

Object Detection and Recognition for Assistive Robots

Experimentation and Implementation



By Ester Martinez-Martin
and Angel P. del Pobil

Digital Object Identifier 10.1109/MRA.2016.2615329
Date of publication: 21 March 2017

Technological advances are being made to assist humans in performing ordinary tasks in everyday settings. A key issue is the interaction with objects of varying size, shape, and degree of mobility. 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 article, we present a vision system for assistive robots that is able to detect and recognize objects from a visual input in ordinary environments in real time. The system computes color, motion, and shape cues, combining them in

©ISTOCKPHOTO.COM/STAFENKOOLENA

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 the number of target objects in the scene.

Approaches for Object Perception

Today, robots have found their way from sealed workstations 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, health-care 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 multifunctional 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 those elements 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 analyzed in [11]. 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 during its life. Among all these capabilities, this article is focused on perception for object detection and recognition, a key task for meaningful assistance.

In this context, vision is considered a primary cue because of the information it can provide. Vision has been used in numerous robotic applications to successfully achieve a task (e.g., obstacle avoidance for navigation [12]–[15], human recognition for human-robot interaction [16], [17], activity recognition for cooperative behavior

[18]–[20], and object identification for manipulation [21]–[23], to name only a few). However, despite significant achievements, the problem of detecting and recognizing 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 and size. 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, as considerably more data need 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 recognize shapes and objects is by means of model-based approaches [24]–[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 learned 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 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 kinds 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, color, or corner cues [27]–[29]; steerable filters [30]; Haar-like features [31]; or scale-invariant descriptors (e.g., scale-invariant feature transform, speeded-up robust features) [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. 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 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 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. These data are 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 effort, 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. Depending on the perceived elements, a task can be executed or not because 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 color; 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 presegmentation 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 recognized 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, color-based invariant gradients have been combined with a histogram of oriented gradient 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 previously, we present new scientific results with a focus on working systems. Our robot system is capable of detecting and recognizing 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 color, motion, and shape cues and combines them in a probabilistic manner to adequately detect and recognize objects. Moreover, a 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.

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 a biologically inspired object recognition system (BIORecS) emulating the human vision system. 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 (SVM) is used as object shape classifier], recognition [carried out by principal components analysis (PCA)], and image database (containing the objects to be recognized).

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 because different objects can have the same visual shape. For example, a ball, a bracelet, a disk, a coin, and 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 because 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 behavioral reorienting module is responsible for detecting novel and behaviorally 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 that 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 centered 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 recognized should also be integrated. Our vision system consists of three different modules (Figure 1):

- *feature extraction*: generates a saliency map from image segmentation based on three object properties—color, shape, and motion
- *memory*: stores the models of the potential target objects
- *recognition*: responsible for recognizing the objects from the visual input and the data coming from the previous modules.

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

Feature Extraction

Visual features are a key point in any detection and recognition procedure. Deciding what features are required to properly detect and recognize a target object in place of others is not an easy task. The reason lies in the fact that a wide variety of features would result in very time-consuming processing, while a poor feature-based object description would lead to an inefficient recognition. 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 behavior and task relevance. In this work, three distinctive feature types are considered: color, motion, and shape. 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.

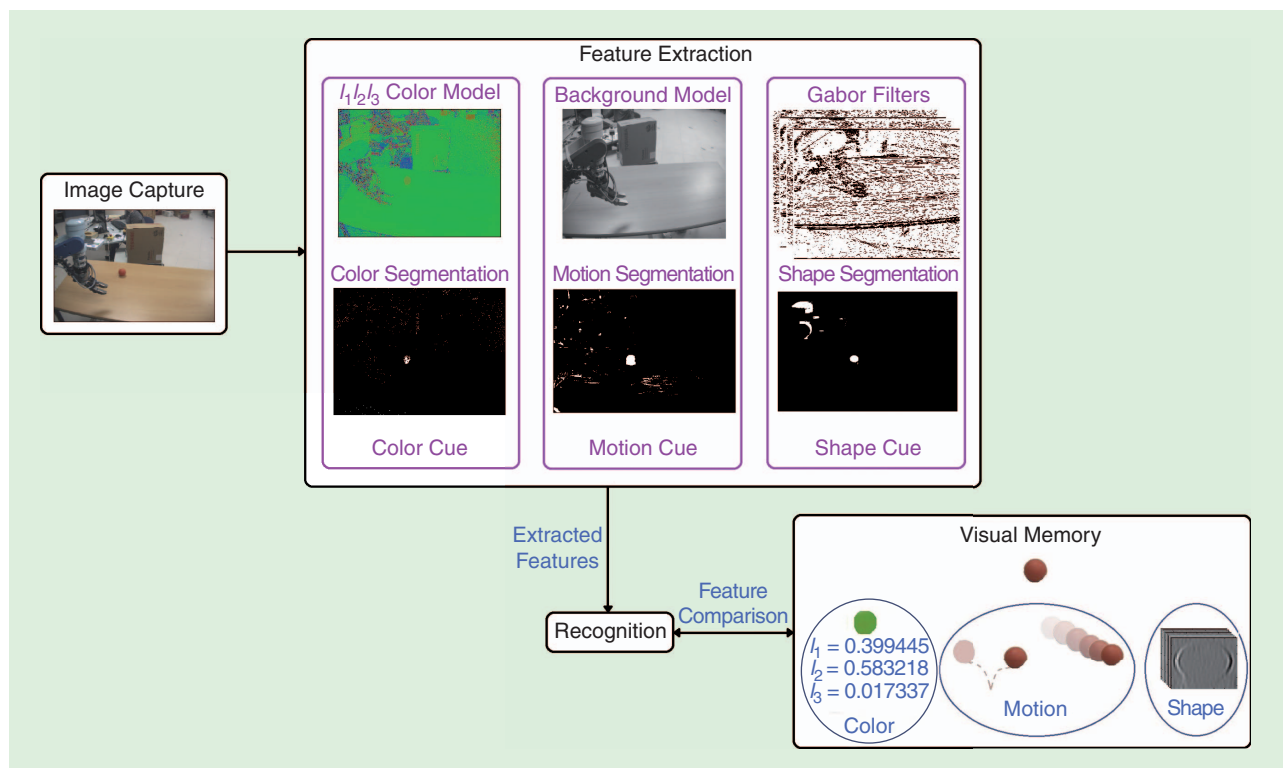


Figure 1. An overview of the system architecture for object detection and recognition, showing its three main modules (feature extraction, memory, and recognition) and the threefold object description (color, motion, and shape).

Color Cues

Vision 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, color histograms can be used to represent and match images or objects. However, despite its simplicity and efficacy, color's accuracy is significantly deteriorated when the illumination conditions change. As an alternative, the color gradient obtained from the addition of channel derivatives could be considered. Nevertheless, given that the color derivatives are separately computed, differences in the color edge directions can make this technique fail miserably.

Another possibility could be to use a different color model. A large variety of color spaces are normally used for different purposes, such as video and television [luminance in-phase quadrature (YIQ) and luminance, blue-luminance, red-luminance (YUV)], display and printing [red, green, blue (RGB) and cyan, magenta, yellow, and key (CMYK)], perceptual uniform spaces ($U^*V^*W^*$, $L^*a^*b^*$, Luv), human perception (hue, saturation, and intensity), or standard primary colors (rgb , xyz). However, a large number of these color models are combinations of RGB (e.g., CMYK, XYZ, and $I_1I_2I_3$) 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^*).

Keeping in mind the goal of a visual system able to accurately detect and recognize multicolored objects in real scenes, existing color models have been analyzed to determine which one is more robust to changes in illumination, object geometry, and camera viewpoint. The aim is a color 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 analyzed diverse color models by evaluating their robustness for object recognition under different image parameters. This comparison concluded that the color model to be chosen depends on the imaging conditions. Indeed, if all the imaging conditions are controlled, RGB is the most invariant color model for object recognition. However, under the constraints of white illumination and no presence of highlights, normalized color rgb and $c_1c_2c_3$ are the most robust color 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.

Given that no environmental and object constraints are established, the $l_1l_2l_3$ color space is used in our system for object recognition due to 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. Nevertheless, with the aim of robustly detecting and recognizing objects in realistic scenarios, other cues must also be used.

Motion Cues

The capability of visually perceiving motion is a key issue in computer vision. 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. These kinds 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, specifically 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 it 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]. Afterward, most of the research focused on methods for background maintenance, i.e., 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 Pfister [52], in which a single Gaussian distribution was used; multimodal statistical models such as a mixture of Gaussians [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; and statistical models based on the minimum and maximum intensity values and the maximum interframe change (temporal derivative [59]). Other methods incorporate spatial region-based scene information, such as kernel density estimation, a Parzen-window estimate with a kernel [60], eigenbackground (eigenspace decomposition based on images of motionless backgrounds [61]), or independent component analysis [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 mentioned earlier [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 resolved, such as how to arrange for a training period with foreground objects in dynamic, real environments; the adaptation to minor dynamic, uncontrolled changes like the passage of time, blinking of a 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, we 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.

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 costs are critical issues, this control is performed by means of a combination of difference techniques: frame difference with reference frame subtraction. Frame difference allows the system to identify objects that have moved from one frame to the next one. However, it is important to take into account that both the previous position and the current one are detected. This problem was solved by using background subtraction because 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 to solve this kind of situation. Furthermore, the used thresholds for those subtraction approaches are automatically set for each pixel from pixel neighborhood 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. 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 color 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:

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

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 that depends on the point distribution.

2) The second process is the 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, as 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:

$$\begin{cases} \mu_{r,c}(t+1) = \begin{cases} (1-\alpha)\mu_{r,c}(t) + \alpha i_{t+1}(r, c) & \text{if background} \\ \mu_{r,c}(t) & \text{otherwise} \end{cases} \\ \sigma_{r,c}(t+1) = \begin{cases} (1-\alpha)\sigma_{r,c}(t) + \alpha i_{t+1}(r, c) & \text{if background} \\ \sigma_{r,c}(t) & \text{otherwise} \end{cases} \end{cases}. \quad (2)$$

Here, the constant α ($0 < \alpha < 1$) controls the adaptation rate, and it is given by the number of pixels that are part of the Gaussian distribution. However, sometimes the pixel gray 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. During the updating phase, two different tasks are carried out.

- The background model is being updated with each new frame by using (2).

- 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.

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, PCA has been widely used (e.g., [71]–[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 preprocessed 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]–[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 recognized. What is more, phase-based methods have been shown to be robust to changes in contrast, scale, and orientation [78], [79]. Therefore, a symmetrical filter kernel and an antisymmetrical filter kernel 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:

$$\begin{cases} I_{\sin, \sigma}(x, \omega) = \int \omega \left(\frac{x-x'}{\sigma} I(x') \sin(\omega(x-x')) \right) dx' \\ I_{\cos, \sigma}(x, \omega) = \int \omega \left(\frac{x-x'}{\sigma} I(x') \cos(\omega(x-x')) \right) dx' \end{cases} \quad (3)$$

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, (3) describes a convolution with Gabor functions.

The proposed method extracts the object shape using a bank of eight oriented Gabor filters. 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 recognize 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 is shown in Figure 2.

Memory

Memory performs a fundamental role in human object recognition. Similarly, in our system, a memory module stores



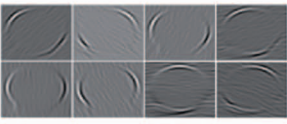





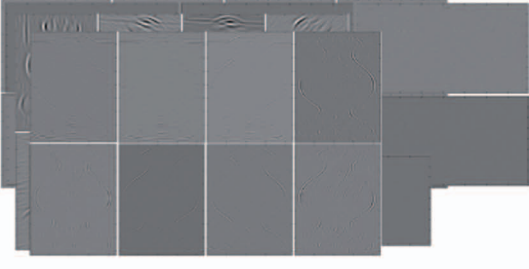
Object	Considered Shapes	Representation
	Front View 	
	Front  Sideways  Top  Perspective 	

Figure 2. The shape cue in terms of Gabor filters based on four shape representations (front, sideways, top, and perspective) used in the proposed approach.

the description of all the potential targets to be recognized. It contains all the features integrating the description of each known object, as shown in Figure 3.

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, 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 article, object identification is addressed because 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 (color, motion, and shape), the object-based likelihood can be obtained as follows:

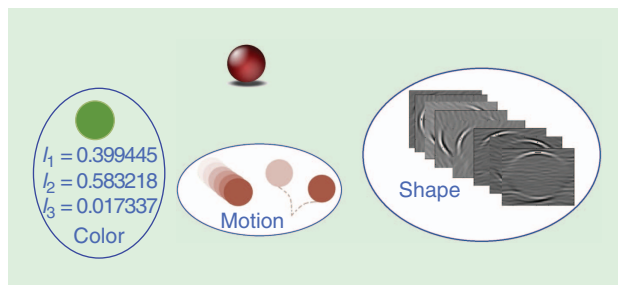


Figure 3. The object description in terms of color, motion, and shape properties saved in the system memory for proper object detection and recognition.

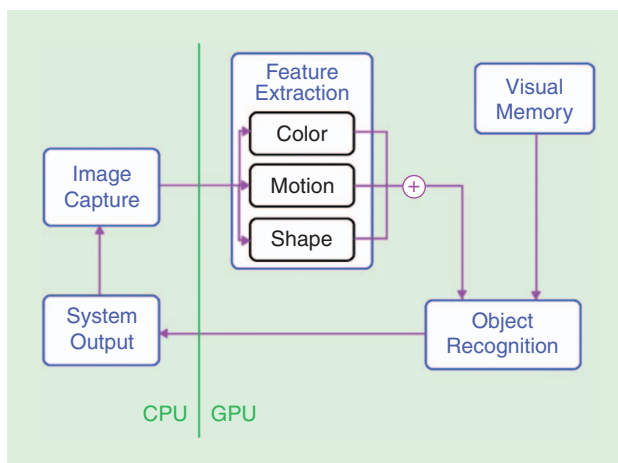


Figure 4. An overview of our CPU-GPU implementation, meeting real-time performance by parallelly implementing on the GPU the computationally intensive task of object detection and recognition.

$$P(I|o) = P(I_c|o) | P(I_m|o) | P(I_s|o), \quad (4)$$

where $P(I_c|o)$, $P(I_m|o)$, and $P(I_s|o)$, respectively, correspond to color-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 color as compared to shape, because a circular shape is very common in real-world scenarios and its discriminative value is lower.

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 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, the higher the computational costs).

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 main memories characterizes these CUDA cores. Moreover, graphics application programming interface 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 central processing unit (CPU)-GPU system implementation is shown in Figure 4. The CPU captures an image and uploads it to the GPU, which will perform the subsequent image processing steps 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.

Because object feature detection and tracking is a computationally intensive but highly parallelizable task, a good parallel solution can be devised to the effect that all image processing is carried out by the GPU (using 1,023 threads per block). As a final system output, the CPU shows on the screen the detected objects.

Experimental Results

The proposed approach for object detection and recognition in real scenarios has been tested in three different kinds of scenarios. First, a semistructured 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.). Second, a set of experiments involved two real, cluttered

environments in which the target objects were to be found among a set of ordinary items such as calendars, books, clocks, or pens. Third, an image data set was 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 was presented.

For the two first experiments, a humanoid torso endowed with a Robosoft TO40 pan-tilt-vergence stereo head and two multijoint arms was used (see Figure 5). The head mounted two Imaging Source DFK 31BF03-Z2 cameras acquiring color images at 30 Hz with a resolution of $1,024 \times 768$ pixels. The baseline between cameras was 270 mm, and the motor positions were provided by high-resolution optical encoders.

Experiment 1: Semistructured Scenes

In the case of semistructured 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. In this experiment, the three visual cues had the same weight when the segmentation result was determined. Four different objects were 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°). As depicted in Figure 6, the implemented approach started with capturing an image. This image was the input of two different processes: the color cue segmentation and the segmentation of the other two considered cues (i.e., motion and shape).

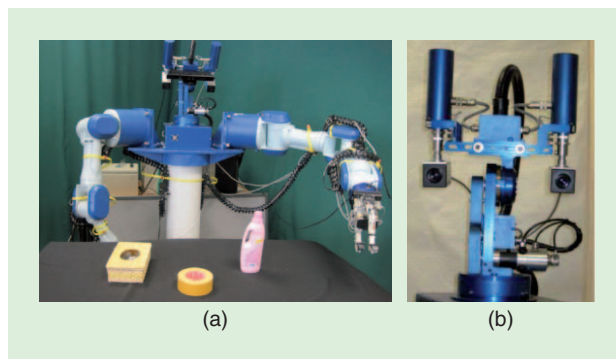


Figure 5. (a) An external view of the humanoid torso employed for the experiments and (b) a detailed view of the pan-tilt-vergence head.

This distinction was for efficiency reasons. Therefore, on the one hand, the image was expressed in $l_1 l_2 l_3$ coordinates and segmented by using the memory information about the different objects to be found. On the other hand, an intensity image was obtained with the purpose of speeding up motion and shape segmentation. Note that shape detection was obtained from the combination of the eight Gabor-filtered images. Once segmentation for each cue was performed, their fusion allowed the system to reduce the search area for object recognition and, despite the presence of factors such as shadows or reflections, the red ball was properly detected in the image.

In a similar way, experiments with the other objects were carried out. Figure 7 shows some of the obtained results (only the final result). Note that the illustrated results correspond to a single trial because there is no randomness in the data. As shown, only one object was searched each time. The reason lies in the performance analysis in the presence of different factors susceptible to making the system fail (e.g., shadows, flickering light sources, variable light reflexes, objects partially visible, etc.). As shown, all the objects were successfully recognized even when they changed their

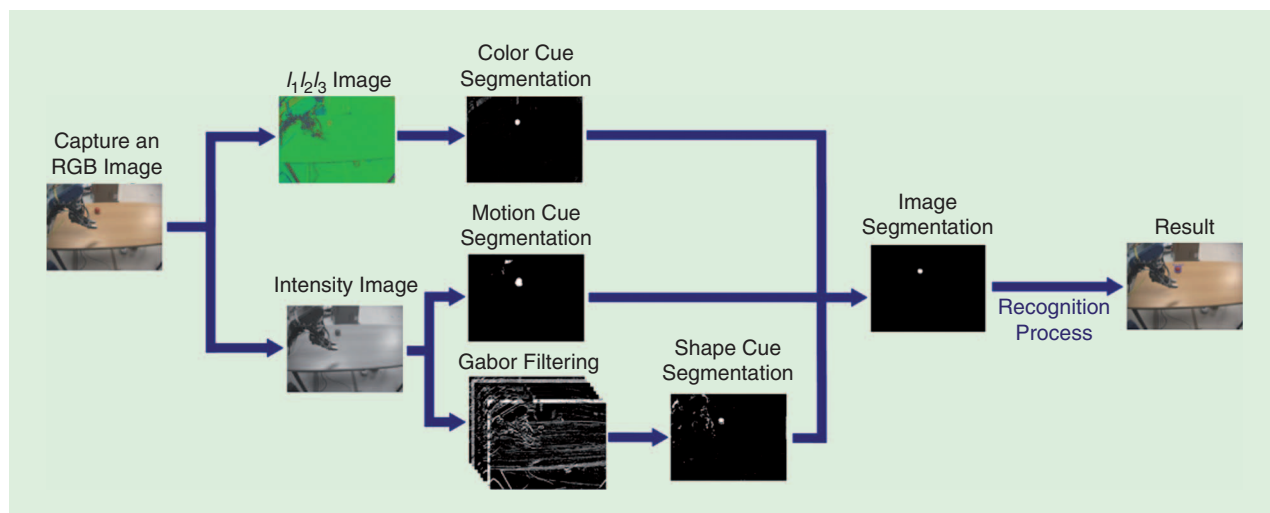


Figure 6. The object detection and recognition process in semistructured scenes.

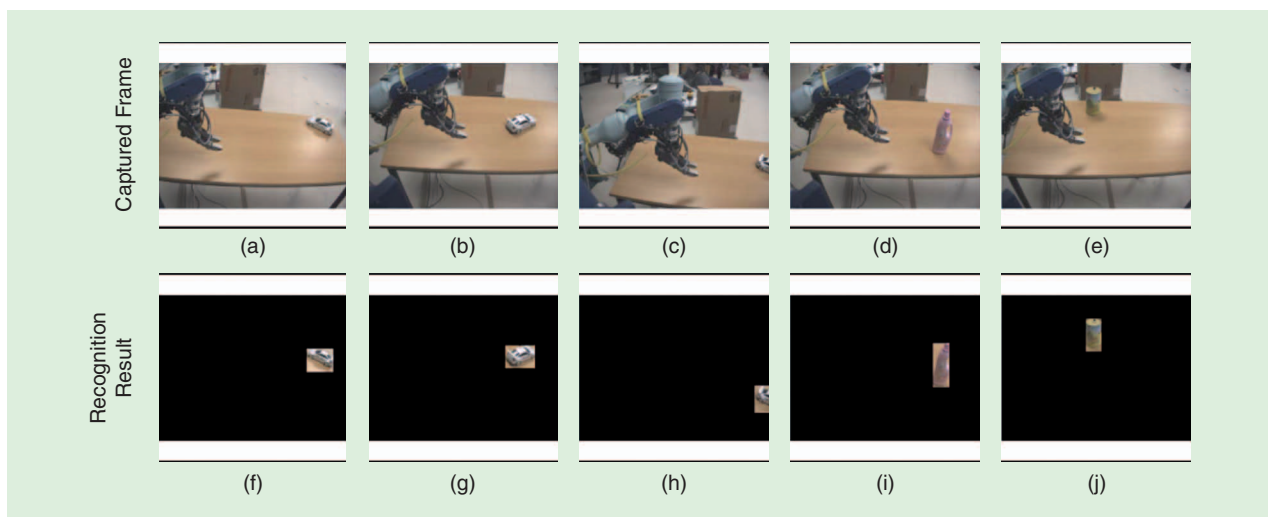


Figure 7. The qualitative experimental results in a semistructured scenario in which both (a)–(e) the camera viewpoint and (f)–(j) the object location and orientation were continuously changed. All of the objects were successfully recognized even when their orientation, location, or the camera’s viewpoint was changed.

orientation or location in the scene or the cameras changed their viewpoint.

With the purpose of validating the obtained qualitative results, a quantitative evaluation was carried out. In this case, the true positive rate (TPR) and false positive rate (FPR) measurements were used [80], i.e., the proportion of correctly classified positives (TPR) and the proportion of incorrectly classified negatives (FPR). So, a good performance was obtained when the TPR was close to 1 while the FPR was near 0. The performance analysis of the proposed approach resulted in a 0.780 TPR and a 0.001 FPR, which indicates good performance and that the system was successful in the object detection/recognition task.

Experiment 2: Real Scenarios

In this experiment, the objects to be detected and recognized were placed on a desk. Two unstructured environments were used, composed of everyday objects of different nature and features, such as textured books, pens, and clocks. In this context, the objects to be detected and recognized included 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 in partial occlusion in some cases. As in the previous case, a human was continuously interacting with the target objects but not with the robot system, so that the motion cue triggered a visual attention focus. However, the other two visual cues were required to distinguish between the target objects and other moving elements in the scene, such as the person. For this reason, the three cues had 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 8. 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 color similarity with its background. Figure 8(i) also shows an example of the simultaneous recognition of two objects in the same image, in which the car and the stapler were correctly detected and recognized. In a similar way, the developed approach adequately focused 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 Figure 8(j).

In the subsequent experiment, the visual system was aimed at detecting and recognizing four objects (a red ball, a green bulb, a yellow ball, and a wooden generalized cylinder) while a person was interacting with the objects in the scene, changing their positions on the desk. As a consequence, the motion cue again played a main role in the object recognition process. Some of the obtained one-trial results are presented in Figure 9. In this case, unlike in previous examples, the binary image is shown, highlighting the detected objects, especially when they are partially occluded or color similarity with the background is considerably high.

Once more, a quantitative analysis validates the prior mentioned qualitative results. In this case, the detection and recognition resulted in a 0.898 TPR and a 0.0009 FPR, confirming the quality of the approach performance when real-life scenarios are considered.

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. Given that the ability to recognize objects is crucial for many applications, a wide range of public image repositories are available. These data sets allow researchers to evaluate their approaches with a large number of objects and under different conditions as well as

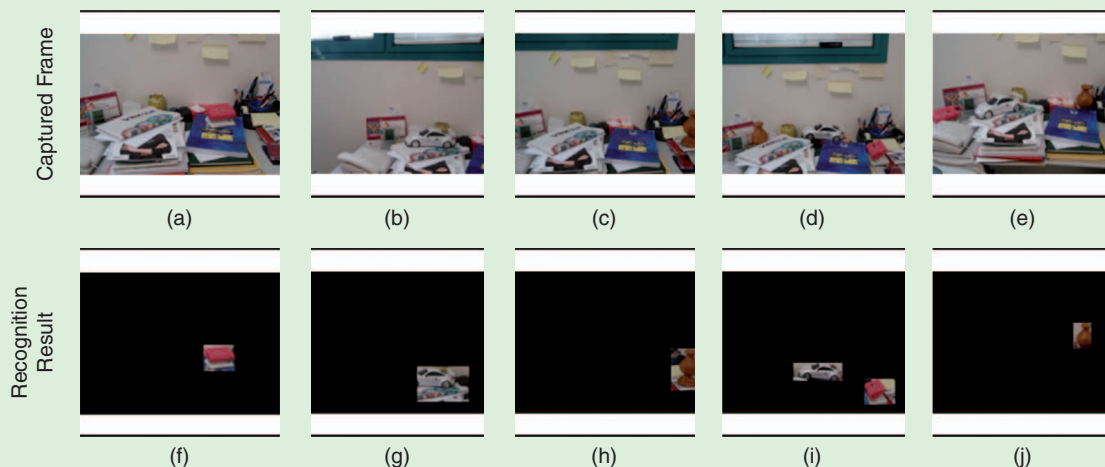


Figure 8. The qualitative experimental results when a real scenario is considered. (a) and (f) A red book is correctly identified. (b) and (g) A toy car is identified even while its color scheme is greatly similar to its background. (c) and (h) A generalized cylinder is correctly identified. (d) and (i) Two objects are simultaneously identified. (e) and (j) Focus is maintained on the target object, the generalized cylinder, even when several objects are added to the scene.

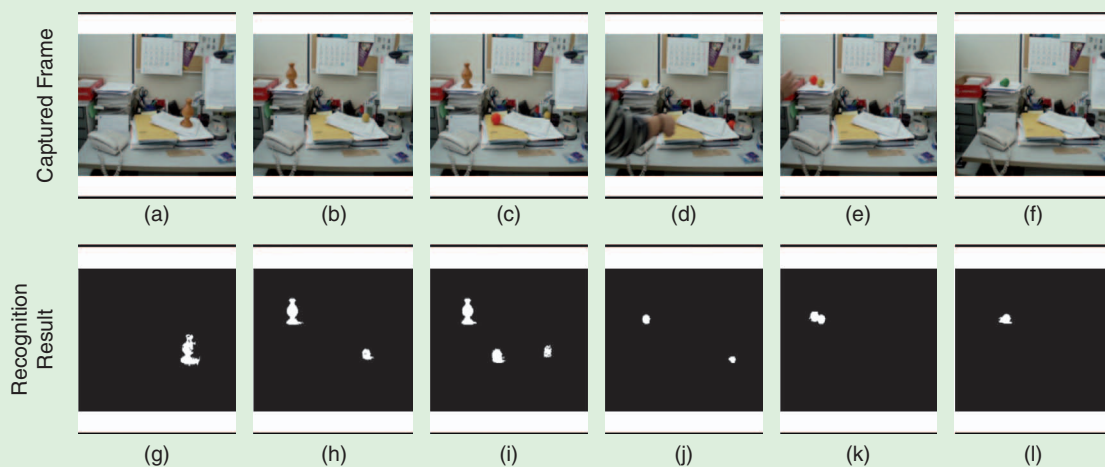


Figure 9. The qualitative experimental results when (a)–(f) a real scenario is considered. (g)–(l) In these binary images, the detected objects are highlighted, especially when they are partially occluded or if the object has a color scheme similar to its background.

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 data set must correspond to the needs of a particular approach. This is why the RGB-D Object Data Set [81], publicly available at <http://www.cs.washington.edu/rgbd-dataset>, has been used for this validation. This data set 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. Objects are organized into a

hierarchy of 51 categories composed of a number of instances between three and 14, so that each object belongs to only one category. In addition, ground truth images are provided to adequately assess the segmentation process. Consequently, this image data set allows object recognition techniques to be evaluated at at two levels: the category level and the instance level.

Despite the fact that the ability to recognize objects at both levels is a key point in the context of robotic tasks, in this work only the instance recognition is considered because no category abstraction was carried out. The task for the recognition algorithm was 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

Table 1. Accuracy comparison on the RGB-D Object Data Set when using alternating contiguous frames.

Approach	Accuracy Based on RGB Information
Exemplar-based local	84.5
Linear SVM	90.2
Nonlinear SVM	90.6
Random forest	90.5
Instance distance learning	91.3
CKM Desc	92.1
The proposed approach	96.1

on color and shape cues was used to build an object model that would be used for object detection and recognition. Note that, in this case, the motion cue was not used because objects were not moving, although the camera was.

For comparison reasons, we considered the cropped RGB-D frames that tightly included the object, exactly as used in the object recognition evaluation of the article introducing the RGB-D Object Data Set [81] (i.e., subsampled every fifth video frame). These were the images used for obtaining the different results over this image repository.

Table 1 compares the obtained results with those from different state-of-the-art approaches: exemplar-based local [82], linear SVM, Gaussian kernel SVM, random forests, kernel descriptors, convolutional k-means descriptors (CKM Desc), hierarchical matching pursuit, and instance distance learning, described in [83]–[86]. As can be observed from the results, our technique substantially

improved upon the performance of the several considered state-of-the-art classification approaches.

In addition, the RGB-D image data set 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. The proposed algorithm was tested in those common indoor environments. Some of the obtained results are illustrated in Figure 10. The first two images show an office and, although the scene illumination and the point of view changed, they corresponded to the same video sequence. As shown, the cellophane box was recognized in both of them, highlighting the approach robustness to lighting changes. Furthermore, Figure 10(j)–(l) refers to different scenarios with the same target object, a green bowl. As is apparent, it was properly detected, even when it was partially occluded.

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 nonparallel computing depending on the image resolution and the number of potential targets.

First, the execution time was analyzed for different image resolutions. Our results show that a similar performance was obtained with the two methods when the image resolution was low. However, when the image resolution was increased, the nonparallel computing time drastically climbed, while the GPU implementation showed a gradual,

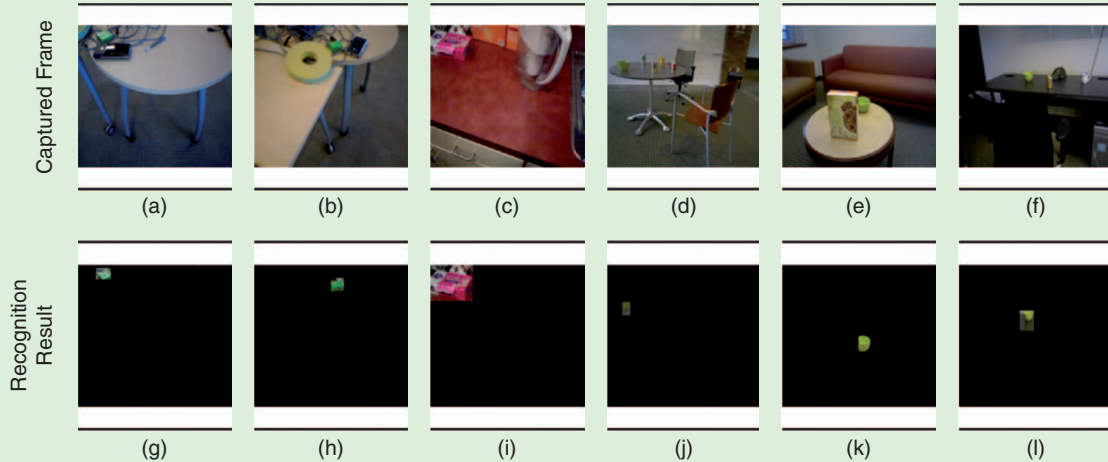


Figure 10. The object recognition results on the real-life scenarios provided by the RGB-D image data set. (a) and (b) Although there was a change in the camera's point of view for this office scene, (g) and (h) the cellophane box was recognized both times. (c) and (i) The pink box was successfully recognized. The same target object, (d)–(f) a green bowl, was (j)–(l) properly detected each time, even when it was partially occluded.

much slower, growth. This is apparent in Figure 11, which plots the speedup with respect to image size. In fact, the execution time for the GPU remained virtually constant (around 0.48 s) for the first ten image resolutions considered because the thread loads remained similar. Given that the number of threads was limited, when the image resolution was increased, both the thread workload and, consequently, the execution time rose, resulting in 0.95 s for our higher resolution ($1,600 \times 1,200$).

Another key issue in practical object recognition is that of scalability, and our last experiment analyzed the execution time when the number of potential target objects was increased. With that aim, different image sequences from the RGB-D image data set were used. The results, shown in Figure 12, illustrate the speedup evolution for an averaged image resolution of 84×85 pixels when the number of objects that could be found in the scene increased. As can be observed, our results highlight the efficiency when parallel computing is used; computation times remained almost unchanged between one object (0.46 s) and 50 target objects (0.47 s). 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 was the case, and, ideally, this response time should be the same as that of human beings. As our results show, the obtained response time was similar in all the studied cases (up to 50 target objects) and below 0.5 s, approximately twice the average human reaction time (between 200 and 250 ms [88]–[90]). In the context of human–computer interaction [91]–[93], a response time below 0.1 s is regarded as an instantaneous reaction, whereas a response delay between 0.1 and 1.0 s is considered as fast enough for a fluent interaction, even though the user would notice the delay. Consequently, a response time of 0.5 s 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 than $1,600 \times 1,200$, execution times would be presumably beyond 1.0 s. In this case, image resolution could be decreased by using, for instance, pyramidal images to obtain real-time performance.

Conclusions and Future Work

During recent 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.

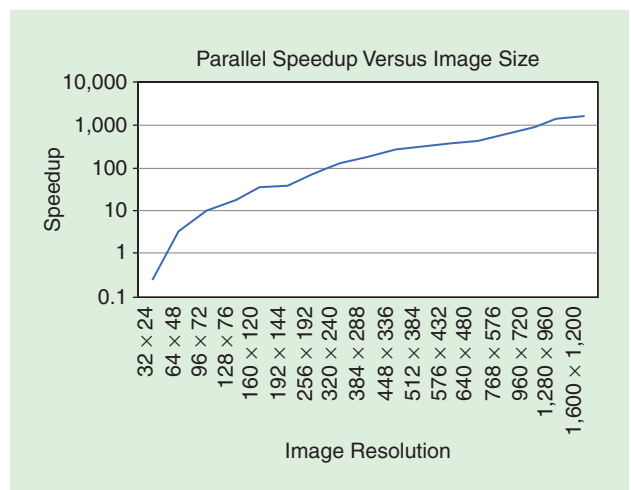


Figure 11. The speedup versus image size for parallel (GPU) and nonparallel (CPU) computing.

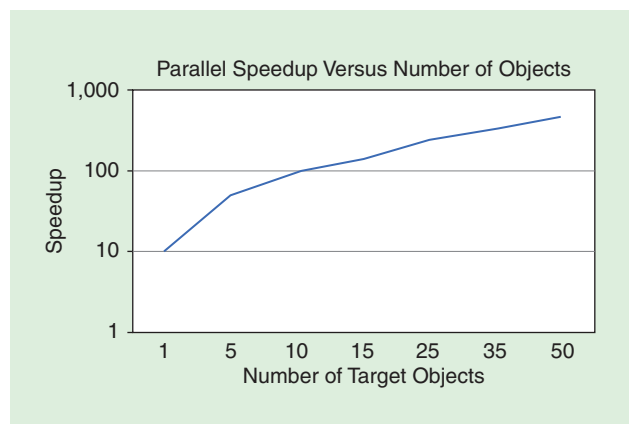


Figure 12. The speedup versus the number of potential target objects for parallel (GPU) and nonparallel (CPU) computing.

In this article, we 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 recognize objects from a visual input in unconstrained scenes in real time. We took inspiration from biology and used a rich object description based on color, motion, and shape cues. Robust color information was obtained thanks to an adequate color model choice that made visual data invariant to changes in viewpoint, object geometry, and illumination. The second considered cue was motion, which was perceived by means of a novel background maintenance technique overcoming the environmental constraints of existing methods. Finally, a phase-based representation of shape concluded the object description presented in this article.

Once the visual features were properly extracted, the system analyzed the statistical similarity between the detected

objects and those whose description were stored in the system's visual memory. This estimated joint likelihood allowed the system to successfully discriminate between

Assistive robots must be endowed with different abilities, such as localization, mapping, path planning, obstacle avoidance, object detection, recognition, and manipulation.

several objects. Furthermore, with the purpose of effectively achieving real-time computation in visual data processing, a GPU was used by taking into account that irregular workloads are common in the task at hand.

The proposed approach was implemented on a robotic platform and tested by considering different parameters that might make the system fail. This large number of parameters allowed us to analyze the robustness of the proposed method. For further experimental vali-

dation, a public image repository for object recognition was 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 was able to accurately detect and recognize 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 to automatically learn new objects, going a step further in emulating the human visual system.

Acknowledgments

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

References

- [1] E. Martinez and A. del Pobil, "Visual surveillance for human-robot interaction," in *Proc. IEEE Int. Conf. Systems, Man and Cybernetics (SMC)*, 2012, pp. 3333–3338.
- [2] E. Martinez and A. del Pobil, "Safety for human-robot interaction in dynamic environments," in *Proc. IEEE Int. Symp. on Assembly and Manufacturing, 2009 (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. Intell. Robotic Syst.*, vol. 66, pp. 245–272, 2012.
- [4] HOBBIT—The mutual care robot. (2013). [Online]. Available: <http://hobbit.acin.tuwien.ac.at/>
- [5] KSERA—Knowledgeable service robots for aging. (2013). [Online]. Available: <http://ksera.ieis.tue.nl/>
- [6] COGNIRON—The cognitive robot companion. (2007). [Online]. Available: <http://www.cogniron.org/final/Home.php>
- [7] Care-O-bot. (2015). [Online]. Available: <http://www.care-o-bot-4.de/>
- [8] HERB. (2015). [Online]. Available: <http://www.cmu.edu/herb-robot/>
- [9] ACCOMPANY. (2015). [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, no. 3, pp. 5654–5676, Mar. 2014.
- [11] I. Leite, C. Martinho, and A. Paiva, "Social robots for long-term interaction: A survey," *Int. J. Social Robotics*, vol. 5, no. 2, pp. 291–308, Apr. 2013.
- [12] F. Bonin-Font, A. Ortiz, and G. Oliver, "Visual navigation for mobile robots: A survey," *J. Intelligent and Robotic Syst.*, vol. 53, no. 3, pp. 263–296, May 2008.
- [13] J. Antich, A. Ortiz, and G. Oliver, "A control strategy for fast obstacle avoidance in troublesome scenarios: Application in underwater cable tracking," presented at the IFAC Conf. Maneuvering 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 *Proc. 2006 IEEE/RSJ Int. Conf. Intelligent Robots and Systems*, pp. 2302–2307.
- [15] J. Shen and H. Hu, "Visual navigation of a museum guide robot," in *Proc. Sixth World Congr. Intelligent Control and Automation, 2006 (WCICA 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 *Proc. 2010 IEEE Int. Conf. Robotics and Biomimetics (ROBIO)*, pp. 573–578.
- [17] P. Turaga, R. Chellappa, V. Subrahmanian, and O. Udrea, "Machine recognition of human activities: A survey," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 11, pp. 1473–1488, Nov. 2008.
- [18] O. Chang, "Evolving cooperative neural agents for controlling vision guided mobile robots," in *Proc. Int. Conf. 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," *Artif. Intell.*, vol. 110, no. 2, pp. 275–292, Jan. 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 *Proc. IEEE/RSJ/GI Int. Conf. Intelligent Robots and Systems '94 (IROS)*, pp. 925–932.
- [21] D. Kragic and H. Christensen, "Survey on visual servoing for manipulation," Computational Vision and Active Perception Laboratory, Royal Institute of Technology, Stockholm, Sweden, 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 *Robotics Research*. New York: Springer-Verlag, 1998, pp. 45–54.

- [23] D. Ünay, Z. Çataltepe, and S. Aksoy, Eds., *Recognizing Patterns in Signals, Speech, Images and Videos*. Berlin, Germany: Springer-Verlag, 2010.
- [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 *Proc. 2007 IEEE Int. Conf. Robotics and Automation*, pp. 3773–3779.
- [25] N. Sian, T. Sakaguchi, K. Yokoi, Y. Kawai, and K. Maruyama, "Operating humanoid robots in human environments," in *Proc. 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 *Proc. IEEE RAS Int. Conf. Humanoid Robots*, 2006, pp. 94–101.
- [27] C. Urdiales, M. Dominguez, C. de Trazegnies, and F. Sandoval, "A new pyramid-based color image representation for visual localization," *Image Vis. Comput.*, 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 *Proc. 9th Information Technology and Telecommunication Conf.*, 2009.
- [29] Z. Kim, "Real time object tracking based on dynamic feature grouping with background subtraction," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition, 2008 (CVPR 2008)*, pp. 1–8.
- [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 *Proc. Int. Conf. Pattern Recognition*, 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, pp. 127–133, Apr. 2006.
- [32] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
- [33] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "SURF: Speeded up robust features," *Comp. Vision Image Understanding*, 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 *Proc. Int. Cognitive Robotics Workshop (CogRob)*, 2008, pp. 32–41.
- [35] P. Fitzpatrick and G. Metta, "Grounding vision through experimental manipulation," *Philos. Trans R. Soc. London A, Math. Phys. Sci.*, 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," *Knowl.-Based Syst.*, 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*. Berlin, Germany: Springer-Verlag, 2009, pp. 235–244.
- [38] L. Yang, H. Cheng, J. Hao, Y. Ji, and Y. Kuang, "A survey on media interaction in social robotics," in *Advances in Multimedia Information Processing*. New York: Springer-Verlag, 2015, pp. 181–190.
- [39] H. Yan, M. H. Ang Jr., and A. N. Poo, "A survey on perception methods for human-robot interaction in social robots," *Int. J. Social Robotics*, vol. 6, no. 1, pp. 85–119, Jan. 2014.
- [40] A. Palomino, R. Marfil, J. Bandera, and A. Bandera, "A new cognitive architecture for bidirectional loop closing," in *Proc. Robot 2015: Second Iberian Robotics Conf.: Advances in Robotics*, 2015, pp. 721–732.
- [41] S. Wolfson and M. Landy, "Examining edge- and region-based texture analysis mechanisms," *Vis. Res.*, 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 *Proc. 2009 IEEE Int. Conf. Robotics and Automation*, pp. 1997–2002.
- [43] E. Martinez-Martin and A. del Pobil, "Visual object recognition for robot tasks in real-life scenarios," in *Proc. 2013 10th Int. Conf. Ubiquitous Robots and Ambient Intelligence (URAI)*, pp. 644–651.
- [44] H. Al-Absi and A. Abdullah, "Biologically inspired object recognition system," in *Proc. 2010 Int. Symp. in Information Technology*, pp. 1–5.
- [45] P. Lanillos, J. Ferreira, and J. Dias, "Designing an artificial attention system for social robots," in *Proc. 2015 IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS)*, pp. 4171–4178.
- [46] J. Ferreira and J. Dias, "Attentional mechanisms for socially interactive robots—A survey," *IEEE Trans. Auton. Mental Develop.*, vol. 6, pp. 110–125, 2014.
- [47] T. Gevers and A. Smeulders, "Color-based object recognition," *Pattern Recognit.*, vol. 32, no. 3, pp. 453–469, 1999.
- [48] F. Orabona, G. Metta, and G. Sandini, "A proto-object based visual attention model," in *Attention in Cognitive Systems: Theories and Systems from an Interdisciplinary Viewpoint*. Berlin, Germany: Springer-Verlag, 2008, 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 *Proc. 2011 IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS)*, pp. 1526–1532.
- [50] R. Jain and H. Nagel, "On the analysis of accumulative difference pictures from image sequences of real world scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, 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 Univ., Pittsburgh, PA, Tech. Rep., 2000.
- [52] C. Wren, A. Azarbeyejani, T. Darrell, and A. Pentland, "Pfindex: Real-time tracking of the human body," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 780–785, 1997.
- [53] D. Koller, J. Weber, and J. Malik, "Robust multiple car tracking with occlusion reasoning," in *Proc. European Conf. Computer Vision 1994*, pp. 189–196.
- [54] K. Toyama, J. Krum, B. Brumitt, and B. Meyers, "Wallflower: Principles and practice of background maintenance," in *Proc. Seventh IEEE Int. Conf. Computer Vision (ICCV)*, 1999, pp. 255–261.
- [55] Z. Zivkovic, "Improved adaptive Gaussian mixture model for background subtraction," in *Proc. Int. Conf. Pattern Recognition*, 2004, pp. 28–34.
- [56] C. Stauffer and W. Grimson, "Learning patterns of activity using real-time tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 5, pp. 747–757, Aug. 2000.
- [57] N. Friedman and S. Russell, "Image segmentation in video sequences: A probabilistic approach," in *Proc. Conf. 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 *Proc. Int. Conf. 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," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 809–830, 2000.
- [60] A. Elgammal, D. Harwood, and L. S. Davis, "Non-parametric model for background subtraction," in *Proc. European Conf. Computer Vision 2000*, pp. 751–767.
- [61] N. Oliver, B. Rosario, and A. Pentland, "A Bayesian computer vision system for modeling human interactions," *IEEE Trans. Pattern Anal. Mach. Intell.*, 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. Image Process.*, 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 *Proc. Computer Vision—European Conf. Computer Vision 2000*, pp. 336–350.
- [64] D. Kottow, M. Koppen, and J. R. del Solar, "A background maintenance model in the spatial-range domain," in *Proc. European Conf. Computer Vision Workshop 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," *J. Real-Time Imaging*, vol. 11, no. 3, pp. 172–185, 2005.
- [66] V. Mahadevan and N. Vasconcelos, "Background subtraction in highly dynamic scenes," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition, 2008 (CVPR 2008)*, pp. 1–6.
- [67] R. Cucchiara, C. Grana, M. Piccardi, and A. Prati, "Detecting moving objects, ghosts, and shadows in video streams," *IEEE Trans. Pattern Anal. Mach. Intell.*, 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 *Proc. 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," *Mach. Vis. Appl.*, vol. 20, no. 2, pp. 93–101, Nov. 2009.
- [70] E. Martinez-Martin and A. del Pobil, *Robust Motion Detection in Real-Life Scenarios*. New York: Springer-Verlag, 2012.
- [71] I. Aljarrah, A. Ghorab, and I. Khater, "Object recognition system using template matching based on signature and principal components analysis," *Int. J. Digital Information 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 Recognit.*, 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," *IEEE Trans. Pattern Anal. Mach. Intell.*, 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. Neurophysiol.*, 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 *Proc. Int. Conf. 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," *Biol. Cybern.*, 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," *Sci. World J.*, vol. 2014, no. 10–11, p. 179,391, 2014.
- [79] E. Martinez-Martin, A. del Pobil, M. Chessa, F. Solari, and S. Sabatini, "An integrated virtual environment for visual-based reaching," in *Proc. ACM Int. Conf. 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 *Proc. Int. Conf. Pattern Recognition*, 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 *Proc. 2011 IEEE Int. Conf. Robotics and Automation*, pp. 1817–1824.
- [82] T. Malisiewicz and A. Efros, "Recognition by association via learning per-exemplar distances," presented at IEEE Conf. Computer Vision and Pattern Recognition 2008, Anchorage, AK.
- [83] L. Bo, X. Ren, and D. Fox, "Unsupervised feature learning for RGB-D based object recognition," in *Proc. Int. Symp. Experimental Robotics*, 2012, pp. 387–402.
- [84] M. Blum, J. Springenberg, J. Wulfin, and M. Riedmiller, "A learned feature descriptor for object recognition in rgb-d data," in *Proc. 2012 IEEE Int. Conf. Robotics and Automation*, pp. 1298–1303.
- [85] L. Bo, X. Ren, and D. Fox, "Depth kernel descriptors for object recognition," in *Proc. 2011 IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS)*, pp. 821–826.
- [86] K. Lai, L. Bo, X. Ren, and D. Fox, "Sparse distance learning for object recognition combining rgb and depth information," in *Proc. 2011 IEEE Int. Conf. Robotics and Automation*, pp. 4007–4013.
- [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 Proc.*, vol. 6, no. 3, pp. 171–186, 2011.
- [88] K. Amano, N. Goda, S. Nishida, Y. Ejima, T. Takeda, and Y. Ohtani, "Estimation of the timing of human visual perception from magnetoencephalography," *J. Neurosci.*, vol. 26, no. 15, 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," *Int. J. Appl. Basic Med. Res.*, vol. 5, pp. 124–127, 2015.
- [90] Human Benchmark Project. (2016). [Online]. Available: <http://www.humanbenchmark.com>
- [91] J. Nielsen, *Usability Engineering*. New York: Academic, 1993.
- [92] J. Nielsen. (2010, June). Website response times. Nielsen Norman Group. Fremont, CA. [Online]. Available: <https://www.nngroup.com/articles/website-response-times>
- [93] D. Norman, *The Design of Everyday Things*. New York: Basic Books, 2013.

Ester Martinez-Martin, Robotic Intelligence Laboratory, Universitat Jaume I, Castellón de la Plana, Spain. E-mail: emartine@uji.es.

Angel P. del Pobil, Robotic Intelligence Laboratory, Universitat Jaume I, Castellón de la Plana, Spain, and Department of Interaction Science, Sungkyunkwan University, Seoul, South Korea. E-mail: pobil@uji.es. 