environments. One rising application field is assistive robotics, aimed at developing robots that support humans as their daily-life assistants. With that aim, these systems must be endowed with different abilities such as localization, mapping, path planning, obstacle avoidance, object detection, recognition, and manipulation.

In particular, in this paper we have focused on object detection and recognition. Even though this issue is the heart of different robotic assistive abilities, real-time efficient object detection and recognition is still a challenging problem when real scenarios are considered. Part of this problem is due to the presence of cluttered, dynamic backgrounds, with possible occlusions, interactions and additional photometric and geometric variations.

Motivated by these challenges, we presented a framework that is able to detect and recognise objects from a visual input in unconstrained scenes in real time. We take inspiration from biology and use a rich object description based on colour, motion and shape cues. Robust colour information is obtained thanks to an adequate colour model choice that makes visual data invariant to changes in viewpoint, object geometry and illumination. The second considered cue is motion, which is perceived by means of a novel background maintenance technique overcoming the environmental constraints of existing methods. Finally, a phase-based representation of shape concludes the object description presented in this paper.

Once the visual features have been properly extracted, the system analyses the statistical similarity between the detected objects and those whose description is stored in the system's visual memory. This estimated joint likelihood allows the system to successfully discriminate between several objects. Furthermore, with the purpose of effectively achieving real-time computation in visual data processing, a Graphical Processing Unit (GPU) is used by taking into account that irregular workloads are common in the task at hand.

The proposed approach has been implemented on a robotic platform and tested by considering different parameters which might make the system fail. This large number of parameters allows us to analyse the robustness of the proposed method. For further experimental validation, a public image repository for object recognition has been used, allowing a quantitative comparison with respect to other state-of-the-art techniques when real-world scenes are considered. Finally, a temporal analysis of the performance was provided with respect to image resolution and number of target objects in the scene. As shown by these experimental results, the system is able to accurately detect and recognise objects in everyday scenarios where there are no constraints about the environment and the objects.

As future work, new object features will be studied for improving object detection and recognition. In addition, a module for visual attention will be developed and integrated in the current implementation with the purpose of determining which features make an object more interesting for the system. At the same time, we would like to add a new stage in order to automatically learn new objects, going a step further in emulating the human visual system.

Acknowledgment

This work has been partially funded by Ministerio de Economía y Competitividad (DPI2015-69041-R), by Generalitat Valenciana (PROMETEOII/2014/028), and by Jaume-I University (P1-1B2014-52).

References

[1] E. Martinez and A. del Pobil, "Visual surveillance for human-robot interaction," in *SMC*, 2012, pp. 3333–3338.
[2] E. Martinez and A. del Pobil, "Safety for human-robot interaction in dynamic environments," in *ISAM*, 2009, pp. 327–332.
[3] T. Breuer, G. G. Macedo, R. Hartanto, N. Hochgeschwender, D. Holz, F. Hegger, Z. Jin, C. Müller, J. Paulus, M. Reckhaus, J. A. Ruiz, P. Plöger, and G. Kraetzschmar, "Johnny: An autonomous service robot for domestic environments," *J Intelligent and Robotic Systems*, vol. 66, pp. 245–272, 2012.
[4] (2013) Hobbit-the mutual care robot. [Online]. Available: http://hobbit.acin.tuwien.ac.at/
[5] (2013) Ksera-knowledgeable service robots for aging. [Online]. Available: http://ksera.ieis.tue.nl/
[6] (2007) Cogniron-the cognitive robot companion. [Online]. Available: http://www.cogniron.org/final/Home.php
[7] (2015) Care-o-bot. [Online]. Available: http://www.care-o-bot-4.de/
[8] (2015) Herb. [Online]. Available: http://www.cmu.edu/herb-robot/
[9] (2015) Accompany. [Online]. Available: http://www.accompanyproject.eu/
[10] A. Costa, P. Novais, and R. Simoes, "A caregiver support platform within the scope of an ambient assisted living ecosystem," *Sensors*, vol. 14, pp. 5654–5676, 2014.
[11] I. Leite, C. Martinho, and A. Paiva, "Social robots for long-term interaction: A survey," *Intl Journal of Social Robotics*, vol. 5, pp. 291–308, 2013.
[12] F. Bonin-Font, A. Ortiz, and G. Oliver, "Visual navigation for mobile robots: A survey," *J Intelligent and Robotic Systems*, vol. 53, pp. 263–296, 2008.
[13] J. Antich, A. Ortiz, and G. Oliver, "A control strategy for fast obstacle avoidance in troublesome scenarios: application in underwater cable tracking," in *IFAC Conf on Manoeuvring and Control of Marine Craft*, 2006.
[14] H. Morita, M. Hild, J. Miura, and Y. Shirai, "Panoramic view-based navigation in outdoor environments based on support vector learning," in *IROS*, 2006, pp. 2302–2307.

[15] J. Shen and H. Hu, "Visual navigation of a museum guide robot," in *WCICA*, vol. 2, 2006, pp. 9169–9173.

[16] D.-H. Lee and J.-H. Kim, "A framework for an interactive robot-based tutoring system and its application to ball-passing training," in *ROBIO*, 2010, pp. 573–578.

[17] P. Turaga, R. Chellappa, V. Subrahmanian, and O. Udrea, "Machine recognition of human activities: A survey," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 18, pp. 1473–1488, 2008.

[18] O. Chang, "Evolving cooperative neural agents for controlling vision guided mobile robots," in *Intl Conf on Cybernetic Intelligent Systems*, 2010, pp. 1–6.

[19] M. Asada, E. Uchibe, and K. Hosoda, "Cooperative behavior acquisition for mobile robots in dynamically changing real worlds via vision-based reinforcement learning and development," *Artificial Intelligence*, vol. 110, pp. 275–292, 1999.

[20] Y. Kuniyoshi, J. Rickki, M. Ishii, S. Rougeaux, N. Kita, S. Sakane, and M. Kakikura, "Vision-based behaviors for multi-robot cooperation," in *IROS*, vol. 2, 1994, pp. 925–932.

[21] D. Kragic and H. Christensen, "Survey on visual servoing for manipulation," Computational Vision and Active Perception Laboratory, Tech. Rep., 2002.

[22] L. Whitcomb, D. Yoerger, H. Singh, and D. Mindell, "Towards precision robotic maneuvering. survey, and manipulation in unstructured undersea environments," in *Intl Symp on Robotics Research*. Springer-Verlag Publications, 1998, pp. 45–54.

[23] *Recognizing Patterns in Signals, Speech, Images and Videos. ICPR 2010 Contests*, ser. LNCS. Springer Berlin Heidelberg, 2010, vol. 6388.

[24] S. Lee, S. Lee, J. Lee, D. Moon, E. Kim, and J. Seo, "Robust recognition and pose estimation of 3d objects based on evidence fusion in a sequence of images," in *ICRA*, 2007, pp. 3773–3779.

[25] N. Sian, T. Sakaguchi, K. Yokoi, Y. Kawai, and K. Maruyama, "Operating humanoid robots in human environments," in *RSS Workshop: Manipulation for Human Environments*, 2006.

[26] R. Platt, R. Burridge, M. Diftler, J. Graf, M. Goza, and E. Huber, "Humanoid mobile manipulation using controller refinement," in *RSS Workshop: Manipulation for Human Environments*, 2006.

[27] C. Urdiales, M. Dominguez, C. de Trazegnies, and F. Sandoval, "A new pyramid-based color image representation for visual localization," *Image and Vision Computing*, vol. 28, no. 1, pp. 78–91, 2010.

[28] C. Zhang, Y. Qiao, E. Fallon, and C. Xu, "An improved camshift algorithm for target tracking in video surveillance," in *9th IT & T Conf*, 2009.

[29] Z. Kim, "Real time object tracking based on dynamic feature grouping with background subtraction," *CVPR*, vol. 0, pp. 1–8, 2008.

[30] M. Villamizar, A. Sanfeliu, and J. Andrade-Cetto, "Computation of rotation local invariant features using the integral image for real time object detection," in *ICPR*, vol. 4, 2006, pp. 81–85.

[31] P. Wilson and J. Fernandez, "Facial feature detection using haar classifiers," *J Computing Sciences in Colleges*, vol. 21, no. 4, 2006.

[32] D. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, no. 2, pp. 91–110, 2004.

[33] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "Surf: Speeded up robust features," *CVIU*, vol. 110, no. 3, pp. 346–359, 2008.

[34] R. Petrick, D. Kraft, K. Mourão, N. Pugeault, N. Krüger, and M. Steedman, "Representation and integration: Combining robot control, high-level planning, and action learning," in *Intl Cognitive Robotics Workshop (CogRob)*, 2008, pp. 32–41.

[35] P. Fitzpatrick and G. Metta, "Grounding vision through experimental manipulation," *Philosophical Trans of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 361, no. 1811, pp. 2165–2185, 2003.

[36] N. Krüger, M. Ackermann, and G. Sommer, "Accumulation of object representations utilising interaction of robot action and perception," *Knowledge-Based Systems*, vol. 15, no. 1-2, pp. 111–118, 2002.

[37] D. Kraft, R. Detry, N. Pugeault, E. Baseski, J. Piater, and N. Krüger, "Learning objects and grasp affordances through autonomous exploration," in *Computer Vision Systems*, ser. LNCS, M. Fritz, B. Schiele, and J. Piater, Eds. Springer Berlin / Heidelberg, 2009, vol. 5815, pp. 235–244.

[38] L. Yang, H. Cheng, J. Hao, Y. Ji, and Y. Kuang, *Advances in Multimedia Information Processing - PCM 2015: 16rh Pacific-Rim Conf on Multimedia, Gwangju, South Korea, Sep 16-18, 2015, Proc., Part II*. Springer Intl Publishing, 2015, ch. A Survey on Media Interaction in Social Robotics.

[39] H. Yan, M. A. Jr., and A. Poo, "A survey on perception methods for human-robot interaction in social robots," *Intl Journal of Social Robotics*, vol. 6, pp. 85–119, 2014.

[40] A. Palomino, R. Marfil, J. Bandera, and A. Bandera, "A new cognitive architecture for bidirectional loop closing," in *Robot 2015: Second Iberian Robotics Conf: Advances in Robotics*, 2015.

[41] S. Wolfson and M. Landy, "Examining edge- and region-based texture analysis mechanisms," *Vision Research*, vol. 38, no. 3, pp. 439–446, 1998.

[42] M. Villamizar, J. Scandaliaris, A. Sanfeliu, and J. Andrade-Cetto, "Combining color-based invariant gradient detector with hog descriptors for robust image detection in scenes under cast shadows," in *ICRA*, 2009, pp. 1997–2002.

[43] E. Martinez-Martin and A. del Pobil, "Visual object recognition for robot tasks in real-life scenarios," in *URAI*, 2013, pp. 644–651.

[44] H. Al-Absi and A. Abdullah, "Biologically inspired object recognition system," in *Intl Symp in Information Technology*, vol. 1, 2010, pp. 1–5.

[45] P. Lanillos, J. Ferreira, and J. Dias, "Designing an artificial attention system for social robots," in *IROS*, 2015.

[46] J. Ferreira and J. Dias, "Attentional mechanisms for socially interactive robots - a survey," *IEEE Trans on Autonomous Mental Development*, vol. 6, pp. 110–125, 2014.

[47] T. Gevers and A. Smeulders, "Color-based object recognition," *Pattern Recognition*, vol. 32, no. 3, pp. 453–469, 1999.

[48] F. Orabona, G. Metta, and G. Sandini, "Attention in cognitive systems. theories and systems from an interdisciplinary viewpoint," L. Paletta and E. Rome, Eds. Berlin, Heidelberg: Springer-Verlag, 2008, ch. A Proto-object Based Visual Attention Model, pp. 198–215.

[49] C. Ciliberto, F. Smeraldi, L. Natale, and G. Metta, "Online multiple instance learning applied to hand detection in a humanoid robot," in *IROS*, 2011, pp. 1526–1532.

[50] R. Jain and H. Nagel, "On the analysis of accumulative difference pictures from image sequences of real world scenes," *TPAMI*, vol. 1, no. 2, pp. 206–214, 1979.

[51] R. Collins, A. Lipton, T. Kanade, H. Fijiyoshi, D. Duggins, Y. Tsin, D. Tolliver, N. Enomoto, O. Hasegawa, P. Burt, and L. Wixson, "A system for video surveillance and monitoring," Carnegie Mellon University, Tech. Rep., 2000.

[52] C. Wren, A. Azarbeyejani, T. Darrell, and A. Pentland, "Pfinder: Real-time tracking of the human body," *TPAMI*, vol. 19, no. 7, pp. 780–785, 1997.

[53] D. Koller, J. Weber, and J. Malik, "Robust multiple car tracking with occlusion reasoning," in *ECCV*, 1994, pp. 189–196.

[54] K. Toyama, J. Krum, B. Brumitt, and B. Meyers, "Wallflower: Principles and practice of background maintenance," in *ICCV*, vol. 1, 1999, pp. 255–261.

[55] Z. Zivkovic, "Improved adaptive gaussian mixture model for background subtraction," in *ICPR*, vol. 2, 2004, pp. 28–34.

[56] C. Stauffer and W. Grimson, "Learning patterns of activity using real-time tracking," *TPAMI*, vol. 27, no. 5, pp. 747–757, August 2000.

[57] N. Friedman and S. Russell, "Image segmentation in video sequences: A probabilistic approach," in *Conf on Uncertainty in Artificial Intelligence*, 1997, pp. 175–181.

[58] C. Ridder, O. Munkelt, and H. Kirchner, "Adaptive background estimation and foreground detection using kalman-filtering," in *Intl Conf on Recent Advances in Mechatronics*, 1995, pp. 193–199.

[59] I. Haritaoglu, D. Harwood, and L. Davis, "W4: Real-time surveillance of people and their activities," *TPAMI*, vol. 22, no. 8, pp. 809–830, 2000.

[60] A. Elgammal, D. Harwood, and L. S. Davis, "Non-parametric model for background subtraction," in *ECCV*, 2000, pp. 751–767.

[61] N. Oliver, B. Rosario, and A. Pentland, "A bayesian computer vision system for modeling human interactions," *TPAMI*, vol. 22, no. 8, pp. 831–843, 2000.

[62] D.-M. Tsai and S.-C. Lai, "Independent component analysis-based background subtraction for indoor surveillance," *IEEE Trans on Image Processing*, vol. 18, no. 1, pp. 158–167, 2009.

[63] J. Rittscher, J. Kato, S. Joga, and A. Blake, "A probabilistic background model for tracking," in *ECCV*, vol. 2, 2000, pp. 336–350.

[64] D. Kottow, M. Koppen, and J. R. del Solar, "A background maintenance model in the spatial-range domain," in *ECCV Workshop on Statistical Methods in Video Processing*, 2004, pp. 141–152.

[65] K. Kim, T. Chalidabhongse, D. Harwood, and L. Davis, "Real-time foreground-background segmentation using codebook model," *Real-Time Imaging*, vol. 11, no. 3, pp. 172–185, 2005.

[66] V. Mahadevan and N. Vasconcelos, "Background subtraction in highly dynamic scenes," in *CVPR*, 2008, pp. 1–6.

[67] R. Cucchiara, C. Grana, M. Piccardi, and A. Prati, "Detecting moving objects, ghosts, and shadows in video streams," *TPAMI*, vol. 25, no. 10, pp. 1337–1342, 2003.

[68] P. Varcheie, M. Sills-Lavoie, and G.-A. Bilodeau, "An efficient region-based background subtraction technique," in *Canadian Conf on Computer and Robot Vision*, 2008, pp. 71–78.
[69] R. Abbott and L. Williams, "Multiple target tracking with lazy background subtraction and connected components analysis," *Machine Vision and Applications*, vol. 20, pp. 93–101, 2009.
[70] E. Martinez-Martin and A. del Pobil, *Robust Motion Detection in Real-Life Scenarios*. Springer, 2012.
[71] I. Aljarrah, A. Ghorab, and I. Khater, "Object recognition system using template matching based on signature and principal components analysis," *Intl Journal of Digital Information and Wireless Communications*, vol. 2, no. 2, pp. 156–163, 2012.
[72] A. Mohammed, R. Minhas, Q. J. Wu, and M. Sid-Ahmed, "Human face recognition based on multidimensional {PCA} and extreme learning machine," *Pattern Recognition*, vol. 44, no. 10–11, pp. 2588–2597, 2011.
[73] J. Yang and D. Zhang, "Two-dimensional pca: A new approach to appearance-based face representation and recognition," *TPAMI*, vol. 26, no. 1, 2004.
[74] J. Jones and A. Palmer, "An evaluation of the two-dimensional gabor filter model of simple receptive fields in cat striate cortex," *J Neurophysiology*, vol. 58, no. 6, pp. 1233–1258, 1987.
[75] Y. Cho, S. Bae, Y. Jin, K. Irick, and V. Narayanan, "Exploring gabor filter implementations for visual cortex modeling on fpga," in *Intl Conf on Field Programmable Logic and Applications (FPL)*, 2011, pp. 311–316.
[76] N. Pinto, D. Cox, and J. DiCarlo, "Why is real-world visual object recognition hard?" *PLoS Comput Biol*, vol. 4, no. 1, p. e27, 2008.
[77] N. Petkov and P. Kruizinga, "Computational models of visual neurons specialised in the detection of periodic and aperiodic oriented visual stimuli: bar and grating cells," *Biological Cybernetics*, vol. 76, no. 2, pp. 83–96, 1997.
[78] E. Martinez-Martin, A. del Pobil, M. Chessa, F. Solari, and S. Sabatini, "An active system for visually-guided reaching in 3d across binocular fixations," *The Scientific World Journal*, vol. 2014, 2014.
[79] E. Martinez-Martin, A. del Pobil, M. Chessa, F. Solari, and S. Sabatini, "An integrated virtual environment for visual-based reaching," in *ACM Intl Conf on Ubiquitous Information Management and Communication*, 2011.
[80] Y. Benezeth, P. Jodoin, B. Emile, H. Laurent, and C. Rosenberger, "Review and evaluation of commonly-implemented background subtraction algorithms," in *ICPR*, 2008, pp. 1–4.
[81] K. Lai, L. Bo, X. Ren, and D. Fox, "A large-scale hierarchical multi-view rgb-d object dataset," in *ICRA*, 2011, pp. 1817–1824.
[82] T. Malisiewicz and A. Efros, "Recognition by association via learning per-examplar distances," in *CVPR*, 2008.
[83] L. Bo, X. Ren, and D. Fox, "Unsupervised Feature Learning for RGB-D Based Object Recognition," in *Intl Symp on Experimental Robotics*, 2012.
[84] M. Blum, J. Springenberg, J. Wulfing, and M. Riedmiller, "A learned feature descriptor for object recognition in rgb-d data," in *ICRA*, 2012, pp. 1298–1303.
[85] L. Bo, X. Ren, and D. Fox, "Depth Kernel Descriptors for Object Recognition," in *IROS*, 2011.
[86] K. Lai, L. Bo, X. Ren, and D. Fox, "Sparse distance learning for object recognition combining rgb and depth information," in *ICRA*, 2011, pp. 1817–1824.
[87] J. Ferreira, J. Lobo, and J. Dias, "Bayesian real-time perception algorithms on gpu. real-time implementation of bayesian models for multimodal perception using cuda," *J Real-Time Image Processing*, vol. 6, no. 3, pp. 171–186, 2011.
[88] K. Amano, N. Goda, S. ya Nishida, Y. Ejima, T. Takeda, and Y. Ohtani, "Estimation of the timing of human visual perception from magnetoencephalography," *J Neuroscience*, vol. 26, pp. 3981–3991, 2006.
[89] A. Jain, R. Bansal, and K. Singh, "A comparative study of visual and auditory reaction times on the basis of gender and physical activity levels of medical first year students," *Intl J Applied and Basic Medical Research*, vol. 5, pp. 124–127, 2015.
[90] (2016) Human benchmark project. [Online]. Available: http://www.humanbenchmark.com
[91] J. Nielsen, *Usability Engineering*. Academic Press, Inc., 1993.
[92] (2010) Website response times. [Online]. Available: https://www.nngroup.com/articles/website-response-times/
[93] D. Norman, *The Design of Everyday Things*. Basic Books, 2013.

Ester Martinez-Martin is an Assistant Professor at Universitat Jaume I and member of the Robotic Intelligence Laboratory. She has a M.Sc. in Computer Science with highest honors from Jaume-I University (2004), awarded for Academic Excellence. In 2011, she received the Ph.D. in Computer Science with a thesis titled Computer Vision Methods for Robot Tasks: Motion Detection, Depth Estimation and Tracking, supervised by Prof. Angel P. del Pobil. Her education is completed with postgraduate certificates in different topics such as computer design, programming languages and web design. She has been involved in robotics research, being part of several national and international projects. Her research background has been extended by participating in several relevant international conferences and schools, as well as with research stays at University of Genoa, SungKyunKwan University and University of Vienna. She is author and co-author of relevant scientific publications including a whole book (*Robust motion detection in real-life scenarios* (Springer 2012)) and editor of two books. She has been the Organization Chair of the IEEE-RAS Summer School on Experimental Methodology, Performance Evaluation and Benchmarking in Robotics (2015), SAB 2014, and 12th International UJI Robotics School on Perceptual Robotics for Humanoids (2012), and has collaborated in some conference committees and outreach activities. She recently received the Santander Award for Young Researchers in the field of engineering and architecture.

Angel P. del Pobil Angel P. del Pobil is a professor at Jaume I University (Spain), where he was the founding director of RobInLab, the UJI Robotic Intelligence Laboratory, and a visiting professor at Sungkyungkwan University (Korea). He is currently co-Chair of the IEEE RAS Technical Committee on Performance Evaluation & Benchmarking of Robotic Systems, and a member of the Governing Board of the Intelligent Autonomous Systems (IAS) Society and EURON (European Robotics Research Network of Excellence, 2001-2009). He has over 250 publications, including four authored and nine edited books. Prof. del Pobil was co-organizer of some 50 workshops and tutorials at ICRA, IROS, RSS, ROMAN, ICAR and HRI. He has been Program or General Chair of international conferences such as Adaptive Behaviour (SAB 2014), or Artificial Intelligence and Soft Computing. He serves regularly as Associate Editor for ICRA and IROS, and on the program committee of over 150 international conferences, such as IJCAI, ICPR, IAS, SAB, ICDL-EPIROB, ROMAN, etc. RobInLab has organized 12 consecutive editions of IURS, the International UJI Robotics School, and he has also served as General Chair of the 2015 IEEE-RAS Summer School on Experimental Methodology, Performance Evaluation and Benchmarking in Robotics. He has been involved in robotics research for the last 30 years. Professor del Pobil has been invited speaker of 63 tutorials, plenary talks, and seminars in 15 countries. He serves as associate or guest editor for 12 journals, and as expert for research evaluation at the European Commission. He has supervised 16 Ph.D. Thesis, including winner and finalists of the Georges Giralt PhD Award and the Robotdalen Scientific Award. He has been Principal Investigator of 30 research projects. Del Pobil is an active member of RAS and a lifetime member of the Association for the Advancement of Artificial Intelligence (AAAI).