International Journal of Advanced Robotic Systems

OPEN ACCESS ARTICLE

A Taxonomy of Vision Systems for Ground Mobile Robots

Invited Feature Article

Jesus Martínez-Gómez[1], Antonio Fernández-Caballero[1,*], Ismael García-Varea[1], Luis Rodríguez[1] and Cristina Romero-González[1]

1 Universidad de Castilla-La Mancha, Departamento de Sistemas Informáticos, Albacete, Spain
* Corresponding author E-mail: Antonio.Fdez@uclm.es

Received 22 May 2014; Accepted 15 Jul 2014

DOI: 10.5772/58900

© 2014 The Author(s). Licensee InTech. This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract This paper introduces a taxonomy of vision systems for ground mobile robots. In the last five years, a significant number of relevant papers have contributed to this subject. Firstly, a thorough review of the papers is proposed to discuss and classify both past and the most current approaches in the field. As a result, a global picture of the state of the art of the last five years is obtained. Moreover, the study of the articles is used to put forward a comprehensive taxonomy based on the most up-to-date research in ground mobile robotics. In this sense, the paper aims at being especially helpful to both budding and experienced researchers in the areas of vision systems and mobile ground robots. The taxonomy described is devised from a novel perspective, namely in order to respond to the main questions posed when designing robotic vision systems: why?, what for?, what with?, how?, and where? The answers are derived from the most relevant techniques described in the recent literature, leading in a natural way to a series of classifications that are discussed and contextualized. The article offers a global picture of the state of the art in the area and discovers some promising research lines.

Keywords Ground Mobile Robots, Vision Systems, Taxonomy, Review

1. Introduction

A mobile robot is an automatic machine that is capable of movement in any given environment. Unlike industrial robots, which usually consist of a jointed arm (multi-linked manipulator) and a gripper assembly (or end-effector) that is attached to a fixed surface, mobile robots are able to move around in their environment. Therefore, they are not fixed to one physical location. Specifically, a ground mobile robot (GMR) is a robotic platform that operates while being in contact with the ground and which does not rely upon on-board human presence. GMRs are used in many applications where the presence of a human operator may be inconvenient, dangerous or even impossible. Generally, the robot incorporates a set of sensors to perceive the environment and either makes decisions autonomously or pass the information on to a remote human operator who controls the robot via teleoperation. In both cases (autonomous and teleoperated GMRs), the more information that is provided, the better the decisions that are made.

While a teleoperated GMR relies on humans for decision-making, autonomous robots need to incorporate artificial intelligence (AI) capabilities to perform this process. In this sense, AI has been roughly divided into two schools of thought since its beginnings: symbolic and sub-symbolic. These two approaches have also had a strong influence on the robotics field [1]. For robotic

systems to navigate through an environment, autonomous planning and deliberation offer a number of examples. In these kinds of tasks, an accurate environmental representation is needed. The representation can be adequately obtained using a computer or machine vision system, which provides the robot with the relevant information about the environment and its current state. Visual perception plays a fundamental role in the behaviour of human beings. Unfortunately, robots still do not 'see' as humans do. To date, no robot has been able to replicate any of the fundamental human abilities. For example, jointly coordinating 'eyes' and 'hands', which provides flexibility, dexterity and strength in movement, is not yet possible in robotics at present. Moreover, humans usually rely upon their sense of sight to locate, identify (both static and moving) and follow objects (or even track extremity movements). Vision is also crucial in grabbing and manipulating objects, allowing these tasks to be performed quickly and reliably. As a consequence, these capabilities are especially helpful when developing robotic systems able to successfully address the types of tasks mentioned above.

In general, vision in robotics primarily refers to the ability of a robot to visually perceive the environment. Compared to the classical definition of computer visions, robotic vision has to go further in order to accomplish tasks entrusted to robotic platforms. These tasks typically involve: navigating to a specific location while avoiding obstacles; finding agents (either humans or other robots) while interacting with them; locating, classifying and manipulating objects in the scene, and so on. Thus, the goal of robot vision is to exploit the power of visual perception to adequately perceive the environment aimed at while being able to properly react to it. In contrast to computer vision, where sensing is an isolated task and most efforts focus on the scene comprehension and object recognition, robot vision involves dealing with all the internal components/modules available in the platform. In other words, in robot vision, sensing is driven by global tasks where all the system modules play their part [2]. This allows the robot to perceive the environment in order to interact with it appropriately.

Vision has been used in robotics applications for more than 30 years. Some examples include applications in industrial settings, services, medicine and underwater robotics, to name a few. In this paper, the proposals for robot vision from the last five years for GMRs are reviewed. Moreover, a taxonomy of vision systems for GMRs is proposed in studying the most recent journal articles. In this sense, the following main questions addressed in this paper have led to the proposed taxonomy (see Fig. 1): (a) 'why' is a vision system incorporated into a GMR?, (b) 'what' physical components are needed in such a vision system?, (c) 'for what purpose' are vision systems used in GMRs?, (d) 'how' is a vision system for GMRs to be developed?, and (e) 'where' should a vision system for GMRs be exploited? All these questions are answered by discussing some of the most influential examples from the last few years.

The rest of the article is organized as follows. Section 2 provides an overview of the reasons behind the use of vision systems for GMRs. In Section 3, the resources available for building vision systems are presented and classified. A review of the different applications of these systems is provided in Section 4. Next, the internal parameters on their application to GMRs are described in Section 5. Section 6 discusses the different environments where GMRs are used to work, as well as their influence in the development of vision systems. Finally, some conclusions are drawn in Section 7.

2. Why incorporate a vision system into a GMR?

Autonomous robots need to have a set of capabilities that allows them to move and interact with the environment. Among all the skills needed, perception constitutes one of the cornerstones. The word 'perception' refers to, among other things, sensory awareness. Historically, the human sensory system has been used as inspiration to build autonomous vehicles or mobile robots. From the five different senses that humans have, vision is arguably the most important for safely moving and interacting with the world. Fundamental things such as sense of direction, obstacle avoidance and object recognition mainly rely upon the use of the visual system.

In the case of a GMR, it should be able to perceive its own state as well as the state of the environment where it moves around. In this regard, there are different tasks to be performed in which vision plays a crucial role [2]. Self-localization is a good example of this. GPS-based systems are not accurate enough to provide a global solution to the localization problem in most cases. Vision is usually irreplaceable, since a visual recognition of the place where the robot is turns out to be, in many cases, the only suitable solution to this problem.

We can also cite the navigation problem. Although tools like infrared sensors or lasers are employed here to some extent, there are some limitations that restrict their use to low-range obstacle detection only. Cameras provide a global picture of the environment where the destination and near- or mid-range obstacles can be identified. The previous discussion is also applicable to mapping. Building a map entails identifying points of interest within the environment. Such points are usually static objects and rigid structures that are appropriately (and often solely) described by means of their visual features. Actually, self-localization and mapping are usually addressed as a single problem, under the approach known as 'SLAM'. Despite the fact that non-visual sensors are used here, these are typically employed to complement the information given by the vision system. In fact, the SLAM problem is usually solved by relying only upon visual sensors and odometry, and much effort has been made in this direction (e.g., [3], [4], [5]).

Apart from the problems described above, autonomous robots usually have to perform some kind of interaction with different types of objects. Computer vision is by far the most suitable way to address the problem of identifying the objects to interact with while recognizing those to be avoided by the robot (dangerous objects or sensitive/fragile items). Usually, these two categories

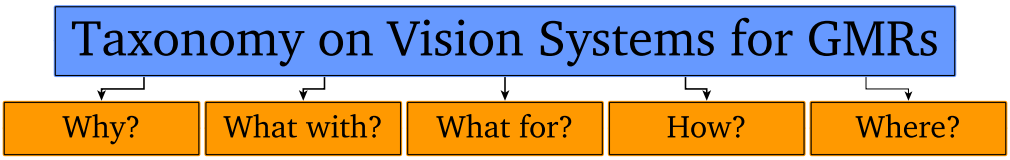

Figure 1. A taxonomy of vision systems for GMRs

of objects are characterized by features like size, shape and colour, which makes visual recognition the most appropriate way to perceive them. Vision is not only useful in this scenario to identifying objects but also to approaching them effectively, in order to grab them or else to carry out any other kind of operation. Grabbing objects requires a detailed visual inspection of the object along with its close environment in order to determine the best way to go near to the object and manipulate it. Besides, identifying objects is not only useful to interacting with them but also to solving other problems, such as place-classification (e.g., [6], [7], [8]).

In social robotics [9], vision also plays an important role. A social robot has to be able to detect humans [10] and, in many cases, to identify them. Face recognition [11] is by far the best method to recognizing a person and, hence, it is widely used in tasks related to care, rehabilitation, surveillance and personal assistance [12]. In addition, gestures or postures [13] are sometimes required to be recognized.

Decision-making and planning also require an appropriate interpretation of the scene around the robot [14]. Deciding the best way to perform a specific operation (for instance, delivering an object) depends heavily upon having an accurate representation of the proximate environment. On the other hand, in some situations the robot could be forced to change its operational mode if there are humans around so as to prevent accidents or other kinds of situations that could cause harm. Although there can be different data sources in appropriately interpreting the environment, a visual description is typically the most useful way to extract relevant information from the outside world. Therefore, if the robot is required to behave autonomously and to make decisions at a high-level of abstraction, the use of a precise vision system seems to be essential.

3. Vision sensors in GMR: what with?

As mentioned above, perception is one of the most important tasks of an autonomous mobile system, which serves as an interface between the robot and the environment. Perception basically works by taking measurements using different sensors and then extracting meaningful information from those measurements. In general, the data extracted by the robot's sensor are used to predict or to model the internal state of the robot, as well as the state of the environment in which the robot is to operate. In particular, this information is used to perform more complex tasks, such as localization, navigation, mapping and human-robot interaction.

Despite the fact that this work is centred on robot vision, most of the perception systems implemented today in real GMRs make use of different sources of information by fusing the measurements provided by different types of sensors. Therefore, we can proceed with a short review of the most widely-used sensors in the majority of GMRs.

- **Tactile sensors**: one of the simplest and most important sensors as regards safety is the tactile sensor. A tactile sensor (also called a 'bumper') is typically used to detect proximity to or contact with different objects - even people - in order to prevent any damage to them or else to the robot itself.
- **Wheel encoders**: an encoder is an electromechanical device that converts the angular position of a shaft into a digital signal. It is used to count the number of turns of the wheels of GMRs, and then to estimate the motion of the robot or else to measure the position of a joint of a robotic manipulator (e.g., arms or legs).
- **Global positioning systems (or GPS)**: these are today mainly used for autonomous outdoor navigation, relying upon the information provided by at least three satellites.
- **Heading sensors**: these are sensors that determine the robot orientation and inclination with respect to a given reference. Some examples are gyroscopes or compasses. Together with appropriate velocity information, they allow the integration of the movement to a position estimate. This procedure is called 'deduced reckoning' and it used in navigation tasks.
- **Accelerometers**: an accelerometer is a device used to measure all the external forces acting upon it, including gravity. Conceptually, an accelerometer is a spring mass damper system, in which the three-dimensional positions of the proof mass relative to the accelerometer casing are measured with some mechanism. When an external force is applied, the proof mass deflects from its natural position; depending upon the physical principle used to measure this deflection, there are different types of accelerometers, such as capacitive or piezoelectric accelerometers.
- **Inertial measurement unit (or IMU)**: an IMU uses gyroscopes and accelerometers to estimate the relative position, velocity and acceleration of a GMR. To estimate the velocity, the initial speed of the vehicle must be known.
- **Ranging finder sensors**: this type of sensor includes the most popular sensors used today in GMRs. Among these, we have sonars (or ultrasonic sensors) to detect and avoid close objects, and laser range-finders (with a higher sensing range), which are used today for obstacle avoidance, scene interpretation and mapping.
- **Digital cameras**: which are also used to complement lasers (or other sensing modalities) with intensity,

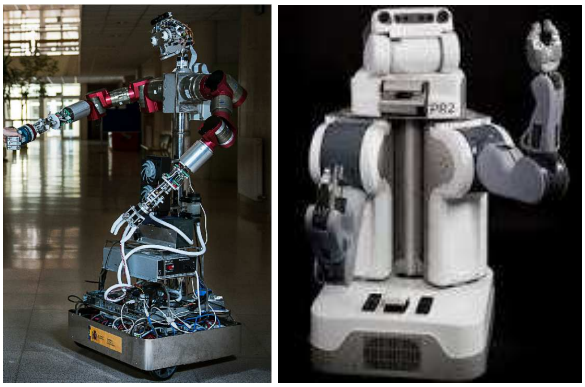

Figure 2. Two examples of social robots: the PR2 robot developed at Willow Garage (right), and the Loki social robot developed at the universities of Castilla-La Mancha and Extremadura (left)

texture and colour information. Cameras are the most important vision sensors used in autonomous mobile systems today.

In Figure 2 and Figure 3, two examples of GMRs are shown: two humanoid prototypes and two autonomous cars, all of them including most of these types of sensors.

Most of the research works published today in the field of robot vision present fusion techniques from camera sensors and other types of sensor to improve their performance and functionality (see, for example, [15], [16]).

We can classify sensors into two main groups according to their functional properties: proprioceptive or exteroceptive, and passive or active. Proprioceptive sensors measure a value internal to the system which is related to the internal state to the robot. Some examples of proprioceptive sensors are motor speed, wheel load, robot arm joint angles and battery voltage. Exteroceptive sensors acquire information from the robot's environmental state. Some examples are distance measurements, light intensity and sound amplitude. We can say that exteroceptive sensor measurements are interpreted by the robot in order to extract meaningful information from the environment. Examples of exteroceptive sensors are tactile sensors, compasses, GPSs, lasers and camera sensors.

Passive sensors measure ambient environmental energy entering into the sensor. Examples of passive sensors include temperature probes, microphones and cameras. On the other hand, active sensors emit energy into the environment. Hence, they measure the environmental reaction to its actions. Since active sensors can manage more controlled interaction with the environment, they often achieve superior performance over passive ones. However, active sensing introduces several risks. The outbound energy may affect the very characteristics that the sensor is attempting to measure. For example, signals emitted by other, nearby, robots, or similar sensors on the same robot, may influence the resulting measurements. Examples of active sensors include wheel encoders, ultrasonic sensors and laser range-finders.

3.1. Vision sensor taxonomy

The use of vision sensors can be classified according to different criteria ([17], [18]), but we here focus only on the most widely used vision sensors, i.e, cameras. A camera is an optical instrument that records images that are stored and/or transmitted to another location. The term 'camera' comes from the phrase *camera obscura* (a Latinism for 'dark chamber'), an early mechanism for projecting images. Modern cameras have evolved from the *camera obscura* device, and its functionality is now very similar to the functionality of the human eye, i.e., taking photographic images or moving images, such as videos and movies. The term 'camera' is also used for devices producing images (or image sequences) from measurements of the physical world, even when the image formation cannot be described as 'photographic'.

According to the type of information that GMRs cameras capture from the environment, we can roughly classify them according to three main categories: colour, thermal and range cameras (see Figure 4).

Colour cameras produce images that are comprehensible to humans, and their most important internal parameter is the colour space. RGB (red, green and blue) might be considered the most standard parameter, but it is not appropriate for dealing with lighting changes. This is because it uses three chromatic components where the luminance is present. Therefore, lighting changes would result in variations for the three R, B and G components. In order to cope with such challenging scenarios, some colour spaces where the luminance is concentrated in just a single component have been proposed, such as YUV or YCbCr. YCbCr and YUV are colour spaces used interchangeably. Y is the luminance component and Cb and Cr are the blue-difference and red-difference chroma components.

While most colour cameras present a directional field of view, omnidirectional cameras are widely used in robotics. An omnidirectional camera is characterized by a very large field of view - ultimately, a spherical field of view. This type of camera can be classified according to three different categories: dioptric cameras, which are formed by a system of lenses to achieve a very large field of view, typically a hemispherical field of view; catadioptric cameras, which use a combination of lenses and mirrors; and polydioptric cameras, which consist of a system of multiple, overlapping cameras, such as those used today in *Google Street View* cars.

Thermal cameras are the second type of vision cameras considered in this 'what with?' taxonomy. Thermal cameras capture the infrared radiation emitted by any object with a temperature greater than zero. Although they are not very common for general purpose GMRs, their use in specific applications becomes essential [19]. These applications include surveillance [20], fire control [21] and a wide range of medical analysis applications [22, 23].

Finally, the third family of cameras comprises range cameras, which produce images of the distance to each point in the scene. They are one of the most important sensing modalities in the field of GMRs, up to the point

Figure 3. Two examples of autonomous self-driving car prototypes: the Smarter Car developed at ETHL Zurich (left), and the Junior Car developed at Stanford University (right)

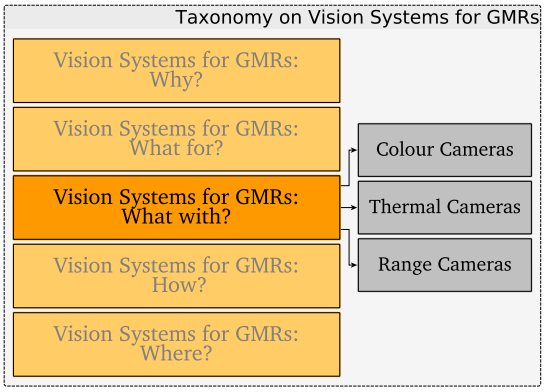

Figure 4. A 'what with' taxonomy

whereby *range imaging* has become one of the most important fields of research in recent years, not only in the specific case of robotic vision but also in computer vision. Taking this into account, we describe the main range imaging techniques in the following.

3.2. Range imaging techniques

'Range imaging' is the name for a collection of techniques used to produce a 2D images representing the distance to points in a scene from a specific point, normally associated with some types of sensor device. This is broadly referred to as a 'range camera'. The resulting image, the range image, stores pixel values corresponding to these distances.

Range cameras operate according to a number of different techniques. Below, a taxonomy of range cameras according to these criteria is presented (see [24], [25], [26], [27], [28]):

- Stereo vision. A stereo camera system can be used to determine the depth of points in the scene, for example, from the centre point of the line between their focal points. In order to solve the depth measurement problem using a stereo camera system, it is necessary to find corresponding points in the two images. Solving the correspondence problem is one of the main problems when using this type of technique, up to the point whereby range imaging based on stereo triangulation usually produces reliable depth estimates only for a subset of all the points visible for the cameras.
- Structured light. Structured light is the projection of a light pattern (a plane, grid or more complex shape, typically named 'structured light') at a known angle onto an object. This technique is very useful for imaging and acquiring dimensional information. The most commonly used light pattern is generated by fanning out a light beam into a sheet-of-light. When the sheet-of-light intersects with an object, a bright line of light can be seen on the object surface. By viewing this line of light from an angle, the observed distortions in the line are translated into height variations. Scanning the object with the light, by moving either the light source (and normally also the camera) or the scene in front of the camera, a sequence of depth profiles of the scene is generated. As a result, 3D information about the shape of the object can be obtained.
- Time-of-flight (ToF). Depth can also be measured using the standard ToF technique, similar to radar or lidar. In this technique, a light pulse is used instead of a radio frequency pulse. For instance, a scanning laser (as a rotating laser head) is employed to obtain a depth profile for points which lie in the scanning plane. This approach also produces a type of range image similar to a radar image. ToF cameras are devices that capture a whole scene in three dimensions with a dedicated image sensor, and therefore moving parts are not needed.
- Structure from motion. 'Structure from motion' refers to the process of estimating three-dimensional structures from two-dimensional image sequences which may be coupled with local motion signals. This technique presents the same problem as a structure from the stereo technique: the correspondence between images and the reconstruction of a 3D object needs to be found. To find correspondences between images, features such as corner points (edges with gradients in multiple directions) need to be tracked from one image to the next. The feature trajectories over time are then used to reconstruct their 3D positions and the camera motion.

Some specific types of cameras include ToF cameras, like the SwissRanger or the Kinect version 2, and structured

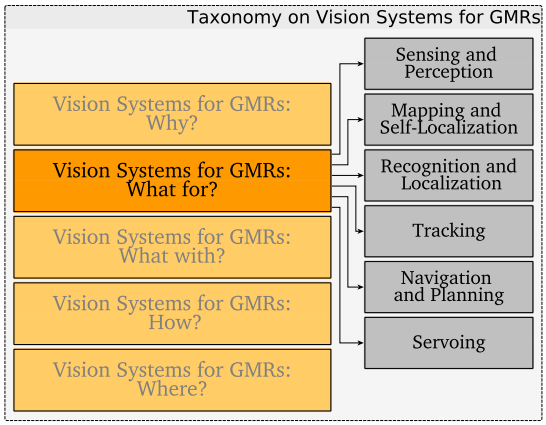

Figure 5. A 'what for' vision system taxonomy

light cameras, like the Kinect version 1. Most range sensors, such as sonars, laser range finders, and the SwissRanger or the Kinect 2, work by measuring the ToF that it takes for an emitted signal to return back to the source emission sensor. In general, the travel distance of a sound or electromagnetic wave is given by the product of the speed and the ToF.

Nowadays, 3D scanners have become very popular, and their use is irreplaceable in autonomous self-driving cars or for tasks such as 3D SLAM. One of the most popular current 3D scanners is the Velodyne sensor. This sensor is a 3D lidar that employs 64 laser emitters instead of the single sensor used in common laser range finders. This device spins at rates of 5 to 15 hertz and delivers more than 1.3 million data points per second.

Structured light sensors rely upon an emitter to project a known pattern or structured light onto the environment. They either project light textures or emit collimated light by means of a rotating mirror. Yet another popular alternative is to project a laser stripe by turning a laser beam into a plane, thanks to the use of a prism. However, new possibilities for applications in robotics have recently been opened up by Kinect, a sensor released in 2010 as a part of the Microsoft Xbox 360 video games console, and produced by *Prime Sense*. Kinect is a very cheap range camera that relies upon the structured light principle explained above. An infrared laser emitter is used to make the projected pattern invisible to the human eye.

4. What should a vision system in GMRs be used for?

Robot vision-based mobility has been the source of countless research contributions within the domains of both vision and control. Vision is becoming increasingly common in applications such as localization, automatic map construction, autonomous navigation, path following, inspection, monitoring and risky situation detection [29]. Figure 5 shows the taxonomy described in this section. The 'what for' vision system taxonomy proposes a decomposition into six classes: sensing and perception; mapping and self-localization; recognition and localization; tracking; navigation and planning; and servoing.

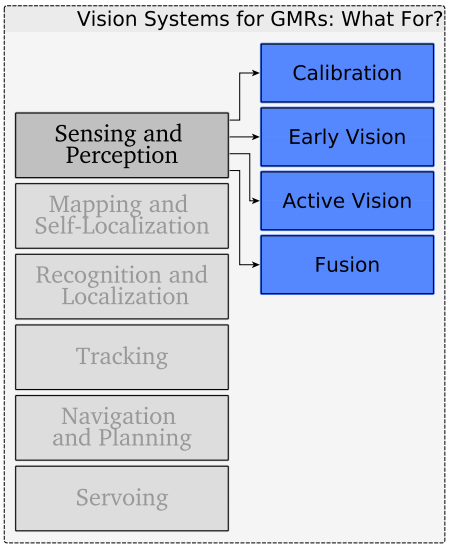

Figure 6. A visual sensing and perception taxonomy

4.1. Visual sensing and perception

There are four specific classes that have been identified as part of visual sensing and perception: calibration, early vision, active vision and fusion (see Figure 6). These are discussed next.

4.1.1. Calibration

Camera calibration is one of the most important components of computer vision. Indeed, the intrinsic and extrinsic parameters of cameras are obtained from camera calibration. Intrinsic parameters encompass focal length, image format and principal point. Extrinsic parameters denote the coordinate system transformations from 3D world coordinates into 3D camera coordinates. Equivalently, the extrinsic parameters define the position of the camera's centre and its heading in world coordinates. Camera calibration is a necessary step in 3D computer vision in order to extract metric information from 2D images. It has been extensively studied in computer vision and photogrammetry.

Some relevant examples of the need for accurate camera calibration are described next. For instance, in [30] a classification of visual servos is presented and used to derive their possible structures. An important cue in this approach is the analysis of the influence of robot-model and robot-camera calibration on the derived control structures. Furthermore, the use of ToF cameras in mobile robotics is suitable for real-time 3D tasks, such as tracking, visual servoing or object pose estimation. Obviously, their usability mainly depends upon accurate camera calibration. A calibration process for ToF cameras with respect to intrinsic parameters, depth measurement distortion and the pose of the camera relative to a robot's end-effector has been described in [31]. Lastly, hand-eye calibration has emerged as a hot topic. In this sense, a very recent paper considers conventional techniques for vision robot system calibration where the camera and robot hand-eye parameters are computed

separately, i.e., first performing camera calibration and then carrying out hand-eye calibration based upon the calibrated parameters of the cameras [32]. A joint algorithm is proposed, combining camera and hand-eye calibration. The proposed algorithm gives the solutions of the camera and hand-eye parameters simultaneously by using nonlinear optimization.

According to the dimensions of the calibration objects, calibration techniques are roughly classified into three categories [33]. (a) In 3D reference object-based calibration, camera calibration is performed by observing a calibration object whose geometry in 3D space is known with very high precision. (b) Techniques in 2D plane-based calibration require the observation of a planar pattern shown at a few different orientations. Such a method for camera calibration is described in an approach which describes how to obtain the position of a chequerboard corner at sub-pixel accuracy from digital images [34].(c) In 1D line-based calibration, the calibration objects used are composed of a set of collinear points. (d) Lastly, self-calibration techniques do not use any calibration object and can be considered as 0D approach because only image point correspondences are required. A prototypical paper introduces the widespread application of camera calibration in robot navigation, three-dimensional reconstruction, bio-medicine, virtual reality and visual surveillance [35]. The paper summarizes the methods in different applications, such as traditional calibration, self-calibration and active vision calibration.

Let us also highlight the importance of providing data sets for calibration purposes. An example is the data set collected by the MIT autonomous vehicle Talos during the 2007 DARPA Urban Challenge [36]. Data from a high-precision navigation system, five cameras, 12 SICK planar laser range scanners and a Velodyne high-density laser range scanner were synchronized and logged to disk for 90 km of travel. In addition to documenting a number of large loop closures useful for developing mapping and localization algorithms, this data set also records the first robotic traffic jam and two autonomous vehicle collisions. In a more recent paper, large, accurately calibrated and time-synchronized data sets, gathered outdoors in controlled and variable environmental conditions, using an GMR and equipped with a wide variety of sensors, are presented [37]. These include four 2D laser scanners, a radar scanner, a colour camera and an infrared camera.

4.1.2. Early vision

The first processing stage in computational vision, also called 'early vision', consists of decoding two-dimensional images in terms of the properties of 3D surfaces. Perceiving the environment is crucial in any application related to mobile robotics research. Early vision includes problems such as the recovery of motion and optical flow, shape from shading, surface interpolation and edge detection. The results of this processing stage are used for higher-level tasks such as navigation in the environment, the manipulation of objects and, of course, object recognition, as well as reasoning about objects. Conventionally, vision is said to be early when it implies little or no semantic interpretation of the scene. Therefore, early vision excludes higher cognitive aspects like object recognition and event interpretation [38]. Unlike high-level vision, early vision is mostly considered as a set of bottom-up processes that do not rely upon specific high-level information about the scene to be analysed.

Early vision faces new challenges due to new imaging techniques and algorithms. For instance, log-polar imaging consists of a type of method that represents visual information with a space-variant resolution inspired by the visual system of mammals. It has been studied for about three decades, and has surpassed conventional approaches in robotics applications [39]. Moreover, real-time human detection through processing video-captured by a thermal infrared camera mounted on an autonomous mobile platform has been introduced [40]. A big challenge arises with multi-sensor systems. Multi-sensor systems consist of several types of sensors, which are installed on fixed or mobile devices. These components provide a huge quantity of information that has to be contrasted, correlated and integrated in order to recognize and react on special situations [41].

4.1.3. Active vision

An area of computer vision is active vision, sometimes also called 'active computer vision'. An active vision system is one that manipulates the viewpoint of the camera in order to investigate the environment and acquire better information from it. Examples of active vision systems usually involve a robot-mounted camera, and applications include automatic surveillance, SLAM, route planning, and so on. Active vision is based on the controlled movement of the viewpoint of the imaging camera as an integral part of the image-processing task. Previously in computer vision research, fixed camera geometry and static images have been beneficial in constraining and simplifying image-processing tasks in order to reduce the enormous complexity of visual data. Active vision takes a different approach and, by analogy with animal vision, does not avoid movement but rather gains information from the dynamics of changing viewpoints to resolve ambiguities, gain depth information and establish relationships between visual sensing and action [42]. Active vision has the goal of improving visual perception; therefore, the investigation of ocular motion strategies must play an important role in the design of robot eyes [43].

Due to its inherent interest, developments in active vision in robotic applications over the last 15 years have been surveyed in [44]. A major challenge to the widespread deployment of mobile robots is the ability to function autonomously, learning useful models of environmental features, recognizing environmental changes and adapting the learned models in response to such changes. The main contribution of [45] is a survey of vision algorithms that are potentially applicable to colour-based mobile robot vision. A first example of a general system for autonomous localization using active vision was described over 10 years ago [3]. It is enabled by a high-performance

stereo head, addressing such issues as uncertainty-based measurement selection, automatic map-maintenance and goal-directed steering.

Visual attention is a complex phenomenon. A particular example of a practical robotic vision system that employs certain attentive processes is presented in [46]. In addition, the problem of actively searching for an object in a 3D environment is studied under the constraint of a maximum search time using a visually guided humanoid robot with 26 degrees of freedom [47]. Another study follows the standard pattern recognition approach based on four main steps [48]: (i) preprocessing to achieve colour constancy and stereo pair calibration; (ii) segmentation using depth-continuity information; (iii) feature extraction based on visual saliency; and (iv) classification using a neural network. The main novelty of the approach lies in the feature extraction step, where the authors propose novel features derived from a visual saliency mechanism. Moreover, in [49] the results of an investigation and pilot study into an active binocular vision system that combines binocular vergence, object recognition and attention control in a unified framework are presented. The prototype developed is capable of identifying, targeting, verging on and recognizing objects in a cluttered scene without the need for calibration or other knowledge of the camera geometry.

Lastly, Kalman filters have received much attention with the increasing demand for robotic automation. The recent developments in robot vision are briefly surveyed in [50]. Among the many factors that affect the performance of a robotic system, Kalman filters have made great contributions to vision perception. Kalman filters solve uncertainties in robot localization, navigation, following, tracking, motion control, estimation and prediction, visual servoing and manipulation, and structure reconstruction from a sequence of images.

4.1.4. Multi-sensor data fusion

Cameras are one of the most relevant sensors in autonomous robots. One challenge with them is to manage the small field of view of regular cameras. A method of coping with this, similar to the attention systems in humans, is to use mobile cameras to cover all the robot surroundings and to perceive all the objects of interest to the robot tasks, even if they do not lie in the same snapshot [51]. Data fusion is the process of the integration of multiple data and knowledge representing the same real-world object in a consistent, accurate and useful representation. Data fusion processes are often categorized as 'low', 'intermediate' or 'high', depending upon the processing stage at which fusion takes place. Low-level data fusion combines several sources of raw data to produce new raw data. The expectation is that fused data is more informative and synthetic than the original inputs. For example, sensor fusion is also known as 'multi-sensor' data fusion. Multi-sensor data fusion is the process of combining observations from a number of different sensors to provide a robust and complete description of an environment or process of interest. Data fusion finds wide application in many areas of robotics, such as object recognition, environment mapping and localization [52].

There exists the possibility of fusing data from cameras alone. In this sense, the possibilities of using monocular SLAM algorithms in systems with more than one camera are explored [53]. The idea is to combine, within a single system, the advantages of both monocular vision (bearings-only, infinite range observations but no 3D instantaneous information) and stereo vision (3D information up to a limited range). A recent paper [54] proposes an approach to combine data from multiple low-cost sensors to detect people in a mobile robot. The work is based on the fusion of Kinect and a thermal sensor mounted on top of a mobile platform.

Due to their wide field of view, omnidirectional cameras are becoming ubiquitous in many mobile robotic applications. A challenging problem consists of using these sensors, mounted on mobile robotic platforms, as visual compasses to provide an estimate of the rotational motion of the camera/robot from the omnidirectional video stream. In this sense, [55] presents a multiple-view geometry constraint for paracatadioptric views of lines in 3D, that are used to design a visual compass algorithm that does not require either the knowledge of the camera calibration parameters or the 3D scene geometry.

Camera information can be fused with other sensor data. The work proposed in [56] suggests how to improve the accuracy of a mobile robot's localization by using the sensor network information, which fuses the machine vision camera, an encoder and an IMU sensor. In another work [57], a context-based multi-sensor system applied to pedestrian detection in urban environments is presented. The proposed system comprises three main processing modules: (i) a lidar-based module acting as the primary object detector, (ii) a module which supplies the system with contextual information obtained from a semantic map of the roads, and (iii) an image-based detection module, using sliding window detectors, with the role of validating the presence of pedestrians in the regions of interest generated by the lidar module. A Bayesian strategy is used to combine information from sensors on-board the vehicle with information contained in a digital map of the roads.

4.2. Visual mapping and self-localization

In this subsection, we distinguish between mapping and self-localization when both are applied independently as well as when they are applied simultaneously (see Figure 7). Here, the term 'self-localization' is used instead of 'localization' in order to differentiate it from the localization of a given object of the robot's environment.

4.2.1. Self-localization

Localization (self-localization) is the process of determining the robot's location within its environment. More precisely, it is a procedure which takes as input a geometric map, a current estimate of the robot's pose and sensor readings, and produces as output an improved

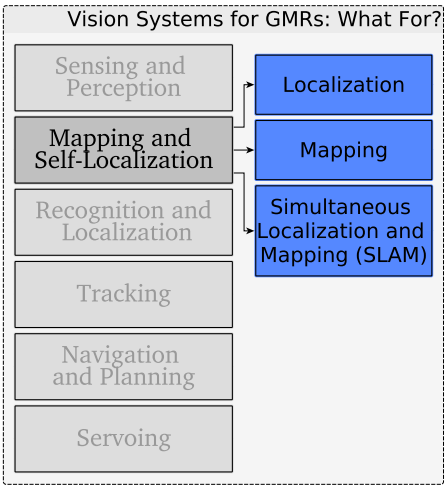

Figure 7. A visual mapping and self-localization taxonomy

estimate of the robot's current pose (position and orientation) [58]. Indeed, the analysis and classification of places, and the ability to actively collect information, are necessary to realizing the autonomous navigation of intelligent robots in a variety of settings. For instance, visual data are organized into an orientation histogram to roughly express input images by extracting and cumulating straight lines according to a direction angle [59]. In addition, behavioural data are organized into a histogram by cumulating motions performed to avoid obstacles encountered while the robot is executing specified behavioural patterns. The location of the robot is classified by merging the probabilities for visual and behavioural data.

Identifying the location of unmanned vehicles is a very important task for automatic navigation, as described in [60]. Conventional positioning sensors may fail to work properly in some real-world situations due to internal and external interference. Given a digital surface map, the location of the vehicle is estimated by the registration of the map and multi-view range images obtained from the vehicle. In [61], environmental information acquired from two sensors is combined and fused by a Bayesian sensor fusion technique based on the probabilistic reliability function of each sensor predefined through experiments for the self-localization of a mobile robot with a monocular camera and a laser-structured light sensor. In [62], the authors describe a system for mobile robot localization in an indoor environment, using concepts like homography and matching borrowed from the context of stereo- and content-based image retrieval techniques. A group of points of interest (POIs) is extracted to represent the image for robust matching in order to deal with variations with respect to viewpoint and camera positions.

The work presented in [5] is related to the application of a visual odometry approach to estimate the location of mobile robots operating under off-road conditions. The visual odometry approach is based on template matching, which deals with the estimation of the robot's displacement through a matching process between two consecutive images. Standard visual odometry has been improved using the visual compass method for orientation estimation. For this purpose, two consumer-grade monocular cameras have been employed. One camera is pointed at the ground underneath the robot, while the other is looking at the surrounding environment. In [63], a more natural approach is presented which dynamically determines a subset of images that best describes the complete image data in the space of all previously seen images. The actual problem of finding such a subset is called the 'connected dominating set', which has been well-studied in the field of graph theory. Lastly, in [64], features extracted from omnidirectional panoramic images are used in a method for the localization of a mobile robot equipped with an omnidirectional camera. Nodes around the robot are extracted by the correlation coefficients of a circular horizontal line between the landmark and the current captured image.

4.2.2. Mapping

Robotic mapping addresses the problem of acquiring spatial models of physical environments through mobile robots [65]. Obviously, mapping is performed in both 2D and 3D, depending upon the camera technologies used.

As regards 2D mapping, there is, for instance, an autonomous navigation system for an indoor mobile robot based on monocular vision [66]. The navigation system is composed of online and offline stages. During the offline learning stage, the robot records an image frame sequence. From this sequence, a hybrid environment map is built with Rao-Blackwell particle filters. The map is partitioned into topological locations characterized by a set of geometrical scale-invariant key-points. During the online navigation stage, the robot recognizes the most likely location through a robust location recognition algorithm, estimates the relative pose between the locations, and then navigates the environment autonomously. In another approach [67], the environment is represented as a collection of modular occupancy grids which are added to the map as far as the mobile robot finds objects outside the existing grids. Under this approach, a ToF camera is exploited as a range sensor for mapping.

A 3D mapping technique that learns high-fidelity models for a geo-specific lidar simulation directly from pose tagged lidar data has been introduced in [68]. The approach introduces a stochastic, volumetric model that captures and reproduces the statistical interactions of lidar with the terrain. The model is automatically learnt from 3D mapping data collected by a GMR in the target environment. On the other hand, RGB-D cameras (such as the Microsoft Kinect) are quite novel sensing systems that capture RGB images along with per-pixel depth information. Of course, such cameras are used for building dense 3D maps of indoor environments [69]. Such maps have applications in robot navigation, manipulation, semantic mapping and telepresence. The authors present RGB-D mapping, a full 3D mapping system that utilizes a novel joint optimization algorithm combining visual features and shape-based alignment. Visual and depth information are also combined for view-based loop-closure detection, followed by pose optimization, to achieve a globally consistent maps.

4.2.3. Simultaneous localization and mapping

An autonomous mobile robot must have the ability to navigate in an unknown environment. The simultaneous localization and mapping (SLAM) problem is related to this autonomous ability. SLAM is a technique used by autonomous robots to build up a map within an unknown environment (i.e., without *a priori* knowledge) or else to update a map within a known environment (i.e., with *a priori* knowledge from a given map), while at the same time keeping track of their current location.

Visual SLAM (VSLAM) is probably still one of the most challenging problems in mobile robotics. Since 2005, there has been intense research into VSLAM, using primarily visual sensors, because of the increasing ubiquity of cameras such as those in mobile devices [4]. Here, we cite some recent work using different visual sensors. In [70] a SLAM method is proposed that uses vertical lines extracted from an omnidirectional camera image and horizontal lines from range sensor data. Due to the large field of view of the omnidirectional camera, features remain in the image for a length of time sufficient to estimate the pose of the robot and the features accurately. A real-time hierarchical (topological/metric) VSLAM system focuses on the localization of a vehicle in large-scale outdoor urban environments in [71]. It is exclusively based on the visual information provided by a cheap wide-angle stereo camera. Additionally, in [72] a successful real-world implementation of an extended Kalman filter-based (EKF-based) SLAM algorithm for indoor environments uses two web-cam based stereo vision-sensing mechanisms. Lastly in [73], a system called 'continuous appearance-based trajectory simultaneous localization and mapping' (CAT-SLAM) is proposed. The system augments sequential appearance-based place recognition with local metric pose filtering to improve the frequency and reliability of appearance-based loop closure.

The 3D information provided by range sensors has often been introduced in SLAM solutions, where we can find scan-matching [74] and probabilistic approaches [75]. These solutions have to cope with a large amount of data to be processed and stored. Therefore, compact map representations are also proposed in conjunction with SLAM solutions [76].

4.3. Visual recognition and localization

As represented in Figure 8, four different fields of application are identified within the recognition and localization taxonomy: feature detection and segmentation, recognition and classification, localization and pose estimation and, finally, inspection.

4.3.1. Feature detection and segmentation

In computer vision, 'feature detection' refers to methods that seek to compute abstractions of image information. These abstractions are then classified as certain types of pre-defined features. Feature detection is a low-level image-processing operation - that is, it is usually performed as the first operation on an image, examining

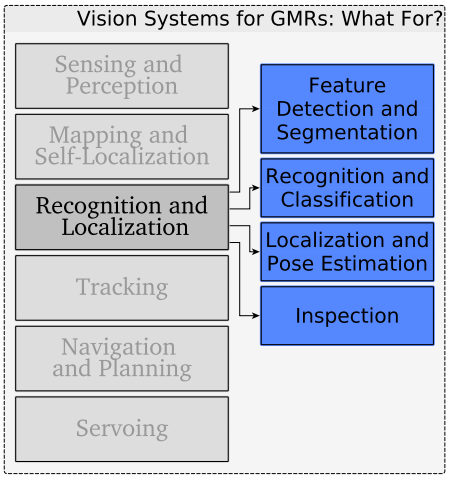

Figure 8. A visual recognition and localization taxonomy

every pixel to see if there is a feature present at that pixel. If this is part of a larger algorithm, then the algorithm will typically only examine the image within the region of interest of the features. The goal of segmentation is to simplify and/or change the representation of an image into something that is more meaningful and easier to analyse. Image segmentation is typically used to locate objects and boundaries (lines, curves, etc.) in images.

In general terms, there are numerous approaches to object segmentation using robot visual sensors. As an example, consider a paper in which a simultaneous 3D volumetric segmentation and reconstruction method, based on the so-called 'generic fitted shapes' method, proposed [77]. The aim of this work is to cope with the lack of volumetric information encountered in visually controlled mobile manipulation systems equipped with stereo or RGB-D cameras. Another work [78] introduces a novel probabilistic method for robot based object segmentation. The method integrates knowledge of the robot's motion to determine the shape and location of objects. This allows a robot without prior knowledge of its environment to isolate objects from their surroundings by moving them and observing the visual feedback. It is also worthwhile noting point cloud feature extraction algorithms which are currently being used for the perception to filter outliers from noisy data, stitch 3D point clouds together, segment relevant parts of a scene, extract key points and compute descriptors to recognize objects in the world based on their geometric appearance, and create surfaces from point clouds (e.g., [79] and [80]).

In recent years, the spaces where robots work have generally been expanding into human spaces (traditional industrial robots aside, which work only at fixed positions apart from humans). When a robot vision system is being employed to monitor humans for surveillance applications, each person in the scene has to be identified. Humans, however, often move together, and occlusions between them occur frequently. A probabilistic neural network is employed to learn the patterns of the best dividing-position along the top pixels of an image region of partly occluded people [81]. Furthermore, an in-depth study has been carried out regarding the possibilities

of a colour camera placed on top of a robot in order to discriminate between humans and thus achieve more reliable person-following behaviour with the robot. In particular, the authors have reviewed and analysed the possibility of using the most popular colour and texture features used in object and texture recognition, to identify and model the target [82]. At present, real-time human detection through processing video captured by a thermal infrared camera is also being considered. A usual approach starts with a phase of static analysis for the detection of human candidates through some classical image processing techniques, such as image normalization and thresholding. Next, the proposal uses Lucas and Kanade optical flow without the pyramids algorithm to filter moving foreground objects from moving scene backgrounds [83].

In addition, face detection and recognition have wide applications in robot vision and intelligent surveillance. However, face identification at a distance is very challenging because long-distance images are often degraded by low resolution, blurring and noise. A person-specific face detection method uses a nonlinear optimum composite filter and subsequent verification stages [84]. The filter's optimum criterion minimizes the sum of the output energy generated by the input noise and the input image. In another paper [11], an efficient facial-feature detection approach based on local image region and direct pixel-intensity distributions is presented. Furthermore, an algorithm based on fuzzy rough-sets is proposed for the recognition of hand postures and faces in [13].

4.3.2. Recognition and classification

Pattern recognition is the assignment of a label to a given input value. Object recognition, in computer vision, comprises the tasks of finding and identifying objects in an image or video sequence. An example of pattern recognition is classification, which attempts to assign each input value to one of a pre-defined set of classes. However, this set of pre-defined classes can suffer variations (e.g., the introduction of new objects into the environment), and novel categories should be discovered, as is the case in the proposal presented in [85].

Let us start with some articles related to recognition. Firstly, there is object recognition for use in mobile robotics. Deformable models have been studied in image analysis over the last decade and have been used for the recognition of flexible or rigid templates under diverse viewing conditions. The work in [86] addresses the question of how to define a deformable model for a real-time colour vision system for mobile robot navigation. Instead of receiving the detailed model definition from the user, the algorithm extracts and learns the information from each object automatically. The resulting perception module has been integrated successfully into a complex navigation system. Additionally, in [87], an image-understanding system and methods targeting automatic, lighting-independent and reliable colour-based object recognition under real-time conditions is presented. Its application test bed is global vision robot soccer, but it has many other applications in the colour-based supervision of moving objects. Another work, [88], presents a systematic approach to the problem of autonomous 3D object searches in indoor environments using a two-wheeled non-holonomic robot fitted with an actuated stereo-camera head and with processing done on a single laptop. A probabilistic grid-based map encodes the likelihood of an object's existence in each cell and is updated after each sensing action. The updating schema incorporates characteristic parameters modelled after the robot's sensing modalities and allows for sequential updating via Bayesian recursion methods.

Another challenging proposal explores the concept of interactive perception - in which sensing guides manipulation - in the context of extracting and classifying unknown objects within a cluttered environment [89]. Under the proposed approach, a pile of objects lies on a flat background and the goal of the robot is to isolate, interact with and classify each object so that its properties are obtained. The algorithm considers each object to be classified using colour, shape and flexibility.

On the other hand, place recognition and environmental reconstruction are hot topics. For instance, a method is proposed for visual place recognition using a bag of words obtained from an accelerated segment to test so-called '(FAST)+BRIEF' features [90]. With this method, a vocabulary tree that discretizes a binary descriptor space is built and is used to speed up correspondences for geometrical verification. Environmental 3D reconstruction is a strategic task in many contexts, above all, in infrastructure inspection and automatic vehicular motion. In [91], a sensor is presented that is capable of recovering 3D data with a very high profile acquisition rate and which performs omnidirectional, highly accurate environmental reconstruction: these skills are allowed by a profilometric laser approach coupled to a catadioptric system.

Lastly, the automatic detection and description of events, particularly human behaviour, is one of the most challenging issues, as event interpretation is highly dependent upon the subject of the robot's attention, which is not uniquely specified. To tackle this problem, the concept of cognitive ontology as a framework for a system that automatically decides upon the attentive focus and which describes the events for a robot has been introduced in [14].

4.3.3. Localization and pose estimation

In computer vision and robotics, a typical task is to identify specific objects in an image and to determine each object's position and orientation relative to some reference coordinate system. This information is then used, for example, to allow a robot to manipulate an object or to avoid moving into the object. The combination of position and orientation is referred to as the 'pose' of an object. The specific task of determining the pose of an object in an image (or stereo images or image sequence) is referred to as 'pose estimation'. The pose estimation problem may be solved in different ways depending upon the image sensor configuration and choice of methodology.

Object localization is probably the simplest and most widespread problem in robot visual localization. Consider, for instance, a landmark detection and localization approach using an integrated laser-camera sensor [92]. Another paper presents a combined machine learning and computer vision approach for robots to localize objects. It allows a humanoid robot to quickly learn to provide accurate 3D position estimates of objects seen [93]. In addition, the approach localizes objects robustly when utilized in a robot's workspace at arbitrary positions, even while the robot is moving its torso, head and eyes.

The problem of estimating the position and orientation (pose) of an object in real-time constitutes an important issue for the vision-based control of robots. Many vision-based pose-estimation schemes in robot control rely upon an EKF that requires the tuning of filter parameters. A new algorithm, namely an iterative adaptive EKF, is proposed by integrating mechanisms for noise adaptation and iterative-measurement linearization [94]. A recent paper describes a binary search pose estimation technique for poses constrained to three degrees of freedom (DOF) [95]. The technique requires three fiduciary marker points and operates by minimizing the angular DOF through a binary search-driven algorithm. The aim of [96] is to improve the skills of robotic systems in their interaction with nearby objects. The basic idea is to enhance the visual estimation of objects in the world through the merging of different visual estimators of the same stimuli. A neuroscience-inspired model of stereoptic and perspective orientation estimators, merged according to different criteria, is implemented on a robotic setup and tested in different conditions. In a similar manner, [97] presents a version of the camera-space manipulation method (CSM). The set of nonlinear view parameters of the classic CSM is replaced with a linear model.

Place and scene recognition comprise the next relevant problem. Although, mobile robotics has achieved notable progress in increasing the complexity of the tasks that mobile robots perform in natural environments, we need to provide them with a greater semantic understanding of their surroundings. As a distinguishing feature, [8] uses common objects, such as doors and furniture, as a key intermediate representation to recognize indoor scenes. The authors frame the method as a generative probabilistic hierarchical model, whereby they use object category classifiers to associate low-level visual features with objects, and contextual relations to associate objects with scenes. Another work [98] presents a technique for place categorization from visual cues called 'place labelling' through image sequence segmentation. It uses change point detection to temporally segment image sequences which are subsequently labelled. Change point detection and labelling are performed inside a systematic probabilistic framework. In addition, [99] proposes a robust real-time camera pose and a scene structure estimation system. First, the pose of the camera is estimated through the analysis of the so-called 'tracks'. The tracks include key features from the imaged scene and geometric constraints which are used to solve the pose estimation problem. Second, based on the calculated pose of the camera, the scene is analysed via a robust depth segmentation and object classification approach.

4.3.4. Inspection

In engineering activities, inspection involves the measurements, tests and gauges applied to certain characteristics with regard to an object or activity. The results are usually compared to specified requirements and standards for determining whether the item or activity is in line with these targets. The use of image processing systems in industrial manufacturing and assembly has significantly increased in recent years. Used as inspection systems, they enhance product quality and minimize loss through waste [100].

Three representative examples of visual inspection are presented. In the first place, a vision-based crack detection system and algorithm to inspect the base side of bridges has been developed [101]. After a human operator makes a decision based on the vision images captured, if the lines on the base side are cracks or dirt, the algorithm automatically finds the length, the width and the shape of the cracks. Another paper [102] describes a mobile inspection robot with an automatic pipe-tracking system for feeder pipe inspection. An automatic pipe-tracking system is proposed based on machine vision techniques to make the mobile robot follow an exact outer circumference of a curved feeder pipe as closely as possible, which is one of the requirements of a thickness measurement system for a feeder pipe. Lastly, a mobile robot equipped with two lasers and a charge-coupled device (CCD) camera for pipe inspection is proposed in [103]. Circular laser streaks that appeared on the inner surface of the pipe reveal the shape of the pipe. The 3D shape of a sewer pipe is reconstructed considering the movement of the mobile robot along the pipe. Since the tilt of the mobile robot with respect to the axis of the pipe appears as the deformation between two circular streaks, the shape of a sewer pipe is measured accurately, regardless of the tilt of the robot.

4.4. Visual navigation and planning

The fourth 'what for?' taxonomy is devoted to applications within the field of navigation and path planning (see Figure 9), and we identify three different families: path planning and exploration, trajectory estimation, and motion planning.

4.4.1. Path planning and exploration

Path planning is the act of finding a path to go from one location to another one. Exploration occurs when a robot is placed in an unknown environment and is asked to construct a map, which will be used for subsequent navigation, as it moves through the world. The decision as to where to go next is informed only by data contained in the partially complete map. Path finding is a key element in the navigation of a mobile robot. To find a path, a robot should know its position exactly, since the position error exposes a robot to many dangerous conditions.

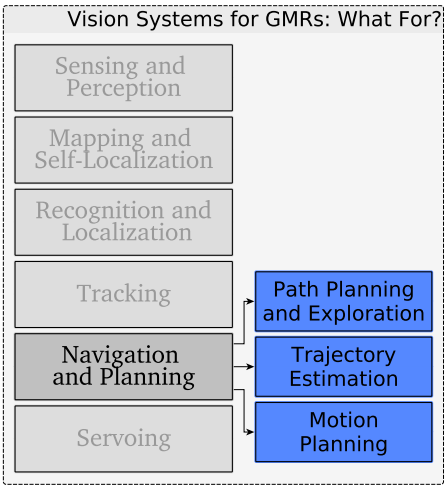

Figure 9. A visual navigation and planning taxonomy

In relation to path planning, one approach considers planning paths that are within the sensing and actuation limits of industrial hardware and software [104]. Building upon recent advances in path planning, the planner augments probabilistic road maps with vision-based constraints. The resulting planner finds collision-free paths that simultaneously avoid occlusions of an image target and keep the target within the field of view of the camera. In addition, virtual reality has been considered for path planning. In this sense, the localization of a mobile robot in its working environment is performed using a vision system and virtual reality modelling language (VRML) [105]. The robot identifies landmarks located in the environment. Image processing and neural network pattern-matching techniques are applied to find location of the robot. After the self-positioning procedure, the 2D scene of the vision is overlaid onto a VRML scene.

As regards exploration, in [106] a technique is presented for mobile robot exploration in unknown indoor environments using just a single forward-facing camera. Rather than processing all of the data, the method intermittently examines only small greyscale images. The method keeps the robot centred in the corridor by estimating two state parameters: the orientation within the corridor and the distance to the end of the corridor. Furthermore, a novel and efficient auto-navigation system based on machine vision for an unknown environment has been developed in [107]. A 3D model using only the measurements of just a single image is implemented for 3D object measurement and map building. For path planning, the well-known A* algorithm is combined with Floyd's shortest path to determine the optimal sub-goals within the image sensing range.

4.4.2. Trajectory estimation

Considering obstacle avoidance for mobile robots, it is useful to estimate/generate an optimal trajectory dynamically in terms of safety and efficiency [108].

The work presented in [109] addresses a local environment recognition system for obstacle avoidance. In vision systems, obstacles that are located beyond the field of view (FOV) cannot be detected accurately. To deal with the FOV problem, the authors propose a 3D panoramic environment map using a modified SURF algorithm. Moreover, in order to determine the avoidance direction and motion automatically, they also propose a complexity measure (CM) and a fuzzy logic-based avoidance motion selector (FL-AMS). The CM is utilized to decide an avoidance direction for obstacles. The avoidance motion is determined using FL-AMS, which considers environmental conditions such as the size of obstacles and the available space. Another paper [110] presents an obstacle avoidance method for a scout robot or an industrial robot in an unknown environment using an IR sensor and a vision system. In the proposed method, robots share information as to where the obstacles are located in real-time; thus, the robots choose the best path for obstacle avoidance.

4.4.3. Motion planning

Motion planning is a term used in robotics for the process of detailing a task into discrete motions. Motion planning for GMRs constitutes a domain of research in which several disciplines meet, ranging from AI and machine learning to robot perception and computer vision [111]. A basic motion planning problem looks to produce a continuous motion that connects a start configuration and a goal configuration, while avoiding collisions with known obstacles. The robot and obstacle geometries are described in a 2D or a 3D workspace, while the motion is represented as a path within the configuration space.

A novel variable, multi-baseline, omnidirectional, stereo vision system for outdoor mobile robot navigation is presented in [112]. The proposed algorithm is implemented on a GPU based on the Nvidia CUDA libraries. Another paper presents a vision- and lidar-based approach to autonomous driving on rural and desert roads that has been tested extensively in a closed-loop system [113]. The vision component uses Gabor wavelet filters for texture analysis to find ruts and tracks from which the road vanishing point is inferred via Hough-style voting, yielding a direction estimate for steering control. In a different work, a vision-based control interface for commanding a robotic wheelchair is presented [114]. The interface estimates the orientation angles of the user's head and it translates these parameters into commands for manoeuvres for different devices.

4.5. Visual tracking

Three different types of applications are considered as regards visual tracking purposes (Figure 10): feature tracking, target tracking and, finally, path following and tracking.

4.5.1. Feature tracking

Feature tracking is one of the most fundamental operations in computer vision. It is probably the most popular means of extracting motion information from a sequence

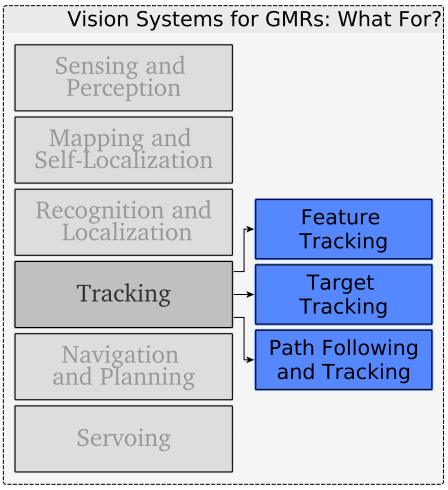

Figure 10. A visual tracking taxonomy

of images. Robust feature tracking is a requirement for many computer vision tasks, such as indoor robot navigation. However, indoor scenes are characterized by poorly localizable features. As result of this, indoor feature tracking without artificial markers is challenging and remains an attractive problem. The work presented in [115] proposes to solve this problem by constraining the locations of a large number of non-distinctive features by several planar homographies which are strategically computed using distinctive features.

Key point detection and matching is of fundamental importance for many applications in computer and robot vision. The association of points across different views is problematic due to image features that undergo significant changes in appearance. Unfortunately, state of the art methods (like SIFT) are not resilient to the radial distortion that often arises in images acquired by cameras with micro-lenses and/or a wide field of view. In [116], modifications to the SIFT algorithm are proposed to substantially improve the repeatability of detection and the effectiveness of matching under radial distortion while preserving the original invariance to scale and rotation. The scale-space representation of the image is obtained using adaptive filtering that compensates for the local distortion, and the key point description is carried after implicit image gradient correction.

4.5.2. Target tracking

Target tracking is the process of locating a moving object (or multiple objects) over time using a camera. The objective of target tracking is to associate target objects in consecutive images.

A general approach is proposed for the simultaneous tracking of multiple moving targets using a generic active stereo setup in [16]. The problem is formulated on a plane, in which cameras are modelled as line scan cameras and targets are described as points with unconstrained motion. Another paper [117] presents the implementation of a real-time tracking algorithm in following and evaluating the 3D position of a generic spatial object. The key issue of our approach is the development of a new algorithm for pattern recognition in machine vision - the least constrained square-fitting of ellipses.

Human tracking is considered as a particular application area for target tracking. In this sense, [118] introduces a multi-agent system approach using the detailed process provided by the Prometheus methodology for the design of a moving robot application for the detection and following of humans. Another article proposes an efficient system which integrates multiple vision models for robust multi-person detection and tracking, used for both service and social mobile robots in public environments. The core technique is a novel maximum likelihood based algorithm which combines multi-model detection in mean-shift tracking [10]. Lastly, a paper addresses the problem of real-time vision-based human tracking to enable mobile robots to follow a human co-worker [119]. Here, an approach to combine stereo vision-based human detection with human tracking using a modified Kalman filter is presented. Stereo vision-based detection combines features extracted from 2D stereo images with reconstructed 3D object features to detect humans in a robot's environment.

4.5.3. Path following and tracking

The path following and tracking problems are chiefly concerned with providing stable motion along a given path with no *a priori* time parametrization associated with movement along a path. More specifically, the control objective is to drive the output of a control system to the path in such a way that the path is traversed in the desired direction [120].

A recent paper presents a novel line of sight control system for a robot vision tracking system which uses a position feedforward controller to preposition a camera and a vision feedback controller to compensate for the positioning error [121]. The camera is rotated in the direction opposite to the motion of the robot. The disturbance compensator consists of two EKFs and a slip detector. A simple approach for vision-based path following for a mobile robot is presented in [122]. Based upon a novel concept called the 'funnel lane', the coordinates of feature points during a replay phase are compared with those obtained during a teaching phase in order to determine the turning direction. The algorithm is qualitative in nature, requiring no map of the environment, no image Jacobian, no homography, no fundamental matrix and no assumption of a flat ground plane.

A proposal to confront the challenge of designing a high-performance dynamic visual servo control scheme is introduced in [123]. Two versatile control laws are developed: a position-based dynamic visual servoing and an image-based dynamic visual servoing. Both control laws are designed to compute the control torques exclusively from a sequential acquisition of regions of interest containing the visual features and in order to achieve accurate trajectory tracking.

4.6. Visual servoing

Visual servoing, also known as 'vision-based robot control', is a technique which uses feedback information extracted from a vision sensor to control the motion of a robot. Visual servoing consists primarily of two techniques. One involves using information from the image to directly control the degrees of freedom of the robot, and is thus referred to as 'image-based visual servoing' (IBVS). The other involves the geometric interpretation of the information extracted from the camera, such as estimating the pose of the target and the parameters of the camera.

A novel approach for the IBVS of a robot manipulator with an eye-in-hand camera is presented in [124]. The camera parameters are not calibrated and the 3D coordinates of the features are not known. Both point and line features are considered. In another paper [125], the image-based regulation control of a robot manipulator with an uncalibrated vision system is discussed. To achieve the control objectives, a Lyapunov-based adaptive control strategy is employed. Another paper proposes an optimal three-dimensional coordinate implementation of a vision sensor using two CCD cameras [126]. Position-based visual servoing is implemented using the positional information obtained from images. The IBVS is also implemented using the difference between the reference and the obtained images. Lastly, [127] presents an image-based visual servoing strategy for the autonomous navigation of a mobile holonomic robot from a current pose towards a desired pose, specified only through a current image and a desired image acquired by the on-board central catadioptric camera. This kind of vision sensor combines lenses and mirrors to enlarge the field of view.

5. How are vision systems applied in GMRs?

There exist several viewpoints as to how to approach vision in GMRs. In most of them, we can identify the two main stages that are depicted in the following: perception and problem modelling. The perception stage receives as input the data provided by robot sensors and extracts relevant information for the problem being solved. This information is then processed to devise models able to solve the task at hand. Figure 11 shows this 'how' taxonomy.

5.1. Perception

Perception in vision-based systems can be defined as the process of extracting information from sensed data. This information can be then sent to the next processing stages. As a first alternative, images that directly encode colour or depth information can be used. These original images are organized as isolated images [128] or video sequences [129] with temporal continuity. The second alternative is related to the use of specific features extracted from the acquired images. The three main points that affect the perception process are shown in Figure 12.

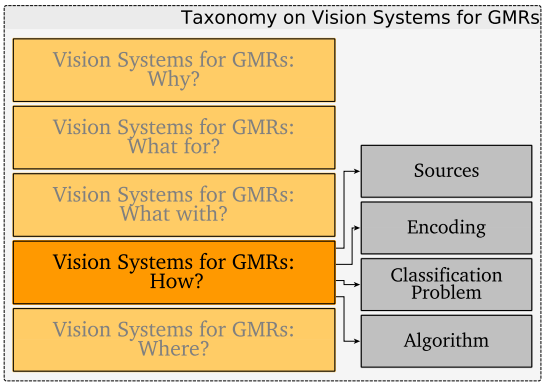

Figure 11. A 'how' taxonomy

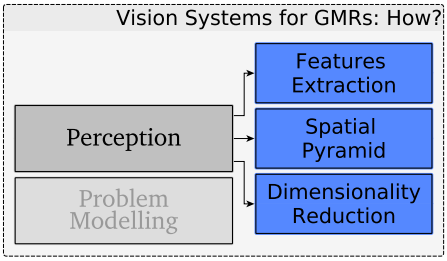

Figure 12. A perception taxonomy

5.1.1. Feature extraction

The main advantage of extracting features from the input images is that redundant or non-relevant information can be removed or properly reduced. This removal allows for a decrease in the amount of data sent to classifiers, which would reduce processing times. However, in some cases it may be necessary to choose some features based on their distinctiveness despite the fact that the new features are larger than the input images.

Feature extraction is performed using general or specific approaches. In other words, this problem is addressed by either relying or not relying upon the intrinsic characteristics of the task environment. The use of such knowledge is especially interesting for places with artificial landmarks or previously defined colour-coded objects, such as the RoboCup competition [130]. There are several examples of the application of this specific knowledge in RoboCup (e.g., [87], [131]). Here, explicit information about players' marks or localization landmarks is introduced into the vision systems aimed at enabling player-localization. Artificial markers (such as ArToolkit [132] and AprilTags [133]) can also be physically attached to the robot, which has proved to be extremely useful for obtaining the real robot position from external cameras [134].

Regarding the use of standard features, there are two main alternatives: local features and global features. Local features are based on extracting information from specific regions of interest (ROIs) or (previously detected) points in the image. The number of ROIs varies from one image to another. Therefore, the number of features and the dimensionality of the descriptor obtained is not fixed.

Global features can be seen as a single local feature where the entire image is always represented as a ROI.

Most global features are based on the use of histograms or vocabularies which are built from colour [135], gradients [136] or more complex information, such as composed receptive field histograms [137]. Since spatial information is completely removed, global features perform better for scene recognition [138] than for object detection.

Local features are expected to present several properties: repeatability, distinctiveness, locality, accuracy and efficiency [139]. Briefly, local features should generate similar descriptors for different visualizations of the same object (people or a scene), but different descriptors for other objects. Some of the most common local features are summarized in the following:

- Harris Laplace Corner Detector [140].
- Local Binary Patterns [141].
- Scale Invariant Feature Transform (SIFT) [142].
- Speed-up Robust Features (SURF) [143].
- Fast Corner Detector (FAST) [144].
- Binary Robust Independent Elementary Features (BRIEF) [145].
- Binary Robust Invariant Scalable Key points (BRISK) [146].
- Oriented Fast and Rotated Brief (ORB) [147].

Scan lines [148] are also considered as a local feature-based approach. This technique is useful in estimating the real distance to colour-coded elements by computing the size of the object projected onto an image. It was widely employed in the Standard Platform League (a former four-legged league) of the RoboCup competition [130], where AIBO robots were fitted with hardware colour filters [149].

The feature extraction process can be also applied to 3D point cloud files instead of visual images. In [150], the authors present a 3D registration procedure that is based on the use of fast point features histograms (FPFHs), which can be considered global features. With respect to the extraction of 3D local features, a remarkable goal was achieved with the release of the normal aligned radial feature (NARF) descriptor [151].

5.1.2. Spatial pyramid

There is an intermediate proposal for generating global features encoding spatial information. It was introduced in 2006 [152] and consists of creating a spatial pyramid where each level corresponds to a new spatial partition of the original image. A histogram is then computed for each one of these regions and, finally, all the histograms are concatenated to produce the global feature. Figure 13 shows an example of a spatial pyramid with two levels and a partition in both the x and y axes. $H_i(j)$ denotes the histogram computed for the i_{th} level and subregion j.

In the example shown in Figure 13, the final features consist of five histograms (of the same size) concatenated together. Although this technique has proven to be optimal for several tasks [153] [154], [155], the large dimensionality of the final descriptor (increasing exponentially with level) constitutes its main drawback. However, the dimensionality of the extracted features can be reduced in an additional step. This step helps to increase the speed of future learning and the classification stages at the expense of deteriorating the performance of those stages.

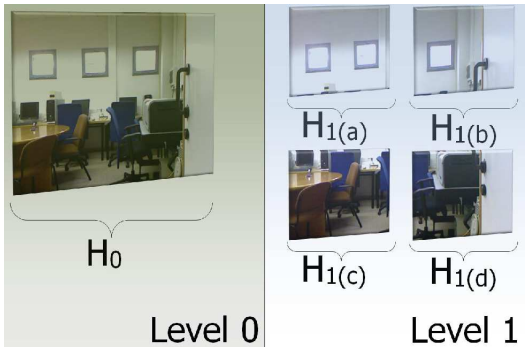

Figure 13. Spatial pyramid encoding with two levels

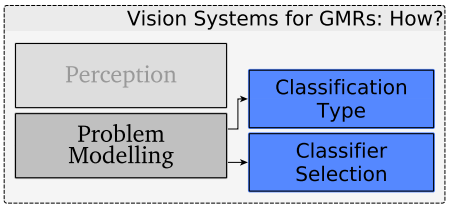

Figure 14. Problem modelling taxonomy

5.1.3. Dimensionality reduction

Principal component analysis (PCA) [156] is a common technique for reducing the large size of pyramidal visual features. Locality-preserving projections [157] are an alternative to PCA whereby the neighbourhood structure of the data set is maintained. Thanks to these approaches, a features dimensionality is reduced with minimal precision loss using an unsupervised method. The dimensionality reduction problem with application to vision systems is exhaustively studied in [158].

5.2. Problem modelling

The second 'how to?' taxonomy is devoted to describing how the vision-based problem is modelled. This stage maps the specific task into a machine learning approach. In this way, we have to firstly identify both the number of classes and the labels for the classes to work with. Next, it is necessary to construct an appropriate classification model for the task to be solved. Figure 14 shows the sub-taxonomy for problem modelling.

5.2.1. Types of classification

There are two main classification types. The first one corresponds to a standard classification problem - that is, the problem is concerned with learning from instances associated with a single label or class. Several robotic applications fit this approach - some of them in the

field of semantic scene classification [159]. The class can be binary (e.g., an indoor/outdoor image classifier) or multi-label (an sunny/cloudy/night image classifier). Binary classifiers can be used to create multi-label classifiers by using one-versus-all or one-versus-one strategies [160].

Multi-label classification [161] is the second type of classification problem. In multi-label classification, instances are linked to two or more classes. This classification type is motivated by problems such as image annotation [162] and object recognition [163], where images present the appearance of several objects (e.g., a bird, clouds and a mountain) or annotations (e.g., music, indoor, classic). Images with multiple labels or tags are nowadays easy to collect through social networks. A good example is the MIR Flickr data-set [164], which was released in 2008 and contains 25,000 annotated images with 1,386 different classes. The classes in this kind of classification can be either binary (as in previous examples) or multi-label. Multi-label classes are used to solve multi-proposal tasks like global scene comprehension [6], where there is a list of objects to be recognized (n binary classes), and the semantic category of the room has to be detected (multi-label class) as well.

5.2.2. Classifier selection

Once the classification problem has been identified, an appropriate classification model has to be adopted. In most cases, such models are trained from a set of samples in an initial stage (a training stage) to be posteriorly used during the robot operation mode to identify visual elements (a classification or recognition stage). In some cases, however, the information extracted during perception or feature extraction explicitly encodes the solution, and no proper classification models are actually needed (e.g., the use of scan lines for distance estimation problems).

The first family of models are denoted as 'lazy classifiers'. The word 'lazy' comes from the fact that no actions are performed until a test instance (an image) is present. We can informally say that the training samples themselves constitute the classification model. Classification is then carried out by computing the similarity between the image to be classified and the set of training instances. K-nearest neighbours [165] is arguably the simplest lazy classifier. More complex approaches include locally weighted regression [166]. Lazy classifiers, despite being simple and having quite reasonable results for categorization tasks [167], present one important drawback: the large amount of data to be stored in memory (the complete training data sets) during the classification stage as well as the computational complexity associated with processing every training sample.

Bayesian classifiers [168] are an alternative. Bayesian approaches have proven to be optimal for information fusion [169], which make them appropriate for working with multiple sources of information (e.g., with several cameras). Two examples of algorithms coming from the family of recursive Bayesian estimators are Kalman filters

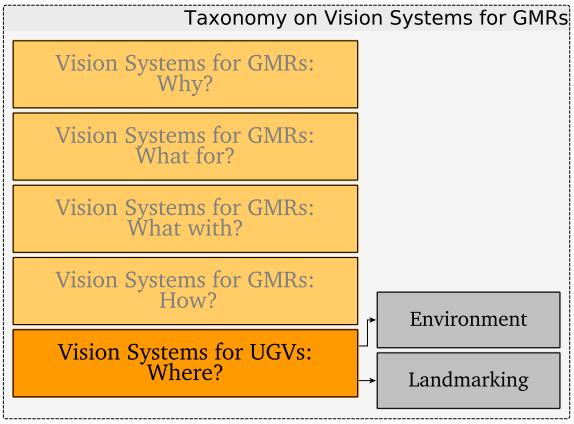

Figure 15. A 'where' taxonomy

[170] and the Monte Carlo method [171]. Both algorithms are standard solutions for solving some of the most common vision-related robotics challenges, such as SLAM [172] and object tracking [173]. Bayesian approaches have also been successfully used in object recognition [174] and navigation [175].

The third family of classifiers considered in this sub-taxonomy are support vector machines [176]. This supervised method is widely used for classification and regression analysis, and it has multiple applications for robotic vision systems. Some examples include object manipulation [177], people detection and tracking [178] or place recognition [179].

6. Where are camera-fitted mobile robots used?

There are two main considerations concerning the location at which a mobile robot should work. First, environmental settings determine whether the robot is located indoors or outdoors. Second, there is additional information given to the vision system through artificial or natural landmarks within the scene. A schema of these types of classification is presented in Figure 15.

These classifications are non-exclusive, as landmarks are situated in either outdoor or indoor scenes. In this section, we will review the specific characteristics of outdoor and indoor environments. We will also see how to use different types of landmarks - as codes or detected objects - to provide additional information to the vision system.

6.1. Environment

The environment for a mobile robot is closely related to its application. It provides useful information for different tasks, such as localization, mapping and navigation. Figure 16 shows the three types of environments considered in this taxonomy.

6.1.1. Indoor locations

The main characteristic of indoor locations is the geometrical distribution of its elements. This helps to find geometric models to represent the information provided

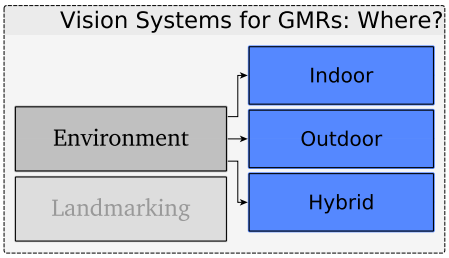

Figure 16. An environment taxonomy

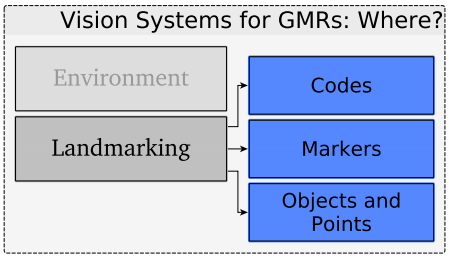

Figure 17. A landmarking taxonomy

by the vision system and interpret it. Taking advantage of this geometrical distribution, [180] proposes a position recognition system based on the straight lines at the sides of a corridor. A similar approach can be found where an obstacle distance estimation for corridor navigation is introduced [181]. Based on a related idea, ceiling vision is widely used for SLAM applications. The parallel lines and corners on the ceiling serve can be easily detected and are relatively useful in localizing an indoor mobile robot, as in [15] and [182].

Moreover, it might be interesting for a GMR to identify structured building elements (e.g., ceilings, walls, columns, and so on). In [183], the structure of the environment is partially reconstructed with the edge-information extracted from the image. This information is then used to detect and track obstacles. In [184], the authors benefit from the shape of the doors to develop door detection used to detect room transition in a localization application. There are several novel examples of vision systems for GMRs in indoor environments. A vision system for manipulation is presented in [185], an obstacle avoidance system is shown in [186], and the authors in [187] exhibit a vision-based people-tracking system.

Despite the fact that lighting conditions are not as extreme as for outdoor environments, vision-based systems have to be prepared to cope with changing conditions, such as sunny or cloudy days and even artificial illumination. This is clearly represented in the RobotVision@ImageCLEF competition [188], where participant proposals [7] have to deal with these conditions.

6.1.2. Outdoor locations

Outdoor locations are characterized by irregular terrains as well as by a high variability in conditions. Terrain irregularities must be correctly detected to avoid the mobile robot becoming involved in dangerous situations. Besides, environmental changes caused by the weather can have an extreme impact on the vision system due to illumination variations. On account of this, vision systems have to be aware of lighting conditions in order to minimize performance degradation. Given the irregularities of outdoor locations, terrain classification is an important task in order for the robot to avoid moving away from the path [189]. In [190], the authors present an image-based path planning method for outdoor terrain environments. Moving objects (e.g., animals, pedestrians and cars) are also something to consider in these kinds of environments.

To cope with all these different challenges (and also to find standard vision-based solutions), general feature-based approaches are commonly used for SLAM, as in [191] and [71], and also for navigation applications [112]. In [192], an algorithm for adaptive image segmentation is proposed. This vision-based solution is used to robustly identify plants from images captured with a GMR under uncontrolled outdoor lighting conditions.

6.1.3. Hybrid scenarios

Finally, there exist some GMRs required to work in both indoor and outdoor locations. In this case, the vision system must be adaptive enough to properly cope with both situations. Generally, vision systems for these types of approaches are based on local feature detection. In [193], an approach to learning efficient navigation policies for mobile robots is introduced. It is based on using SURF visual features for localization. These policies are applied both to indoor and outdoor scenarios. For instance, a monocular navigation system based on a map and replay technique is presented in [194]. This system allows the robot to navigate in large outdoor and indoor environments simply by detecting natural landmarks. In this work, the navigation systems have proven to be robust when facing real-world conditions, such as indoor/outdoor changes, changing illumination, minor environmental changes and partial occlusions.

6.2. Landmarking

In addition to the information provided by the environment and its structure, the robot acquires information from landmarks. In order to achieve this goal, the vision system has to recognize these tags in the scene. In this subsection, we will define a sub-taxonomy (see Figure 17) that reviews three of the most common landmarks used in robotic vision systems: codes, markers and detected objects and points.

6.2.1. Information codes

Information codes provide additional information along with the location as which they were extracted. This type of landmark is easily detected because it is specifically designed for it. There are different types of codes to be used, depending upon the information to

represent. Barcodes are one-dimensional representations of data with spaces and lines of variable width that translate into characters. Quick response (QR) codes are two-dimensional codes that allow the representation of more data than linear barcodes. For instance, a QR code with information about its own position and the position of nearby landmarks is used in [195], while in [196], two-dimensional barcodes encode their absolute position.

ARTag markers [197] are a new type of landmarking code. They consist of a pattern with a square border and they are used in [198] to accurately obtain a mobile robot pose estimation. Similar to this concept is the AprilTag marker [133], a visual fiducial system that uses a 2D barcode-style 'tag', allowing for the full six-DOF localization of features from a single image. In addition to these examples, specific landmark codes are designed by using standard shapes. For instance, a new landmark code based on circles to be placed in ceilings and employed to identify position and direction has been presented [199].

6.2.2. Markers

There are some environments where the addition of landmarks is not necessary, because of their specific configuration. These environments present inherent markers that usually consist of geometrical forms at fixed positions in the scene. Inherent markers do not include any additional information beyond its localization, but they are useful for robot localization and navigation.

Some markers are used to limit parts of the scene, as is the case with the well-known colour lines adopted to limit robot tracks [200]. Other markers are utilized to obtain more accurate localization according to the position at which they are found, such as scattered checker boards that form a unique map [185].

6.2.3. Detected objects and points

The last approach in landmark usage is based on taking advantage of fixed objects within the scene and using them as landmarks. This involves an early step whereby the robot performs a training stage. During this stage, the robot recognizes different objects and points (or even regions or areas) and stores this information for later use. These landmarks are based on non-moving (and easily seen) elements within the scene and, therefore, they are often known as 'natural' landmarks.

Natural landmark detection is usually performed by using several descriptors and features [201]. The detection is carried out from a single location and even from multiple viewpoints. A fleet of GMRs was used for mapping an office-like indoor environment - each robot had its own sensor and all the measurements were fused to create a global single map in [202].

Another popular approach is the use of ceiling landmark positions. This method takes advantage of objects like corners, lamps and doors, and uses its information to perform localization and mapping [203]. Other objects like car wheels are also used as landmarks for pose estimation [68]. For instance, natural landmarks are matched to salient regions (areas of the environment that are easily detected) [204]. The landmarks database is built by moving the robot through the entire path in the environment while storing the salient regions and robot location.

7. Conclusions

In this paper we have presented a novel taxonomy of vision systems for GMRs. The goal of this paper was not only to describe some of the relevant work and advances in robotics vision systems, but also to propose a clear categorization of their internal aspects. The taxonomy proposed is intended to facilitate the identification of the main topics related to robotic vision systems. The questions to be answered (why?, what with?, what for?, how?, and where?) have been thoroughly discussed while describing novel and outstanding proposals in the literature. Thanks to this taxonomy, heterogeneous GMR vision systems can be more easily classified in order to compare them better.

While the 'why?' and 'what for?' questions can be helpful in determining the proper application of a GMR vision system (or an event to discard its use), the remaining questions are useful when addressing the development of such systems. Specifically, 'what with?' deals with the advantages and disadvantages of a wide range of alternatives for vision sensors. The 'how?' argumentation depicts several solutions to the most common vision system problems, whereas the 'where?' question discusses where to exploit the potential of GMR vision systems.

The importance of robotic vision systems as well as some of the most promising current research areas have been detailed and discussed in this paper. It is intended to build a novel taxonomy that can help to effectively organize and classify the significantly numerous scientific papers on this topic written in recent years.

8. Acknowledgements

This work was partially supported by the Spanish Ministerio de Economía y Competitividad / FEDER under TIN2013-47074-C2-1-R, TIN2010-20845-C03-01 and TIN2010-20900-C04-03 grants, and by Innterconecta Programme 2011 project ITC-20111030 ADAPTA.

9. References

[1] A. Casals and A. Fernández-Caballero. Robotics and autonomous systems in the 50th anniversary of artificial intelligence. *Robotics and Autonomous Systems*, 55(12):837–839, 2007.

[2] D. Kragic and M. Vincze. Vision for robotics. *Foundations and Trends in Robotics*, 1(1):1–78, 2009.

[3] A.J. Davison and D.W. Murray. Simultaneous localization and map-building using active vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):865–880, 2002.

[4] N. Karlsson, E. Di Bernardo, J. Ostrowski, L. Goncalves, P. Pirjanian, and M.E. Munich. The vslam algorithm for robust localization and mapping. In *Proceedings of the 2005 IEEE International Conference on Robotics and Automation*, pages 24–29, April 2005.

[5] R. Gonzalez, F. Rodriguez, J.L. Guzman, C. Pradalier, and R. Siegwart. Combined visual odometry and visual compass for off-road mobile robots localization. *Robotica*, 30(6):865–878, 2012.

[6] A. Torralba, K. P. Murphy, W. T. Freeman, and M. A. Rubin. Context-based vision system for place and object recognition. In *Proceedings of the Ninth IEEE International Conference on Computer Vision*, pages 273–280. IEEE, 2003.

[7] J. Martinez-Gomez, A. Jimenez-Picazo, J. A. Gamez, and I. Garcia Varea. Combining invariant features and localization techniques for visual place classification: successful experiences in the robotvision@ imageclef competition. *Journal of Physical Agents*, 5(1):45–54, 2011.

[8] P. Espinace, T. Kollar, N. Roy, and A. Soto. Indoor scene recognition by a mobile robot through adaptive object detection. *Robotics and Autonomous Systems*, 61(9):932–947, 2013.

[9] Terrence Fong, Illah Nourbakhsh, and Kerstin Dautenhahn. A survey of socially interactive robots. *Robotics and Autonomous Systems*, 42(3):143–166, 2003.

[10] L. Li, S. Yan, X. Yu, Y.K. Tan, and H. Li. Robust multiperson detection and tracking for mobile service and social robots. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 42(5):1398–1412, 2012.

[11] C.-W. Park and T. Lee. A robust facial feature detection on mobile robot platform. *Machine Vision and Applications*, 21(6):981–988, 2010.

[12] José M. Gascueña, Francisco J. Garijo, Antonio Fernández-Caballero, Marie-Pierre Gleizes, and André Machonin. Deliberative control components for eldercare robot team cooperation. *Journal of Intelligent and Fuzzy Systems*, in press, 2014.

[13] P.P. Kumar, P. Vadakkepat, and A.P. Loh. Hand posture and face recognition using a fuzzy-rough approach. *International Journal of Humanoid Robotics*, 7(3):331–356, 2010.

[14] Y. Wakuda, K. Sekiyama, and T. Fukuda. Dynamic event interpretation and description from visual scene based on cognitive ontology for recognition by a robot. *International Journal of Robotics and Automation*, 24(3):263–279, 2009.

[15] D. Xu, L. Han, M. Tan, and Y.F. Li. Ceiling-based visual positioning for an indoor mobile robot with monocular vision. *IEEE Transactions on Industrial Electronics*, 56(5):1617–1628, 2009.

[16] J.P. Barreto, L. Perdigoto, R. Caseiro, and H. Araujo. Active stereo tracking of $n \leq 3$ targets using line scan cameras. *IEEE Transactions on Robotics*, 26(3):442–457, 2010.

[17] M. Imran. *Analysis of Vision Systems and Taxonomy Formulation: An Abstract Model for Generalization*. Mid Sweden University, 2011.

[18] M. Imran, K. Benkrid, K. Khursheed, N. Ahmad, M. O'Nils, and N. Lawal. Analysis and characterization of embedded vision systems for taxonomy formulation. In *Real-Time Image and Video Processing*, volume 86560J, 2013.

[19] R. Gade and T. Moeslund. Thermal cameras and applications: a survey. *Machine Vision and Applications*, 25(1):245–262, 2014.

[20] A. Torabi, G. Massé, and G. Bilodeau. An iterative integrated framework for thermal–visible image registration, sensor fusion, and people tracking for video surveillance applications. *Computer Vision and Image Understanding*, 116(2):210–221, 2012.

[21] J. Hwang, S. Jun, S. Kim, D. Cha, K. Jeon, and J. Lee. Novel fire detection device for robotic fire fighting. In *Control Automation and Systems (ICCAS), 2010 International Conference on*, pages 96–100. IEEE, 2010.

[22] N Memarian, T Chau, and A.N. Venetsanopoulos. Application of infrared thermal imaging in rehabilitation engineering: Preliminary results. In *Science and Technology for Humanity (TIC-STH), 2009 IEEE Toronto International Conference*, pages 1–5. IEEE, 2009.

[23] T.R. Gault and A.A. Farag. A fully automatic method to extract the heart rate from thermal video. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on*, pages 336–341. IEEE, 2013.

[24] Bernd Jähne. *Practical Handbook on Image Processing for Scientific Applications*. CRC Press, 1997.

[25] David A. Forsyth and Jean Ponce. *Computer Vision: A Modern Approach*. Prentice-Hall, 2002.

[26] Rafael C. Gonzalez and Richard E. Woods. *Digital Image Processing (3rd Edition)*. Prentice-Hall, 2006.

[27] Maja J. Mataric. *The Robotics Primer*. The MIT Press, 2007.

[28] R. Siegwart, I.R. Nourbakhsh, and D. Scaramuzza. *Introduction to Autonomous Mobile Robots, 2nd edition*. The MIT Press, 2011.

[29] F. Bonin-Font, A. Ortiz, and G. Oliver. Visual navigation for mobile robots: A survey. *Journal of Intelligent and Robotic Systems*, 53(3):263–296, 2008.

[30] M. Staniak and C. Zielinski. Structures of visual servos. *Robotics and Autonomous Systems*, 58(8):940–954, 2010.

[31] S. Fuchs and G. Hirzinger. Extrinsic and depth calibration of tof-cameras. In *26th IEEE Conference on Computer Vision and Pattern Recognition*, volume 4587828, 2008.

[32] Z. Zhao and Y. Weng. A flexible method combining camera calibration and hand-eye calibration. *Robotica*, 31(5):747–756, 2013.

[33] Zhengyou Zhang. Camera calibration with one-dimensional objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(7):892–899, 2004.

[34] L. Krüger and C. Wöhler. Accurate chequerboard corner localisation for camera calibration. *Pattern Recognition Letters*, 32(10):1428–1435, 2011.

[35] Q. Wang, L. Fu, and Z. Liu. Review on camera calibration. In *2010 Chinese Control and Decision Conference*, pages 3354–3358, 2010.

[36] A.S. Huang, E. Antone, M. Olson, L. Fletcher, D. Moore, S. Teller, and J. Leonard. A high-rate, heterogeneous data set from the darpa urban challenge. *International Journal of Robotics Research*, 29(13):1595–1601, 2010.

[37] T. Peynot, S. Scheding, and S. Terho. The marulan data sets: Multi-sensor perception in a natural environment with challenging conditions. *International Journal of Robotics Research*, 29(13):1602–1607, 2010.

[38] C. Tomasi. Early vision. In *Encyclopedia of Cognitive Sciences*. Nature Publishing Group, 2002.

[39] V.J. Traver and A. Bernardino. A review of log-polar imaging for visual perception in robotics. *Robotics and Autonomous Systems*, 58(4):378–398, 2010.

[40] A. Fernández-Caballero, J.C. Castillo, J. Martínez-Cantos, and R. Martínez-Tomás. Optical flow or image subtraction in human detection from infrared camera on mobile robot. *Robotics and Autonomous Systems*, 58(12):1273–1281, 2010.

[41] J. Pavón, J.J. Gómez-Sanz, A. Fernández-Caballero, and J.J. Valencia-Jiménez. Development of intelligent multisensor surveillance systems with agents. *Robotics and Autonomous Systems*, 55(12):892–903, 2007.

[42] M. Bertero, T.A. Poggio, and V. Torre. Ill-posed problems in early vision. *Proceedings of the IEEE*, 76(8):869–889, 1988.

[43] G. Cannata and M. Maggiali. Models for the design of bioinspired robot eye. *IEEE Transactions on Robotics*, 24(1):27–44, 2008.

[44] S. Chen, Y. Li, and N.M. Kwok. Active vision in robotic systems: A survey of recent developments. *International Journal of Robotics Research*, 30(11):1343–1377, 2011.

[45] M. Sridharan and P. Stone. Color learning and illumination invariance on mobile robots: A survey. *Robotics and Autonomous Systems*, 57(6–7):629–644, 2009.

[46] K. Shubina and J.K. Tsotsos. Visual search for an object in a 3d environment using a mobile robot. *Computer Vision and Image Understanding*, 114(5):535–547, 2010.

[47] A. Andreopoulos, S. Hasler, H. Wersing, H. Janssen, J.K. Tsotsos, and E. Körner. Active 3d object localization using a humanoid robot. *IEEE Transactions on Robotics*, 27(1):47–64, 2011.

[48] S. Montabone and A. Soto. Human detection using a mobile platform and novel features derived from a visual saliency mechanism. *Image and Vision Computing*, 28(3):391–402, 2010.

[49] G. Aragon-Camarasa, H. Fattah, and J.P. Siebert. Towards a unified visual framework in a binocular active robot vision system. *Robotics and Autonomous Systems*, 58(3):276–286, 2010.

[50] S.Y. Chen. Kalman filter for robot vision: A survey. *IEEE Transactions on Industrial Electronics*, 59(11):4409–4420, 2012.

[51] C.E. Agüero, F. Martín, L. Rubio, and J.M. Cañas. Comparison of smart visual attention mechanisms for humanoid robots. *International Journal of Advanced Robotic Systems*, 9(6):article number 233, 2012.

[52] Hugh Durrant-Whyte and ThomasC. Henderson. Multisensor data fusion. In Bruno Siciliano and Oussama Khatib, editors, *Springer Handbook of Robotics*, pages 585–610. Springer Berlin Heidelberg, 2008.

[53] J. Solá, A. Monin, M. Devy, and T. Vidal-Calleja. Fusing monocular information in multicamera slam. *IEEE Transactions on Robotics*, 24(5):958–968, 2008.

[54] L. Susperregi, A. Arruti, E. Jauregi, B. Sierra, J.M. Martínez-Otzeta, E. Lazkano, and A. Ansuategui. Fusing multiple image transformations and a thermal sensor with kinect to improve person detection ability. *Engineering Applications of Artificial Intelligence*, 26(8):1980–1991, 2013.

[55] G.L. Mariottini, S. Scheggi, B. Morbidi, and D. Prattichizzo. An accurate and robust visual-compass algorithm for robot-mounted omnidirectional cameras. *Robotics and Autonomous Systems*, 60(9):1179–1190, 2012.

[56] J.-Y. Choi and S.-G. Kim. Study on the localization improvement of the dead reckoning using the ins calibrated by the fusion sensor network information. *Journal of Institute of Control, Robotics and Systems*, 18(8):744–749, 2013.

[57] C. Premebida and U. Nunes. Fusing lidar, camera and semantic information: A context-based approach for pedestrian detection. *International Journal of Robotics Research*, 32(3):371–384, 2013.

[58] R.G. Brown and B.R. Donald. Mobile robot self-localization without explicit landmarks. *Algorithmica*, 26:515–559, 2000.

[59] C. Yi, Y.C. Oh, I.H. Suh, and B.-U. Choi. Indoor place classification using robot behavior and vision data. *International Journal of Advanced Robotic Systems*, 8(5):49–60, 2011.

[60] S.-Y. Park, S.-I. Choi, J.-S. Jang, S.K. Jung, J. Kim, and J.S. Chae. Localization of unmanned ground vehicle using 3d registration of dsm and multiview range images: Application in virtual environment. *Journal of Institute of Control, Robotics and Systems*, 15(7):700–710, 2009.

[61] M.Y. Kim, S.T. Ahn, and H. Cho. Bayesian sensor fusion of monocular vision and laser structured light sensor for robust localization of a mobile robot. *Journal of Institute of Control, Robotics and Systems*, 16(4):381–390, 2010.

[62] Y. Feng, J. Ren, J. Jiang, M. Halvey, and J.M. Jose. Effective venue image retrieval using robust feature extraction and model constrained matching for mobile robot localization. *Machine Vision and Applications*, 23(5):1011–1027, 2012.

[63] O. Booij, Z. Zivkovic, and B. Kröse. Efficient data association for view based slam using connected dominating sets. *Robotics and Autonomous Systems*, 57(12):1225–1234, 2009.

[64] J. Kim, M.-S. Lim, and J. Lim. Omni camera vision-based localization for mobile robots navigation using omni-directional images. *Journal of Institute of Control, Robotics and Systems*, 17(3):206–210, 2011.

[65] Sebastian Thrun. *Robotic Mapping: A Survey*. Technical Report, Carnegie-Mellon University, 2002.

[66] M.-H. Li, B.-R. Hong, Z.-S. Cai, S.-H. Piao, and Q.-C. Huang. Novel indoor mobile robot navigation using monocular vision. *Engineering Applications of Artificial Intelligence*, 21(3):485–497, 2008.

[67] S. Almansa-Valverde, J.C. Castillo, and A. Fernández-Caballero. Mobile robot map building from time-of-flight camera. *Expert Systems with Applications*, 39(10):8835–8843, 2012.

[68] R. Ross, A. Martchenko, and J. Devlin. A 3-degree of freedom binary search pose estimation techniqu. *Machine Vision and Applications*, 24(4):769–776, 2013.

[69] P. Henry, M. Krainin, E. Herbst, X. Ren, and D. Fox. Rgb-d mapping: Using kinect-style depth cameras for dense 3d modeling of indoor environments. *International Journal of Robotics Research*, 31(5):647–663, 2012.

[70] S. Kim and S.-Y. Oh. Slam in indoor environments using omni-directional vertical and horizontal line features. *Journal of Intelligent and Robotic Systems*, 51(1):31–43, 2008.

[71] D. Schleicher, L.M. Bergasa, M. Ocaña, R. Barea, and E. López. Real-time hierarchical stereo visual slam in large-scale environments. *Robotics and Autonomous Systems*, 58(8):991–1002, 2010.

[72] A. Chatterjee, O. Ray, A. Chatterjee, and A. Rakshit. Development of a real-life ekf based slam system for mobile robots employing vision sensing. *Expert Systems with Applications*, 38(7):8266–8274, 2011.

[73] W. Maddern, M. Milford, and G. Wyeth. Cat-slam: Probabilistic localisation and mapping using a continuous appearance-based trajectory. *International Journal of Robotics Research*, 31(4):429–451, 2012.

[74] C. Ulas and H. Temeltas. A fast and robust feature-based scan-matching method in 3d slam and the effect of sampling strategies. *International Journal of Advanced Robotic Systems*, 10, 2013.

[75] T.-D. Vu, J. Burlet, and O. Aycard. Grid-based localization and local mapping with moving object detection and tracking. *Information Fusion*, 12(1):58–69, 2011.

[76] K. M. Wurm, A. Hornung, M. Bennewitz, C. Stachniss, and W. Burgard. Octomap: A probabilistic, flexible, and compact 3d map representation for robotic systems. In *Proc. of the ICRA 2010 workshop on best practice in 3D perception and modeling for mobile manipulation*, volume 2, 2010.

[77] T.T. Cocias, F. Florin Moldoveanu, and S.M. Grigorescu. Generic fitted shapes (gfs): Volumetric object segmentation in service robotics. *Robotics and Autonomous Systems*, 61(9):960–972, 2013.

[78] D. Beale, P. Iravani, and P. Hall. Probabilistic models for robot-based object segmentation. *Robotics and Autonomous Systems*, 59(12):1080–1089, 2011.

[79] J. Xiao, J. Zhang, B. Adler, H. Zhang, and J. Zhang. Three-dimensional point cloud plane segmentation in both structured and unstructured environments. *Robotics and Autonomous Systems*, 61(12):1641–1652, 2013.

[80] J. Xiao, B. Adler, J. Zhang, and H. Zhang. Planar segment based three-dimensional point cloud registration in outdoor environments. *Journal of Field Robotics*, 30(4):552–582, 2013.

[81] Y. Do. Dividing occluded humans based on an artificial neural network for the vision of a surveillance robot. *Journal of Institute of Control, Robotics and Systems*, 15(5):505–510, 2009.

[82] V. Alvarez-Santos, X.M. Pardo, R. Iglesias, A. Canedo-Rodriguez, and C.V. Regueiro. Feature analysis for human recognition and discrimination: Application to a person-following behaviour in a mobile robot. *Robotics and Autonomous Systems*, 60(8):1021–1036, 2012.

[83] J.C. Castillo, J. Serrano-Cuerda, A. Fernández-Caballero, and M.T. López. Segmenting humans from mobile thermal infrared imagery. In *Bioinspired Applications in Artificial and Natural Computation*, pages 334–343. Springer, 2009.

[84] S. Yeom and Y.-H. Woo. Person-specific face detection in a scene with optimum composite filtering and colour-shape information. *International Journal of Advanced Robotic Systems*, 10:70, 2013.

[85] J. Zhang, J. Zhang, and S. Chen. Discover novel visual categories from dynamic hierarchies using multimodal attributes. *IEEE Transactions on Industrial Informatics*, 9(3):1688–1696, 2013.

[86] M. Mata, J.M. Armingol, J. Fernández, and A. De La Escalera. Object learning and detection using evolutionary deformable models for mobile robot navigation. *Robotica*, 26(1):99–107, 2008.

[87] N. Weiss. Adaptive supervision of moving objects for mobile robotics applications. *Robotics and Autonomous Systems*, 57(10):982–995, 2009.

[88] J. Ma, T.H. Chung, and J. Burdick. A probabilistic framework for object search with 6-dof pose estimation. *International Journal of Robotics Research*, 30(10):1209–1228, 2011.

[89] B. Willimon, S. Birchfield, and I. Walker. Interactive perception of rigid and non-rigid objects. *International Journal of Advanced Robotic Systems*, 9:227, 2012.

[90] D. Gálvez-López and J.D. Tardós. Bags of binary words for fast place recognition in image sequences. *IEEE Transactions on Robotics*, 28(5):1188–1197, 2012.

[91] F. Marino, P. De Ruvo, G. De Ruvo, M. Nitti, and E. Stella. Hiper 3-d: An omnidirectional sensor for high precision environmental 3-d reconstruction. *IEEE Transactions on Industrial Electronics*, 59(1):579–591, 2012.

[92] D. Amarasinghe, G.K.I. Mann, and R.G. Gosine. Landmark detection and localization for mobile robot applications: A multisensor approach. *Robotica*, 28(5):663–673, 2010.

[93] J. Leitner, S. Harding, M. Frank, A. Förster, and J. Schmidhuber. Learning spatial object localization from vision on a humanoid robot. *International Journal of Advanced Robotic Systems*, 9:243, 2012.

[94] F. Janabi-Sharifi and M. Marey. A kalman-filter-based method for pose estimation in visual servoing. *IEEE Transactions on Robotics*, 26(5):939–947, 2010.

[95] B. Browning, J.E. Deschaud, D. Prasser, and P. Rander. 3d mapping for high-fidelity unmanned ground vehicle lidar simulation. *International Journal of Robotics Research*, 31(12):1349–1376, 2012.

[96] E. Chinellato, B.J. Grzyb, and A.P. Del Pobil. Pose estimation through cue integration: A neuroscience-inspired approach. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 42(2):530–538, 2012.

[97] J.M. Rendón-Mancha, A. Cárdenas, M.A. García, E. González-Galván, and B. Lara. Robot positioning using camera-space manipulation with a linear camera model. *IEEE Transactions on Robotics*, 26(4):726–733, 2010.

[98] A. Ranganathan. Pliss: labeling places using online changepoint detection. *Autonomous Robots*, 32(4):351–368, 2012.

[99] S.M. Grigorescu, G. Macesanu, T.T. Cocias, D. Puiu, and F. Moldoveanu. Robust camera pose and scene structure analysis for service robotics. *Robotics and Autonomous Systems*, 59(11):899–909, 2011.

[100] M. Gauss, A. Buerkle, T. Laengle, J. Woern, J. Stelter, S. Ruhmkorf, and R. Middelmann. Adaptive robot based visual inspection of complex parts. In *Proceedings of the 34th International Symposium on Robotics*, pages 1–9, 2003.

[101] J.-O. Kim and D.-J. Park. Development of a vision-based crack detection algorithm for bridge inspection. *Journal of Institute of Control, Robotics and Systems*, 14(7):642–646, 2008.

[102] C. Choi, B. Park, H. Jung, and S. Jung. Inch-worm robot with automatic pipe tracking capability for the feeder pipe inspection of a phwr. *Journal of Institute of Control, Robotics and Systems*, 14(2):125–132, 2008.

[103] K. Kawasue and T. Komatsu. Shape measurement of a sewer pipe using a mobile robot with computer vision. *International Journal of Advanced Robotic Systems*, 10:article number 52, 2013.

[104] M. Baumann, S. Léonard, E.A. Croft, and J.J. Little. Path planning for improved visibility using a probabilistic road map. *IEEE Transactions on Robotics*, 26(1):195–200, 2010.

[105] E.-H. Son, Y.-C. Kim, and K.-T. Chong. Implementation of path finding method using 3d mapping for autonomous robotic. *Journal of Institute of Control, Robotics and Systems*, 14(2):168–177, 2008.

[106] V.N. Murali and S.T. Birchfield. Autonomous exploration using rapid perception of low-resolution image information. *Autonomous Robots*, 32(2):115–128, 2012.

[107] M.-S. Chang and J.-H. Chou. A novel machine vision-based mobile robot navigation system in an unknown environment. *International Journal of Robotics and Automation*, 25(4):344–351, 2010.

[108] I. Okawa and K. Nonaka. Optimal online generation of obstacle avoidance trajectory running on a low speed embedded cpu for vehicles. In *Proceedings of the IEEE International Conference on Control Applications*, pages 1257–1262, 2010.

[109] T.K. Kang, I.-H. Choi, G.-T. Park, and M.-T. Lim. Local environment recognition system using modified surf-based 3d panoramic environment map for obstacle avoidance of a humanoid robot. *International Journal of Advanced Robotic Systems*, 10:275, 2013.

[110] B.-S. Jeon, D.-Y. Lee, I.-H. Choi, Y.-H. Mo, J.-M. Park, and M.-T. Lim. Obstacle avoidance method for multi-agent robots using ir sensor and image information. *Journal of Institute of Control, Robotics and Systems*, 18(12):1122–1131, 2012.

[111] Panagiotis Papadakis. Terrain traversability analysis methods for unmanned ground vehicles: A survey. *Engineering Applications of Artificial Intelligence*, 26(4):1373–1385, 2013.

[112] W.L.D. Lui and R. Jarvis. Eye-full tower: A gpu-based variable multibaseline omnidirectional stereovision system with automatic baseline selection for outdoor mobile robot navigation. *Robotics and Autonomous Systems*, 58(6):747–761, 2010.

[113] C. Rasmussen. Roadcompass: following rural roads with vision + ladar using vanishing point tracking. *Autonomous Robots*, 25(3):205–229, 2008.

[114] E. Perez, C. Soria, O. Nasisi, T.F. Bastos, and V. Mut. Robotic wheelchair controlled through a vision-based interface. *Robotica*, 30(5):691–708, 2012.

[115] R. Rodrigo, M. Zouqi, Z. Chen, and J. Samarabandu. Robust and efficient feature tracking for indoor navigation. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 39(3):658–671, 2009.

[116] M. Lourenco, J.P. Barreto, and F. Vasconcelos. Srd-sift: Keypoint detection and matching in images with radial distortion. *IEEE Transactions on Robotics*, 28(3):752–760, 2012.

[117] N. Greggio, A. Bernardino, C. Laschi, J. Santos-Victor, and P. Dario. Real-time 3d stereo tracking and localizing of spherical objects with the icub robotic platform. *Journal of Intelligent &Robotic Systems*, 63(3–4):417–446, 2011.

[118] J.M. Gascueña and A. Fernández-Caballero. Agent-oriented modeling and development of a person-following mobile robot. *Expert Systems with Applications*, 38(4):4280–4290, 2011.

[119] E. Petrovic, A. Leu, D. Ristic-Durrant, and V. Nikolic. Stereo vision-based human tracking for robotic follower. *International Journal of Advanced Robotic Systems*, 10:article number 230, 2013.

[120] Christopher Nielsen, Cameron Fulford, and Manfredi Maggiore. Path following using transverse feedback linearization: Application to a maglev positioning system. *Automatica*, 46:585–590, 2010.

[121] J. Park, W. Hwang, H. Kwon, K. Kim, and D.-I.D. Cho. A novel line of sight control system for a robot vision tracking system, using vision feedback and motion-disturbance feedforward compensation. *Robotica*, 31(1):99–112, 2013.

[122] Z. Chen and S.T. Birchfield. Qualitative vision-based path following. *IEEE Transactions on Robotics*, 25(3):749–754, 2009.

[123] R. Dahmouche, N. Andreff, Y. Mezouar, O. Ait-Aider, and P. Martinet. Dynamic visual servoing from sequential regions of interest acquisition. *International Journal of Robotics Research*, 31(4):520–537, 2012.

[124] H. Wang, Y.-H. Liu, and D. Zhou. Adaptive visual servoing using point and line features with an uncalibrated eye-in-hand camera. *IEEE Transactions on Robotics*, 24(4):843–857, 2008.

[125] E. Tatlicioglu, D.M. Dawson, and B. Xian. Adaptive visual servo regulation control for camera-in-hand configuration with a fixed camera extension. *International Journal of Robotics and Automation*, 24(4):346–355, 2009.

[126] J. Lee and J.M. Lee. A study on the visual servoing of autonomous mobile inverted pendulum. *Journal of Institute of Control, Robotics and Systems*, 19(3):240–247, 2013.

[127] G.L. Mariottini and D. Prattichizzo. Image-based visual servoing with central catadioptric cameras. *International Journal of Robotics Research*, 27(1):41–56, 2008.

[128] N. Kumar, P. N. Belhumeur, A. Biswas, D. W. Jacobs, W. J. Kress, I. C. Lopez, and J.V. Soares. Leafsnap: A computer vision system for automatic plant species identification. In *Computer Vision–ECCV 2012*, pages 502–516. Springer, 2012.

[129] T. Jaeggli, E. Koller-Meier, and L. Van Gool. Learning generative models for multi-activity body pose estimation. *International Journal of Computer Vision*, 83(2):121–134, 2009.

[130] H. Kitano, M. Asada, Y. Kuniyoshi, I. Noda, and E. Osawa. Robocup: The robot world cup initiative. In *Proceedings of the First International Conference on Autonomous Agents*, pages 340–347. ACM, 1997.

[131] P. Khandelwal and P. Stone. A low cost ground truth detection system for robocup using the kinect. In *RoboCup 2011: Robot Soccer World Cup XV*, pages 515–527. Springer, 2012.

[132] H. Kato and M. Billinghurst. Marker tracking and hmd calibration for a video-based augmented reality conferencing system. In *Proceedings of the 2nd IEEE and ACM International Workshop on Augmented Reality*, pages 85–94. IEEE, 1999.

[133] Olson E. AprilTag: A robust and flexible visual fiducial system. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 3400–3407. IEEE, May 2011.

[134] S. Ceriani, G. Fontana, A. Giusti, D. Marzorati, M. Matteucci, D. Migliore, D. Rizzi, D.G. Sorrenti, and P. Taddei. Rawseeds ground truth collection systems for indoor self-localization and mapping. *Autonomous Robots*, 27(4):353–371, 2009.

[135] M. J. Swain and D. H. Ballard. Color indexing. *International Journal of Computer Vision*, 7(1):11–32, 1991.

[136] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the IEEE Computer Society Conference on Vision and Pattern Recognition*, volume 1, pages 886–893 vol. 1, 2005.

[137] O. Linde and T. Lindeberg. Object recognition using composed receptive field histograms of higher dimensionality. In *Proceedings of the 17th International Conference on Pattern Recognition*, volume 2, pages 1–6 Vol.2, 2004.

[138] Felix W., Joaquin S., and Frederic D. M. Automatic place determination using colour histograms and self-organising maps. pages 111–116, 2007.

[139] T. Tuytelaars and K. Mikolajczyk. Local invariant feature detectors: a survey. *Foundations and Trends® in Computer Graphics and Vision*, 3(3):177–280, 2008.

[140] K. Mikolajczyk and C. Schmid. Indexing based on scale invariant interest points. In *Proceedings of the Eighth IEEE International Conference on Computer Vision*, volume 1, pages 525–531 vol.1, 2001.

[141] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002.

[142] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[143] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. In *Computer Vision–ECCV 2006*, pages 404–417. Springer, 2006.

[144] E Rosten, R Porter, and T Drummond. Faster and better: A machine learning approach to corner detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(1):105–119, 2010.

[145] M. Calonder, V. Lepetit, C. Strecha, and P. Fua. Brief: Binary robust independent elementary features. In *11th European Conference on Computer Vision*, pages 778–792. Springer, 2010.

[146] S. Leutenegger, M. Chli, and R.Y. Siegwart. Brisk: Binary robust invariant scalable keypoints. In *2011 IEEE International Conference onComputer Vision*, pages 2548–2555. IEEE, 2011.

[147] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. Orb: an efficient alternative to sift or surf. In *2011 IEEE International Conference on Computer Vision*, pages 2564–2571. IEEE, 2011.

[148] M. Jüngel, J. Hoffmann, and M. Lötzsch. A real-time auto-adjusting vision system for robotic soccer. In *RoboCup 2003: Robot Soccer World Cup VII*, pages 214–225. Springer, 2004.

[149] R.A. Palma-Amestoy, P.A. Guerrero, P.A. Vallejos, and J. Ruiz-del Solar. Context-dependent color segmentation for aibo robots. In *Proceedings of the IEEE 3rd Latin American Robotics Symposium*, pages 128–136. IEEE, 2006.

[150] R.B. Rusu, N. Blodow, and M. Beetz. Fast point feature histograms (fpfh) for 3d registration. In *IEEE International Conference on Robotics and Automation*, pages 3212–3217. IEEE, 2009.

[151] B. Steder, R.B. Rusu, K. Konolige, and W. Burgard. Point feature extraction on 3d range scans taking into account object boundaries. In *2011 IEEE International Conference on Robotics and Automation*, pages 2601–2608. IEEE, 2011.

[152] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 2169–2178. IEEE, 2006.

[153] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman. Multiple kernels for object detection. In *Proceedings of the IEEE 12th International Conference onComputer Vision*, pages 606–613. IEEE, 2009.

[154] J. Liu and M. Shah. Learning human actions via information maximization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.

[155] A. Bosch, A. Zisserman, and X. Muñoz. Scene classification using a hybrid generative/discriminative approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(4):712–727, 2008.

[156] I. Jolliffe. *Principal component analysis*. Wiley Online Library, 2005.

[157] X. Niyogi. Locality preserving projections. In *Neural information processing systems*, volume 16, page 153, 2004.

[158] Y.Y. Lin, T.L. Liu, and C.S. Fuh. Multiple kernel learning for dimensionality reduction. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(6):1147–1160, 2011.

[159] C. Siagian and L. Itti. Rapid biologically-inspired scene classification using features shared with visual attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(2):300–312, 2007.

[160] R. Rifkin and A. Klautau. In defense of one-vs-all classification. *The Journal of Machine Learning Research*, 5:101–141, 2004.

[161] G. Tsoumakas and I. Katakis. Multi-label classification: an overview. *International Journal of Data Warehousing and Mining*, 3(3):1–13, 2007.

[162] S. Nowak and M. J. Huiskes. New strategies for image annotation: Overview of the photo annotation task at imageclef 2010. In *CLEF (Notebook Papers/LABs/Workshops)*, volume 1, page 4. Citeseer, 2010.

[163] B.C. Russell, W.T. Freeman, A.A. Efros, J. Sivic, and A. Zisserman. Using multiple segmentations to discover objects and their extent in image collections. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 1605–1614. IEEE, 2006.

[164] M.J. Huiskes and M.S. Lew. The mir flickr retrieval evaluation. In *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval*, pages 39–43. ACM, 2008.

[165] T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967.

[166] W. S. Cleveland and C. Loader. Smoothing by local regression: Principles and methods. In *Statistical Theory and Computational Aspects of Smoothing*, pages 10–49. Springer, 1996.

[167] A. C. Berg, T. L. Berg, and J. Malik. Shape matching and object recognition using low distortion correspondences. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 26–33. IEEE, 2005.

[168] N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine Learning*, 29(2-3):131–163, 1997.

[169] R. P. S. Mahler. *Statistical multisource-multitarget information fusion*, volume 685. Artech House Boston, 2007.

[170] R. E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35–45, 1960.

[171] Frank Dellaert, Dieter Fox, Wolfram Burgard, and Sebastian Thrun. Monte carlo localization for mobile robots. In *Proceedings of the IEEE International Conference on Robotics and Automation*, volume 2, pages 1322–1328. IEEE, 1999.

[172] S. Huang and G. Dissanayake. Convergence and consistency analysis for extended kalman filter based slam. *IEEE Transactions on Robotics*, 23(5):1036–1049, 2007.

[173] D. Angelova and L. Mihaylova. Extended object tracking using monte carlo methods. *IEEE Transactions on Signal Processing*, 56(2):825–832, 2008.

[174] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding*, 106(1):59–70, 2007.

[175] L.F. Posada, K.K. Narayanan, F. Hoffmann, and T. Bertram. Floor segmentation of omnidirectional images for mobile robot visual navigation. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 804–809. IEEE, 2010.

[176] Vladimir Vapnik. *The Nature of Statistical Learning Theory*. Springer, 2000.

[177] R. Pelossof, A. Miller, P. Allen, and T. Jebara. An svm learning approach to robotic grasping. In *Proceedings of the IEEE International Conference on Robotics and Automation*, volume 4, pages 3512–3518. IEEE, 2004.

[178] Junji Satake and Jun Miura. Robust stereo-based person detection and tracking for a person following robot. In *ICRA Workshop on People Detection and Tracking*, 2009.

[179] J. Luo, A. Pronobis, B. Caputo, and P. Jensfelt. Incremental learning for place recognition in dynamic environments. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 721–728. IEEE, 2007.

[180] E. Hayashi and T. Kinoshita. Development of an indoor navigation system for a monocular-vision-based autonomous mobile robot. *Artificial Life and Robotics*, 14(3):324–328, 2009.

[181] Z. Zhou, T. Chen, D. Wu, and C. Yu. Corridor navigation and obstacle distance estimation for monocular vision mobile robots. *International Journal of Digital Content Technology and its Applications*, 5(3):192–202, 2011.

[182] H. Chen, D. Sun, J. Yang, and J. Chen. Localization for multirobot formations in indoor environment. *IEEE/ASME Transactions on Mechatronics*, 15(4):561–574, 2010.

[183] M. Marrón-Romera, J.C. García, M.A. Sotelo, D. Pizarro, M. Mazo, J.M. Cañas, C. Losada, and A. Marcos. Stereo vision tracking of multiple objects in complex indoor environments. *Sensors*, 10(10):8865–8887, 2010.

[184] J. Martinez and B. Caputo. Towards semi-supervised learning of semantic spatial concepts for mobile robots. *Journal of Physical Agents*, 4(3):19–31, 2010.

[185] S.S. Srinivasa, D. Ferguson, C.J. Helfrich, D. Berenson, A. Collet, R. Diankov, G. Gallagher, G. Hollinger, J. Kuffner, and M.V. Weghe. Herb: A home exploring robotic butler. *Autonomous Robots*, 28(1):5–20, 2010.

[186] W. Budiharto, A. Santoso, D. Purwanto, and A. Jazidie. A new method of obstacle avoidance for service robots in indoor environments. *ITB Journal of Engineering Science*, 44 B(2):148–167, 2012.

[187] C. Hu, X. Ma, X. Dai, and K. Qian. Reliable people tracking approach for mobile robot in indoor environments. *Robotics and Computer-Integrated Manufacturing*, 26(2):174–179, 2010.

[188] B. Caputo, H. Muller, B. Thomee, M. Villegas, R. Paredes, D. Zellhofer, H. Goeau, A. Joly, P. Bonnet, J. Martinez-Gomez, et al. Imageclef 2013: the vision, the data and the open challenges. In *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, pages 250–268. Springer, 2013.

[189] Y. Morales, A. Carballo, E. Takeuchi, A. Aburadani, and T. Tsubouchi. Autonomous robot navigation in outdoor cluttered pedestrian walkways. *Journal of Field Robotics*, 26(8):609–635, 2009.

[190] W.H. Huang, M. Ollis, M. Happold, and B.A. Stancil. Image-based path planning for outdoor mobile robots. *Journal of Field Robotics*, 26(2):196–211, 2009.

[191] J. Courbon, Y. Mezouar, and P. Martinet. Autonomous navigation of vehicles from a visual memory using a generic camera model. *IEEE Transactions on Intelligent Transportation Systems*, 10(3):392–402, 2009.

[192] H.Y. Jeon, L.F. Tian, and H. Zhu. Robust crop and weed segmentation under uncontrolled outdoor illumination. *Sensors*, 11(6):6270–6283, 2011.

[193] A. Hornung, M. Bennewitz, and H. Strasdat. Efficient vision-based navigation. *Autonomous Robots*, 29(2):137–149, 2010.

[194] T. Krajník, J. Faigl, V. Vonásek, K. Kosnar, M. Kulich, and L. Preucil. Simple yet stable bearing-only navigation. *Journal of Field Robotics*, 27(5):511–533, 2010.

[195] Y. Xue, G. Tian, B. Song, and T. Zhang. Distributed environment representation and object localization system in intelligent space. *Journal of Control Theory and Applications*, 10(3):371–379, 2012.

[196] G. Lin and X. Chen. A robot indoor position and orientation method based on 2d barcode landmark. *Journal of Computers*, 6(6):1191–1197, 2011.

[197] M. Fiala. Designing highly reliable fiducial markers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(7):1317–1324, 2010.

[198] F. Lamberti, A. Sanna, G. Paravati, P. Montuschi, V. Gatteschi, and C. Demartini. Mixed marker-based/marker-less visual odometry system for mobile robots. *International Journal of Advanced Robotic Systems*, 10, 2013.

[199] A. Rusdinar, J. Kim, J. Lee, and S. Kim. Implementation of real-time positioning system using extended kalman filter and artificial landmark on ceiling. *Journal of Mechanical Science and Technology*, 26(3):949–958, 2012.

[200] E.-J. Jung and B.-J. Yi. Task-oriented navigation algorithms for an outdoor environment with colored borders and obstacles. *Intelligent Service Robotics*, 6(2):69–77, 2013.

[201] A. Gil, O.M. Mozos, M. Ballesta, and O. Reinoso. A comparative evaluation of interest point detectors and local descriptors for visual slam. *Machine Vision and Applications*, 21(6):905–920, 2010.

[202] A. Gil, O. Reinoso, M. Ballesta, M. Juliá, and L. Payá. Estimation of visual maps with a robot network equipped with vision sensors. *Sensors*, 10(5):5209–5232, 2010.

[203] S.-Y. Hwang and J.-B. Song. Monocular vision-based slam in indoor environment using corner, lamp, and door features from upward-looking camera. *IEEE Transactions on Industrial Electronics*, 58(10):4804–4812, 2011.

[204] C. Siagian and L. Itti. Biologically inspired mobile robot vision localization. *IEEE Transactions on Robotics*, 25(4):861–873, 2009.