A state-of-the-art review on mobile robotics tasks using artificial intelligence and visual data

Sergio Cebollada, Luis Payá, Maria Flores, Adrián Peidró,
Oscar Reinoso

PII: S0957-4174(20)30926-X
DOI: https://doi.org/10.1016/j.eswa.2020.114195
Reference: ESWA 114195

To appear in: *Expert Systems With Applications*

Received date : 5 June 2020
Revised date : 10 August 2020
Accepted date : 29 October 2020

A State-of-the-Art Review on Mobile Robotics Tasks Using Artificial Intelligence and Visual Data[⋆]

Sergio Cebollada[a,*], Luis Payá[a], Maria Flores[a], Adrián Peidró[a] and Oscar Reinoso[a]

[a] Department of Systems Engineering and Automation, Miguel Hernández University, Elche, 03202, Spain

ARTICLE INFO

Keywords:
Mobile Robotics
Visual Information
Artificial Intelligence
Mapping
Localization
Navigation
SLAM
Exploration

ABSTRACT

Nowadays, the field of mobile robotics has experienced an important evolution and these robots are more commonly proposed to solve different tasks autonomously. The use of visual sensors has played an important role in mobile robotics tasks during the past few years due to the advances in computer vision hardware and algorithms. It is worth remarking the use of AI tools to solve a variety of problems in mobile robotics based on the use of images either as the only source of information or combining them with other sensors such as laser or GPS. The improvement of the autonomy of mobile robots has attracted the attention of the scientific community. A considerable amount of works have been proposed over the past few years, leading to an extensive variety of approaches. Building a robust model of the environment (mapping), estimating the position within the model (localization) and controlling the movement of the robot from one place to another (navigation) are important abilities that any mobile robot must have. Considering this, this review focuses on analyzing these problems; how researchers have addressed them by means of AI tools and visual information; and how these approaches have evolved in recent years. This topic is currently open and a large number of works can be found in the related literature. Therefore, it can be of interest making an analysis of the current state of the topic. From this review, we can conclude that AI has provided robust solutions to some specific tasks in mobile robotics, such as information retrieval from scenes, mapping, localization and exploration. However, it is worth continuing to develop this line of research to find more integral solutions to the navigation problem so that mobile robots can increase their autonomy in large, complex and heterogeneous environments.

1. Introduction

Over the past few years, the use of mobile robots has significantly increased. Nowadays, they can be used for a wide range of applications and they can be found in diverse kinds of environments, such as industrial, household, educational and healthcare. Regarding mobile autonomous robots, they must be able to navigate through an environment which is usually *a priori* unknown, while simultaneously tackle the task they have been designed for. Hence, the robot must be capable of building a model of the environment, estimating its current position and orientation within the environment by using this model and also navigating throughout the environment to arrive to the target points.

Mapping, localization and navigation are the classical problems in mobile robotics. They have attracted a great attention and nowadays continue being a prominent research area, since a robust solution to these problems is fundamental to increase the autonomy of mobile robots and subsequently expand their use for other applications.

To conduct the mobile robotics tasks, it is necessary to provide the robot with relevant information about the environment. For this purpose, robots are equipped with sensors that allow them to obtain such information. Subsequently, the robots need to process the data captured from the environment and transform them in useful information for their tasks.

Concerning the mapping task, two main frameworks can be highlighted from the related literature: metric and topological maps. On the one hand, the metric maps represent the environment with geometric accuracy. On the other hand, the topological maps lead typically to a graph representation, that is, the environment is described as a graph that contain representative locations and links that connect them. As for the localization task, it tries to estimate the current position and orientation of the robot using a model of the environment. In order to carry out the localization task, the environment must be previously modeled. Hence, in this way, firstly the robot carries out the mapping task and after that, once the map is available, the localization can be done. Nonetheless, the related literature has also studied a blend of both tasks that can be developed at the same time. This concept is known as Simultaneous Localization And Mapping (SLAM) and consists in modeling the environment as the robot moves through it and, at the same time, estimating its position and orientation.

Additionally to the mapping and localization tasks, there are other tasks that are included within the mobile robotics challenges. The navigation task includes the ability of the robot to determine its position within the map, to plan a

[⋆] This work has been supported by the Generalitat Valenciana and the FSE through the grant ACIF/2017/146 and the project AICO/2019/031: "Creación de modelos jerárquicos y localización robusta de robots móviles en entornos sociales"; and by the Spanish government through the project DPI 2016-78361-R (AEI/FEDER, UE): "Creación de mapas mediante métodos de apariencia visual para la navegación de robots."

The authors declare no conflict of interest.

[*] Corresponding author

sergio.cebollada@umh.es (S. Cebollada); lpaya@umh.es (L. Payá); maria.flores03@umh.es (M. Flores); apeidro@umh.es (A. Peidró); o.reinoso@umh.es (O. Reinoso)

ORCID(s): 0000-0003-4047-3841 (S. Cebollada); 0000-0002-3045-4316 (L. Payá); 0000-0003-1117-0868 (M. Flores); 0000-0002-4565-496X (A. Peidró); 0000-0002-1065-8944 (O. Reinoso)

path to reach a target position and to send the necessary commands to the actuators to move the robot while avoiding dynamic obstacles. Hence, the robot must be capable of mapping the environment and interpreting the agents included in it. The main objectives of the robot navigation consist in avoiding collisions such as objects, walls, human beings, etc; and avoiding unsafe places or conditions such as radioactive places, exposing to hazardous places due to high temperatures or other environmentally dangerous conditions. Reinoso & Payá (2020a) present a special issue about the current frameworks in the mobile robots navigation field and a variety of approaches related to this task.

Similar to the concept of SLAM, the mobile robotics field also considers the combination of mapping and navigation. This combination is known as exploration. The exploration task consists basically in guiding a robot in such a way that it covers the environment with its sensors (Stachniss & Burgard, 2003). Exploration approaches are relevant to address surveillance or surface inspection mine sweeping, among others.

Concerning the sensors provided to address the mobile robotics tasks, this work focuses on the use of cameras. This type of sensors have been widely used for these purposes. In this way, Reinoso & Payá (2020b) present a special issue about some of the possibilities that vision systems offer, focusing on the different configurations that can be used and novel applications in fields of application, from mapping for navigation of mobile robots to object recognition or scene reconstruction. Through this kind of sensors, the amount of information collected may be enough to carry out most of the problems related to mobile robotics. Moreover, these sensors present a relatively good relation "quantity of information - cost". However, these approaches present downsides, such as their sensitivity to changes of lighting conditions. For example, underfloor environments may be completely dark and the illumination is only provided by light sources installed either on the robot and/or in a specific position of the environment. Hence, the shadows may generate inaccuracies (Cebollada et al., 2018; Parra et al., 2020).

According to the number of cameras and the field of view, different configurations have been proposed. Some authors such as Okuyama et al. (2011) have used monocular configurations. Others proposed stereo cameras by using binocular (such as Yong-guo et al. (2012) or Gwinner et al. (2016)) or even trinocular systems such as Jia et al. (2003). In order to obtain complete information from the environment, several images must be captured. In this respect, omnidirectional cameras constitute a good alternative. They can provide a big amount of information with a field of view of 360 deg. around them and their cost is relatively low in comparison with other kinds of sensors. Furthermore, omnidirectional vision systems present further advantages. For instance, the features in the images are more stable (because they stay longer as the robot moves) and they permit estimating both the position and the orientation of the robot. Omnidirectional cameras have been successfully used by different authors for mapping and localization (Valiente et al., 2018; Payá et al., 2018; Tardif et al., 2008; A. Murillo et al., 2007; Menegatti et al., 2006). A wide study was carried out by Payá et al. (2017), who introduce a state of the art of the most relevant mapping and localization algorithms developed with omnidirectional visual information.

With this aim, a wide range of approaches have emerged to process the information obtained by the sensory systems. As a result, a variety of methods have been proposed regarding sensory information and processing techniques. Additionally, in recent years, several tools and techniques based on artificial intelligence (AI) have proved their ability to solve a variety of problems with a profound data treatment. The use of these techniques has become very popular among the mobile robotics works and it is worth studying their main applications to solve mapping and localization. Within the AI field, Machine Learning (ML) is a subfield whose algorithms attempt to improve automatically through experience (Mitchell, 1997). More precisely, these algorithms build a mathematical model based on sample data with the aim of carrying out predictions or decisions without being explicitly programmed for a specific purpose (Bishop, 2006). During the past few years, ML has received considerable attention, since this technique is capable of addressing a wide variety of complex problems accurately. There are many machine learning algorithms and they depend on the approach they use, the data type of the input, the data type of the output, and the type of problem that they are designed to solve. On the other hand, supervised learning algorithms consist on building a mathematical model to represent a set of data which contains the inputs and also the desired outputs. By addressing iterative optimization of an objective function, these algorithms learn a function that can be used to predict outputs associated with new inputs (Mehryar et al., 2012). On the other hand, unsupervised learning algorithms take a set of data that contains only inputs. These algorithms try to find a structure in the training data such as grouping or clustering of data points (Hinton et al., 1999).

Currently, deep learning has become the dominant approach for many works in the field of machine learning. Like ML, deep learning algorithms build mathematical models based on sample data to carry out predictions/decisions without being explicitly programmed for a certain purpose using typically an architecture of multiple hidden layers in an artificial neural network. These algorithms try to construct automatically high level data models by using a matrix of initial data and architectures that allow linear, non-linear, multiple and iterative transformations (Bengio et al., 2013). These architectures have been commonly applied during the past few years to fields like computer vision, speech recognition, natural language processing, social network filtering, machine translation, and bioinformatics among others. Its use has proved to provide results comparable to and in some cases surpassing human expert performance (Goodfellow et al., 2016).

The fields of mobile robotics and computer vision have progressed considerably during the past few years. Notwithstanding that, there are still some issues that need to be addressed more robustly to enable mobile robots to move and perform their tasks more autonomously in complex environments, under real operation conditions (Payá et al., 2017). Additionally, the continuous improvement in the capacity and performance of computing devices has extended the use of different AI methods. Therefore, it is worth knowing the contribution of AI to mobile robotics and computer vision, and the research gaps, opportunities and solutions which are contributing to the development of these fields.

Some researchers have presented reviews in the field of mobile robotics. Parker (2000) develops a survey about the methods used to solve tasks in distributed mobile robotics. Muhammad et al. (2009) introduce a review about SLAM methods that make use of visual information. Garcia-Fidalgo & Ortiz (2015) focus on approaches to solve the topological mapping problem. Also, Fuentes-Pacheco et al. (2015) and Kuutti et al. (2018) present a state of the art of the localization techniques for autonomous vehicle applications. Finally, Payá et al. (2017) develop a review of mapping and localization techniques using omnidirectional vision. With respect to these previous reviews, and considering the great development of AI techniques and their use in mobile robotics, in the present work we focus on the use of such techniques. We present the most relevant AI tools in mobile robotics, how they can be used to extract relevant information from the scenes and their application to solve specific problems in mobile robotics.

To summarize, the aim of this review is to present the most relevant works conducted in the field of mapping, localization, navigation, SLAM and exploration by using AI, paying more attention to the developments that are based on visual information. The remainder of the paper is structured as follows. First, section 2 presents the main AI tools used in the robotics field. Then, section 3 presents the main methods to describe the information. After that, section 4 shows the AI techniques that have been proposed to solve the mobile robotics tasks with visual data. Last, section 5 presents a final discussion about the existing approaches.

2. Artificial Intelligence Tools

This section depicts the concept of AI and the areas where this science has been commonly applied in the field of robotics during the last few years. Also, the most relevant techniques are outlined. Subsection 2.1 defines some areas of use in the field of robotics; subsection 2.2 presents a variety of applications that contribute to the autonomy of mobile robots and subsection 2.3 focuses on the problems of mapping and localization and the AI tools used to address these problems.

2.1. Definition and Areas of Use

In the related literature, some definitions of AI can be found. For instance, Schleichert (1970) defined AI as “the science of making machines do things that would require intelligence if done by men”. Charniak et al. (1985) define AI as “the study of mental faculties through the use of computational models”. More recently, according to Schalkoff (1990), AI is “a field of study that seeks to explain and emulate intelligent behavior in terms of computational process”. If we focus on the use of AI to solve robotics tasks, we can describe this science as a set of techniques that are applied in computer programming to solve problems whose difficulty requires a certain degree of intelligence. As for the birth of the AI, many researchers consider that this happened during the Second World War, when the scientist Alan Turing worked to crack the ‘Enigma’ code that was used by German forces to send messages securely. Alan Turing and his team created the Bombe machine, which was used to decipher Enigma’s messages. Both Enigma and Bombe Machines are considered the foundations for Artificial Intelligence (Ray, 2018).

Artificial Intelligence has been widely used in different areas. According to the type of manipulation, we can establish two categories. First, the **physical manipulation**, which covers the fields of computer vision, robotics and control systems. For example, Vyborny & Giger (1994) have used successfully computer vision together with AI in mammography to detect or to characterize abnormalities on digital images. Thanks to it, Radiologists are able to detect anomalies better and then, the errors in mammography interpretation are considerably reduced. Another example of computer vision and AI is proposed by Wachs et al. (2011), who propose a vision based hand gesture recognition for human computer interaction based on an artificial neural network, fuzzy logic, and genetic algorithms. Regarding the use of Artificial Intelligence for Robotics, M. K. Singh & Parhi (2011) propose a neural network to solve the path and time optimization problem of mobile robots. The inputs to the proposed neural controller consist of distances to the obstacles with respect to the position of the robot and target angle. The output of the neural network is the steering angle. De Momi & Ferrigno (2010) propose a backpropagation algorithm in the healthcare field that is used to train the network. The goal is to assist surgeons with a robotic system controlled by an intelligent high-level controller (HLC) able to gather and integrate information from the surgeon, from diagnostic images, and from an array of on-field sensors. Last, regarding control systems, Wong et al. (2010) propose a novel modeling and optimization approach for steady state and transient performance tune-up of an engine at idle speed. In terms of electric control, a genetic algorithm and particle swarm optimization are applied to obtain an optimal control unit setting automatically. Gadoue et al. (2009) present a comparison between four different speed controller design strategies based on AI techniques; two are based on tuning of conventional PI (Proportional–Integral) controllers, the third makes use of a fuzzy logic controller and the last is based on hybrid fuzzy sliding mode control theory.

Regarding the **thinking manipulation**, this branch covers fields such as Natural Language Processing, Data Mining, Neural Networks, Automatic Learning, and Pattern Recognition. There are also a substantial amount of works related to them. To cite some examples, Kohavi & Quinlan (2002) present data mining tasks by using decision-tree discovery for classification. Xing et al. (2015) propose a learning-based framework for robust and automatic nucleus segmentation with shape preservation. Krittanawong et al. (2017) review the recent AI applications in cardiovascular clinical care and discuss its potential role in facilitating precision cardiovascular medicine. Calderon-Cordova et al. (2016) conduct the design and development of the system architecture to recognition of Electromyography signal patterns by using Feedforward backpropagation Artificial Neural Network. Minaei-Bidgoli et al. (2003) introduce an approach to classify students in order to predict their final grade based on features extracted from logged big data in an education Web-based system and they also propose the use of a genetic algorithm to learn an appropriate weighting of the features.

2.2. Applications of AI

In the subsection 2.1, some definitions were introduced as well as some examples of use in a variety of fields. This subsection focuses on the main applications of AI which are closely related to mobile robotics and that have been developed during the past few years.

2.2.1. Self-driving navigation

Self-driving navigation means that a vehicle is able to plan its path and execute its plan without human intervention. This task is carried out through the use of data captured from sensors aboard the vehicle and sometimes through the use of remote navigation aids (Michelson, 2000). This application is quite common in vehicles such as cars and lorries but also for other kinds of ground, underwater or aerial robots. The interest in such applications has increased in recent years due to the desire to develop a full self-driving vehicle with the main aim of reducing the traffic accidents provoked by human beings. These systems basically consist of a mobile platform that integrates a set of sensors. The data collected by the sensors provide the perception of the environment. This information can be processed through AI algorithms with the aim of tackling the path-planning task to move through the environment with minimal human intervention. The autonomous navigation also takes into consideration tasks such as move-on-route, obstacle detection and avoidance and leader/follower capabilities.

Regarding the use of AI for this application, a considerable number of works have been developed in recent years. Li et al. (2017) introduce a fully autonomous navigation system for a smart microvehicle with a microscope-coupled CCD (Charge-Coupled Device) camera, an AI planner, and a magnetic field generator. The AI planner is split into three functional modules: a computer vision module for tracking the microvehicle and detecting obstacles in its environment; a motion planner to generate an optimal obstacle-free path between starting point and destination; and a magnetic motion controller to manipulate the microvehicle movement along a predesigned path. Polvara et al. (2018) propose collision detection and path planning methods for autonomous Unmanned Surface Vehicles by using artificial neural networks and evolutionary algorithms. Sharma et al. (2017) apply AI to secure wireless communications of Connected Vehicles, which facilitates exchange of safety messages for collision avoidance in self-driving cars. The AI system learns to augment its ability to discern and recognize its surroundings. Badue et al. (2019) carry out a survey about the state-of-the-art on self-driving cars focusing on works published since the birth of the DARPA (Defense Advanced Research Projects Agency) challenges. This survey focuses on the perception systems and the decision-making systems based on methods that make use of AI.

2.2.2. Face detection and recognition

Face Detection is the preliminary step for face recognition, and it consists basically in detecting faces in the images. These tasks have played an important role in robotics concerning problems such as surveillance (Ahuja et al., 2018) and home service robots (Jiang & Wang, 2017). According to M. Yang et al. (2002), the solution to the face detection problem can be divided in two steps:

1. Finding out whether there is any face in a given image or not.
2. If there is any face within the image, then, calculate where it is located.

In the related literature, many works can be found in this field. Romdhani et al. (2001) propose a face detector that is based on running an observation window at all possible positions and using a Support Vector Machine (SVM) to determine whether a face is contained within the window. Ahuja et al. (2018) propose Local Binary Patterns (LBP) to detect the ROI (Region Of Interest) of the face inside the image and Haar feature-based cascade classifiers for developing the face recognition. Nevertheless, the revolution in this task arrives when Viola & Jones (2004) introduced a real-time face detector, which is able to detect faces in real-time with high accuracy. This work is based basically on three contributions. The first is the introduction of a new image representation that allows quick computations. The second contribution is a simple and efficient classifier built through the AdaBoost learning algorithm to select a small number of critical visual features from a very large set of potential features (Freund & Schapire, 1995). The third contribution is a method that combines classifiers in cascade, allowing quickly background regions rejection and spending more computation on promising face-like regions.

Face detection frameworks are commonly proposed to identify multiple appearances in smartphone cameras like Hadid et al. (2007) explain. Nowadays, face detection is commonly used in any kind of storing system or social networks. For instance, Facebook, Google, etc. are using face detection in the images uploaded in social networks (Rabbath et al., 2012).

Second, face recognition is defined by Jafri & Arabnia (2009) as a system to verify the identity of a person among a set of identities by using as input a face image and a database of face images of known individuals . This task has attracted an enormous interest in automatic processing of digital images in order to solve a variety of applications such as biometric authentication or surveillance (Jain & Li, 2011). Face recognition has been proposed during the past few years as an identification system in the same way that fingerprint and iris were proposed before. According to Abate et al. (2007), face recognition systems fall into two categories: verification and identification. As for face verification, it is a one-to-one match that compares an image of a face, whose identity has to be recovered, against a template face. On the other hand, concerning face identification, it is a one-to-many problem that compares a candidate face image against all the image templates that are contained in a face database with the objective of determining the identity of the candidate face.

The face recognition task has been used in a high number of applications. For example, Kim (2005) proposes a security system that carries out automatic recognition for verification between the picture of the passport and the face of the individual; this work proposes a clustering algorithm that creates adaptive clusters to the variations of input patterns and it is applied to the extracted areas for the recognition. Regarding surveillance, CCTVs (Closed-circuit television) can be used to look for someone. Y. Wang et al. (2017) use face recognition in real-world surveillance. They propose a convolutional neural network which is trained with a labeled dataset and subsequently proposed to recognize individuals from the campus surveillance system.

2.2.3. Objects recognition and categorization

These tasks have played an important role in robotics concerning building object-based representations of the environment and manipulation of objects. Object recognition basically consists in detecting an object instance and object categorization consists in classifying a specific object (such as a cup of tea) (Loncomilla et al., 2016). For instance, H. Gao et al. (2018) propose an object classification method using RGB-D data to train a Convolutional Neural Network (CNN) with the objective of detecting and categorizing usual objects in an autonomous vehicle environment such as other cars, cyclists, pedestrians and trucks. Zhu et al. (2016) introduce a CNN to detect and classify traffic signs. Furthermore, there are several works that use this application to aditionally carry out a pose estimation of the objects detected. Kanezaki et al. (2018) use a CNN to categorize objects from multi-view images and estimating their position. Wei et al. (2018) developed an end-to-end Mask-CNN model that selects deep convolutional descriptors for fine-grained object recognition. Zaki et al. (2019) propose a multi-scale feature representation based on a convolutional hypercube pyramid (HP-CNN) that is able to carry out viewpoint invariant semantic object and scene categorization.

2.2.4. Objects manipulation

A manipulation planning or object manipulation is a task related to the motion planning, but the focus is not on the movement of the robot, but on the objects to be manipulated. This task consists basically in changing the position and/or the orientation of a specific object (or set of objects), while avoiding collisions or breaking the object/s (Jiménez, 2012). The interest for this application has increased substantially over the past few decades with the aim of replacing human workers in challenging (due to the required accuracy) or hazardous tasks, specially in industrial, health care and domestic environments (Smith et al., 2012).

In the bibliography, many works about manipulations and planning can be found which are sustained by AI tools. For instance, Boularias et al. (2015) introduce a robot system for grasping objects in dense clutter by using depth images. For this purpose, the robot learns to manipulate the objects by trial and error through a decision-making problem based on a reinforcement learning framework. Y. Yang et al. (2015) propose a system that learns manipulation action by processing videos from the internet by using two CNNs, one for classifying the hand grasp type and the other for object recognition. Matas et al. (2018) present a combination of state-of-the-art deep reinforcement learning algorithms to solve the problem of manipulating deformable objects.

2.3. Frameworks Commonly Proposed for Mapping and Localization

After presenting some definitions and the main applications of AI in the robotics field, the present subsection introduces some of the most popular AI tools used to address mapping and localization in mobile robotics.

2.3.1. Machine Learning Classifiers

Classification is a task that predicts the class or category which an ‘object’ belongs to. The object is also known as pattern and it is assumed to pertain to a unique class among a set of categories. Each pattern is represented by a set of measurements known as features, that must provide enough class-discriminatory information to predict the category of the pattern with high probability (Theodoridis, 2015). Usually, n feature variables, $x_1, ..., x_n$, are selected and arranged in a feature vector, $x \in \mathbb{R}^n$. The objective is to train a classifier whose function (or set of functions) $f(x)$ in $\mathbb{R}^n$ is able to predict the class which the pattern belongs to.

This technique has been widely used to solve a range of problems. For example, Atkinson & Campos (2016) propose a feature-based emotion recognition model by using a multi-class SVM with EEG-based Brain–Computer interfaces . Narudin et al. (2016) use different machine learning classifiers to detect malware in mobile phones using the anomaly-based approach. Concerning the computer vision field, there are many works that use classifiers. For instance, Korytkowski et al. (2016) introduce a fuzzy classifier with local image features to carry out objects classification. B. Zhang et al. (2015) propose an automatic defective apple detection method by using a weighted relevance vector machine (RVM) classifier. Aguilar et al. (2017) propose a pedestrian detector for UAVs (Unmanned Aerial Vehicles) based on a combination of Haar-LBP features with Adaboost and using cascade classifiers with Meanshift.

2.3.2. Clustering

Classifiers, as described in the previous subsection, is a supervised technique, that is, it needs correctly labeled data to carry out the training process. In this case, clustering is an unsupervised technique, where class labeling of the training patterns is not available. Hence, the main objective consists in finding out the organization of patterns into clusters (groups). To organize specific data into clusters, a clustering criterion or several clustering criteria must be established. Then, each pattern is categorized in a group and each cluster is characterized by the common attributes of the data that belong to it (Theodoridis & Koutroumbas, 1999).

This AI tool has been commonly proposed in a wide range of problems related to robotics and computer vision. For example, Dhanachandra et al. (2015) propose an image segmentation using k-means clustering. Schroff et al. (2015) propose a clustering and face recognition approach based on a system that directly learns a measure of face similarity. Fan et al. (2018) propose a progressive unsupervised learning method based on pedestrian clustering and fine-tuning of a CNN to transfer pre-trained deep representations to unseen domains. C. Wang et al. (2019) propose an improvement of 3D object recognition by introducing a view clustering and pooling layer based on dominant sets.

2.3.3. Deep Feedforward Networks

Deep feedforward networks, also known as feedforward neural networks or multilayer perceptrons (MLPs), are deep learning models whose objective is to approximate some function f^*. This network defines a mapping $y = f(x; \phi)$ where x and y are the input and output (or target) data respectively. This network learns the value of the parameters ϕ that best approximate the function f (Goodfellow et al., 2016). These models carry out a flow through the function f evaluating from x to the output y. Nevertheless, these models do not provide feedback connections, that is, outputs of the model are not fed back into the model itself. During training, the aim is to drive $f(x)$ to match $f^*(x)$: the training data are provided and $f^*(x)$ is evaluated with those data. Moreover, a label y is included with each example x to achieve $y \approx f^*(x)$.

These networks present an extreme importance in machine learning, since they are the basis of many applications. For example, many object recognition approaches are based on this kind of models, such as the work by Mostajabi et al. (2015), who propose a feed-forward architecture for semantic segmentation to tackle a rich feature representation that is used for object recognition.

2.3.4. Autoencoders

An autoencoder is a neural network architecture composed basically of an encoder and a decoder system whose aim is to find a compressed representation of the given input data. The process consists in finding a representation or code to carry out useful transformations on the input data. Traditionally, autoencoders were proposed for dimensionality reduction or even for feature learning (**?**). Denoising autoencoders try to find a code that can convert noisy data into clean ones. Moreover, autoencoders are also used to perform colorization, feature-level arithmetic, detection, tracking, and segmentation among others. As shown in fig. 1, regarding the encoder, it transforms the input data x into a low-dimensional latent representation $h = f(x)$. This latent representation is a vector of lower dimension. The encoder learns to extract the most important features of the input data. As for the decoder, it recovers the input data from the latent representation, $r = g(h)$ with the objective that $g(f(x)) = r$, being r as close as possible to x. In general the encoder and decoder are non-linear functions and the dimension of the latent representation h is considerably smaller than the input dimensions. Similar to other neural networks, the autoencoder tries to minimize a loss function during the training process. The loss function is established to measure how dissimilar the input x and the reconstructed input r are. For example, the Mean Squared Error (MSE) can be used for this purpose.

$$L(x, r) = MSE = \frac{1}{m} \sum_{i=1}^{i=m} (x_i - r_i) \tag{1}$$

where m the number of components of the input and the output ($m = width \times height \times channels$).

Autoencoders have been a successful tool for dimensionality reduction and also for information retrieval. Regarding dimensionality reduction, this tool has provided reconstructions with an error rate lower than using other techniques such as PCA (Principal Components Analysis) (Hinton & Salakhutdinov, 2006b). Therefore, through the improvement of lower-dimensional representations, other related tasks have also been improved. First, in classification tasks, autoencoders provide a model with less memory requirements and computing time consumption (X. Ma et al., 2016). Second, by means of dimensionality reduction, information retrieval can be carried out more efficiently. With the use of autoencoders and its related dimensionality reduction, exhaustive searching becomes more efficient. For example, Pfeiffer et al. (2018) present a study about learning-to-rank and query refinement approaches for information retrieval in the pharmacogenomic domain. Zhu et al. (2016) propose using an autoencoder for feature learning from 2D images with the aim of carrying out 3D shape retrieval. Moreover, autoencoders have been widely proposed to produce codifications that are low-dimensional and binary. In this way, entries of a database can be stored in a hash table and information retrieval can be carried out by returning all the entries that have the same binary code as the query. To cite one example of this approach, Carreira et al. (2015) introduce a fast search in image databases with binary hashing, where each high-dimensional, real-valued image is mapped with an autoencoder onto a low-dimensional, binary vector and the search is done in this binary space. Apart from these examples, many others can be found in the related literature concerning the use of autoencoders for mobile robotics. Sergeant et al. (2015) address a navigation task by using a deep autoencoder which learns how to navigate from the sensory data stored in a dataset. H. Wang et al. (2018) propose an autoencoder for fusion and extraction of multiple visual features from different sensors with the aim of carrying out motion planning based on deep reinforcement learning.

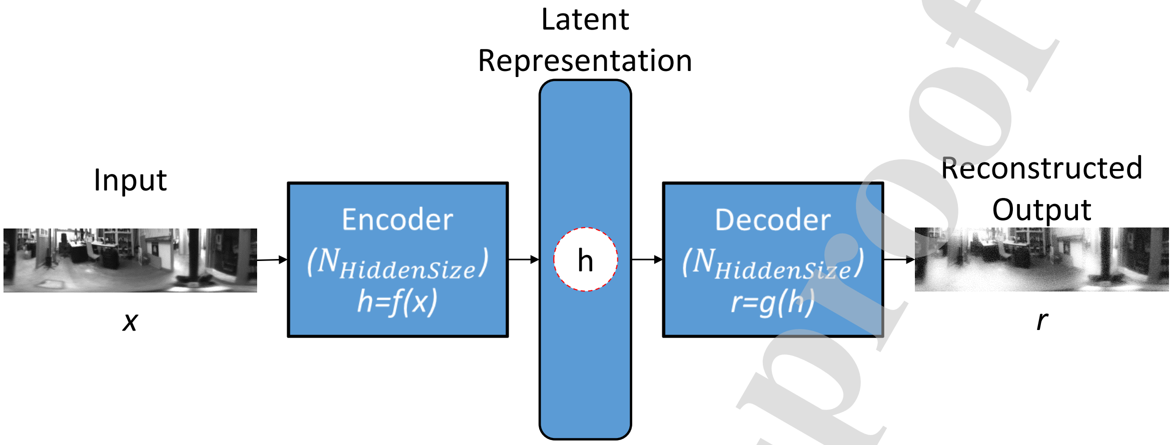

Figure 1: Autoencoder structure. An input x is mapped to obtain a reconstruction r by means of using a latent representation h. f is a function that encodes or maps x to h and g is a function that decodes, that is, maps h to r.

2.3.5. Convolutional Neural Networks

Convolutional Neural Networks, commonly known as CNNs, are currently the most popular tool among the deep learning techniques, since they have led to successful results in many practical applications. They are a specialized kind of neural network for processing data that present an already known topology. These networks are commonly designed to receive images as input and they have different applications such as classification or objects detection. This kind of networks are based on the use of convolutions, which is a specialized kind of linear mathematical operation (Goodfellow et al., 2016). That means, whereas traditional neural networks use matrix multiplication with a separate parameter that describes the interaction between inputs and outputs, CNNs present sparse interactions, i.e., using specific and meaningful features obtained from the input data. CNNs consist of local connections between neurons and hierarchically organized transformations of the data. Basically, CNNs are composed by three types of neural layers: convolutional layers, pooling layers and fully connected layers. Every layer transforms the input and generates an output according to the parameters established. This process is tackled throughout several layers until reaching the last layer, which is a fully connected layer that outputs a 1D feature vector, which provides the most likely prediction.

There are very well known CNNs whose architectures have been used as starting point to develop new computer vision tasks. For instance, AlexNet was introduced by Krizhevsky et al. (2012). This network consists of eight layers (five convolutional layers and three fully connected layers) with a final 1000-way softmax and three pooling layers. The input image has a size of 227×227×3 and the network was trained to identify objects in the input images. It is able to identify 1000 object categories, such as keyboard, pencil, and a variety of animals. Fig. 2 shows the architecture of this network. GoogLeNet was proposed by Szegedy et al. (2015). This network has 22 layers, it is also trained for object classification but it uses 12 times fewer parameters than AlexNet. A wide review of the most outstanding CNNs can be found in Pak & Kim (2017). Moreover, table 1 shows a summary table of the most popular CNNs until the present date.

Additionally, there are other options that permit reusing robust CNNs which have provided successful results, to solve different problems from the input images. On the one hand, the **transfer learning** technique consists in the process of retraining a pre-trained network to classify a new set of images, that is, reusing the architecture, weights and parameters of a CNN which already works properly as starting point to build a new CNN with a different purpose. The main idea is to get profit of most of the intermediate layers, because their parameters have been tuned with a large number of images. The problem, then, is reduced to changing the final layers (in order to re-adapt them to the new task proposed) and, perhaps, the initial layers (if the size of the images does not match the size used previously). Once the "new" network architecture is established, the training process starts through using the new labelled training data. Hence, this technique can save a considerable amount of time for training and even output better results than creating a new network from scratch. This idea has been used by many authors. For example, Han et al. (2018) use CNN transfer learning to-

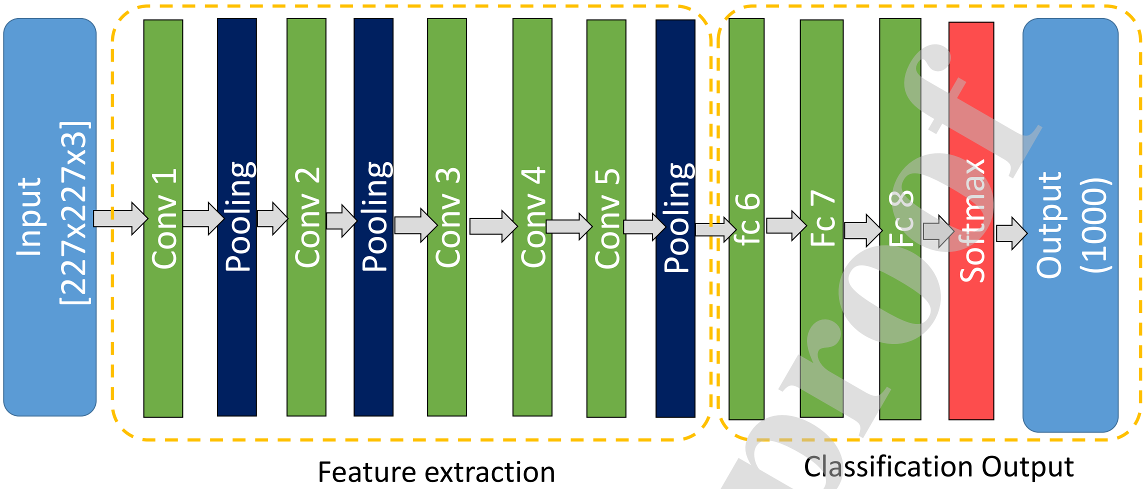

Figure 2: Architecture of the CNN AlexNet. Input images have a size of $227 \times 227 \times 3$ and the output is able to classify objects into 1000 categories.

Table 1
Summary of the most popular CNNs developed during the past few years.

CNN	Year	Developed by	No. of convolutional layers	No. of parameters
LeNet	1998	LeCun et al. (1998)	5	60,000
AlexNet	2012	Krizhevsky et al. (2012)	8	60 million
GoogLeNet	2014	Szegedy et al. (2015) (Google company)	22	4 million
VGG Net	2014	Simonyan & Zisserman (2014)	19	138 million
Inception	2015	Szegedy et al. (2015)	65	5 million
ResNet	2016	He et al. (2016)	152	25.6 million
Xception	2017	Chollet (2017)	42	23 million

gether with data augmentation in order to achieve good solutions despite the small size of the datasets used. Also, as mentioned previously, Wozniak et al. (2018) use the transfer learning technique to retrain the VGG-F network to classify places among 16 rooms acquired by a humanoid robot. On the other hand, many authors have also proposed the use of intermediate layers to generate global-appearance descriptors of the input image. In this sense, once the network is properly available to face the desired task, the hidden layers perform vector description which can be used to characterize the input data. This idea has been exploited by some authors such as Arroyo et al. (2016), who use a CNN that automatically learns to generate visual descriptors which are robust against changes of seasons, in order to carry out a robust topological localization. Wozniak et al. (2018) also use the features extracted from the FC-6 layer to train a linear SVM (Singular Vector Machine) classifier. Mancini et al. (2017) use this visual information to carry out place categorization with a Naïve Bayes classifier.

Regarding the use of CNNs to solve robotics tasks through visual information, there are many works that have provided successful results by using this technique. For instance, Sinha et al. (2018) propose a CNN to process data from a monocular camera and tackle an accurate robot relocalization in GPS-denied indoor and outdoor environments. Payá et al. (2018) propose using CNN-based descriptors to create hierarchical visual models for mobile robot localization. More recently, Chaves et al. (2019) propose a CNN to build a semantic map. Concretely, they use the network to detect objects in images and, after that, the results are placed within a geometric map of the environment. Xu et al. (2019) propose a multi-sensor-based indoor global localization system integrating visual localization aided by CNN-based image retrieval with a Monte Carlo probabilistic localization approach.

2.3.6. *Regression Neural Networks*

Apart from the main techniques based on deep learning showed previously, there are other options that have been used by researchers. Despite its use is less common to solve mobile robotics tasks, they have also provided successful results.

Regression Neural Networks (R-CNN) are among these deep learning techniques. Deep neural networks are well known for classification problems, where the goal is to predict a single discrete label of an input vector. Nevertheless, the regression problem consists in obtaining a continuous value instead. Therefore, this type of network has commonly been proposed for continuous predictions such as forecasting. Bilgili & Sahin (2010) propose an analysis of regression neural network models to predict wind speed; and Kumar et al. (2015) introduce a study about regression neural networks to estimate the monthly average global solar radiation. These models have also been proposed for other types of predictions such as medical diagnoses. For instance, Kayaer & Yıldırım (2003) propose using a general regression neural network to diagnose diabetes. Ferreira et al. (2004) use a general regression neural network to construct the base of an adaptive neuro-fuzzy system and carry out a walking control of an autonomous biped robot. As for mobile robotics, the related literature also presents different examples of application. For example, K. Wang et al. (2007) use a general regression neural network for approximating the functional relationship between high-dimensional map features and states of the robot. Rahman et al. (2012) propose a location estimation algorithm using generalized regression neural network and Wireless Sensor Network (WSN). Dezfoulian et al. (2013) propose a method to interpret the data from various types of 2-dimensional range sensors and a regression neural network to perform the navigation task.

2.4. Recurrent Neural Networks

Recurrent Neural Networks (RNNs) are a kind of neural network specialized in processing a sequence of values. This type of networks share parameters in a different way. Each member of the output is a function of the previous outputs and the connections between nodes form a directed graph along a temporal sequence. RNNs can use their internal state to process sequences of inputs and therefore they exhibit temporal dynamic behavior. RNNs are flexible in their use of context information, because they can learn what to store and what to ignore and they can recognize sequential patterns in the presence of distortions (Sak et al., 2014). More detailed information about this type of networks can be found in the work developed by Graves (2012). An example of RNN is the Long short-term memory (LSTM), which has feedback connections and can process entire sequences of data. For instance, LSTM is commonly applicable to tasks such as unsegmented, connected handwriting recognition, as Messina & Louradour (2015) do to recognize lines of handwritten Chinese text, or speech recognition such as Graves et al. (2013) do. Additionally, this tool has also been proposed to solve mobile robotics tasks. For example, Otte et al. (2016) introduce an extension of Long Short Term Memories (LSTMs) for ground robots based on vibration data classification with the aim of carrying out recognition of the ground reliably condition for mobile robot navigation. Rahmatizadeh et al. (2016) carry out a deep learning controller based on LSTM with the aim of learning manipulation tasks for assistance robotics and wheelchair mobile robots. Otte et al. (2016) propose an extension of LSTMs for classification of 14 different ground types based on vibration data, since recognizing the condition of the ground may be key in mobile robot navigation systems. Sun et al. (2018) present a 3-DOF pedestrian trajectory prediction approach for autonomous mobile robots by means of range-finder sensors with an LSTM network.

2.5. Deep Reinforcement Learning

Reinforcement learning is a branch of machine learning that has gained a lot of attention since it was proposed to play Atari games (Mnih et al., 2013). In reinforcement learning, an autonomous *agent* receives information from the environment and takes actions to maximize a notion of cumulative reward (Chollet, 2017). Deep reinforcement learning consists in the use of deep learning and reinforcement learning principles with the aim of creating efficient algorithms. This field of research has been able to solve complex decision making tasks that were hard to solve by means of conventional methods. François-Lavet et al. (2018) introduce this deep learning model and focus on the aspects related to generalization and how deep reinforcement learning can be used for practical applications. Despite the related algorithms have been scarcely applied to solve real situation tasks, the state of the art already presents some examples, such as Lillicrap et al. (2015), who introduce a deep learning refoircement algorithm that solves more than 20 simulated physics tasks, including classic problems such as cartpole swing-up, dexterous manipulation, legged locomotion and car driving. As for the mobile robotics tasks, J. Zhang et al. (2017) propose a successor-feature-based deep reinforcement learning algorithm that can learn to transfer knowledge from previously mastered navigation tasks to new problem instances. Tai et al. (2017b) propose a learning-based mapless motion planner based on an asynchronous deep reinforcement learning method to estimate the target steering commands with respect to the mobile robot coordinate frame. Zhu et al. (2017) introduce a target-driven visual navigation system in indoor scenes by using deep reinforcement learning with the aim of improving the lack of capability of generalizing new goals and the data inefficiency. Kahn et al. (2018) carry out a self-supervised deep reinforcement learning for robot navigation to improve the need to learn complex policies from the environment with few samples. Their robotic system captures raw monocular images and it is able to tackle the navigation task by means of learning by a fully autonomous reinforcement learning.

3. Description of the Visual Information by Using AI Tools

Vision sensors have been widely used for mobile robotics purposes. However, images are highly dimensional data and they also change for several reasons apart from the movement of the robot, such as change of illumination or position of some objects that constitute the environment. Hence, the approaches that work with these data consist commonly in extracting the most relevant and invariant information from

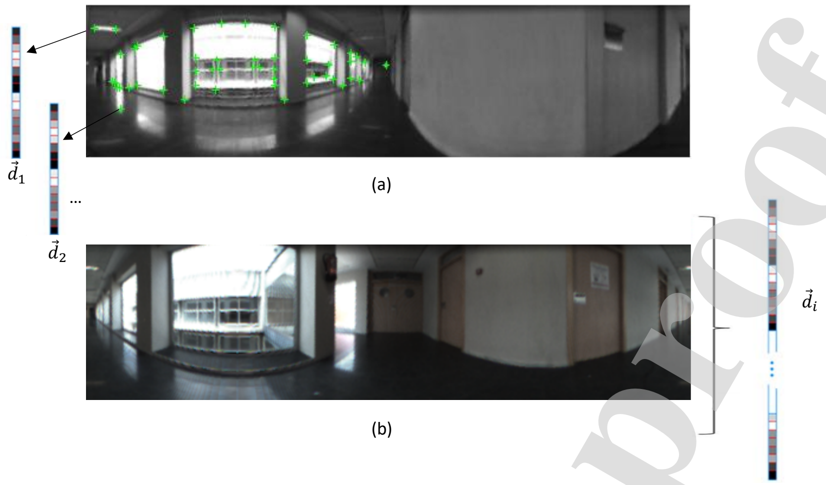

Figure 3: Two main methods to extract the most relevant information from the images for mapping and localization purposes. (a) Detection, description and tracking of some relevant landmarks along a set of scenes. (b) Building a unique descriptor per image that contains information on its global-appearance.

scenes. In this sense, two main approaches have been commonly proposed; either by detection and description of local features, or working with global-appearance extraction methods. On the one hand, the methods based on local features consist in extracting some outstanding points from each scene and creating a descriptor for each point, using the information around it (fig. 3(a)). On the other hand, global-appearance description methods consist in building a unique descriptor per image. Figure 3 illustrates (a) local features extraction and description and (b) global-appearance description.

In the literature, many examples can be found using local features as well as global-appearance descriptors to solve mobile robotics tasks. To cite some examples, concerning local features, Kunii et al. (2017) propose a robust landmark tracking method for mobile robot operation in natural environments, where ORB (Oriented FAST and rotated BRIEF), CenSurE (Center Surround Extremas) are used for feature extraction and SURF (Speeded-Up Robust Features), ORB, FREAK (Fast Retina Keypoint) for feature description. Su et al. (2017) propose a global localization approach with the capability of addressing the kidnapped robot problem, where the ORB local descriptor is used to further improve localization accuracy. Regarding global-appearance descriptors, Payá et al. (2016) present a comparative analysis of some global-appearance descriptors for mapping. Rituerto et al. (2014) propose the use of the *gist* (Oliva & Torralba, 2001) descriptor to build topological maps departing from omnidirectional images. A. C. Murillo et al. (2013) use a panoramic gist descriptor to address the localization task in urban environments. More recently, Faessler et al. (2016) present a vision-based quadrotor system to map a dense three-dimensional area. Korrapati & Mezouar (2017) propose the use of omnidirectional images through global appearance descriptors to build topological maps and also a loop closure detection method. Both local features and global-appearance descriptors have been commonly calculated by means of analytical methods. Traditionally, initial works in mobile robotics tried to extract and describe local features from the scenes. Later, a number of works proposed using the information as a whole, creating a holistic descriptor per scene. More recently, the development of new AI techniques and the evolution of the computing devices has made it possible to extract relevant information by means of such AI techniques. Fig. 4 shows this evolution of methods to extract relevant information from the scenes in mobile robotics, and includes a number of relevant works that make use of each approach. In the next subsections, some of the most popular methods are detailed.

3.1. Local Features

Since the emergence of SIFT (Scale-Invariant Feature Transform) (Lowe, 2004), local features have played an important role in image matching, for example, to solve the image retrieval problem (Zheng et al., 2017; Se et al., 2005). Nevertheless, local features have not been used only for image retrieval, but they have also been proposed as a powerful tool for other computer vision problems such as wide baseline stereo matching or object detection. Typically, methods based on local features comprise two main stages: extracting a set of outstanding points, objects or regions from each scene and creating a descriptor for each. That means that every feature is described by means of a data vector, which is typically invariant against changes in the position and orientation of the camera. Once the extraction and description of the features has been addressed, they are usually tracked and matched along a set of scenes.

A considerable amount of local features extraction and description methods have been developed since the appearance of SIFT. Many subsequent developments focused on reducing its computational requirements or improving the invariability to other effects. For example, SURF (Bay et al., 2008) presents lower computational cost and higher robustness against image transformation and BRIEF (Binary Robust Independent Elementary Features) is designed to be used in real time at expense of a lower tolerance to image distortion and transformations (Calonder et al., 2010). All these examples are known as traditional or hand-crafted features, since they are based on the detection of visual structures such as corners. A deep survey of these tools can be found in Payá et al. (2017) and Mukherjee et al. (2015) carried out an exhaustive comparative experimental study.

Concerning the development of local features based on AI techniques, they are widely known as learned features and like hand-crafted ones, the AI methods typically consist in either detecting or describing local features, or even both (detecting and describing). Within the learned local features, two main blocks can be established: features based on machine learning and based on deep learning techniques. As for the first block, FAST (Features From Accelerated Segment Test) was one of the first successful methods and it is designed for high-speed corners detection (Rosten & Drummond, 2006). Despite being principally constructed for speed purposes, this method also proved to outperform existing corner detectors. Later, simulated annealing was proposed to optimize the parameters of the FAST detector to achieve higher repeatability (Rosten et al., 2008). This work shows that using machine learning produces significant improvements in repeatability, speed and quality. Early attempts were based on genetic algorithms. In this sense, Trujillo & Olague (2006) present an approach for extracting automatically low-level features by applying genetic programming. These authors introduce a Genetic Programming implementation that is capable of discovering a modified version of a feature operator which presents an improved performance. This work also highlights the balance between genetic programming and domain knowledge expertise to obtain results that improve hand-crafted solutions. More recently, machine learning tools have been typically used in feature detection to imitate and/or accelerate previously defined methods. Šochman & Matas (2009) propose a faster version of binary decision algorithms by using a WaldBoost classifier . This classifier learns to minimize the decision time of the classifier while guaranteeing predefined precision. Holzer et al. (2012) address the Interest Point (IP) detection as a regression problem by using machine learning. A regression forest (RF) model learns to detect if there is an IP in the center of a given image patch. Other researchers use machine learning to reduce the size of the descriptor, such as Strecha et al. (2011), who propose metric learning to reduce the size of the descriptors by representing them as short binary strings. In short, they map the descriptor vectors into the Hamming space, which is used to compare the resulting representations. This way, the size of the descriptors is reduced by representing them as short binary strings. Simonyan & Zisserman (2014) develop a learned local feature descriptor by using convex optimization. This work shows that learning the pooling regions for the descriptor can be formulated as a convex optimization problem. It also shows a descriptor dimensionality reduction by using Mahalanobis matrix nuclear norm regularization. Both formulations are based on discriminative large margin learning constraints.

Regarding the learned features based on deep learning techniques, they have been often used to improve rather than to replace hand-crafted local features. For instance, they have been used to learn covariant feature detectors invariant against viewpoint changes without supervision. For example, Lenc & Vedaldi (2016) propose a general machine learning formulation for covariant feature detectors. Moreover, many other improvements can be done, such as including explicitly modelling detection confidence, predicting multiple features in a patch, or jointly training detectors and descriptors. Mishkin et al. (2018) introduce a method for learning local affine-covariant regions. The proposed affine shape estimator is trained considering the loss function, descriptor type, geometric parametrization, etc. Furthemore, the training process does not require aligned patches geometrically accurate.

Most of the related works are focused on feature descriptors, nevertheless, more recently, there has been also an important progress in the development of detectors. For example, Verdie et al. (2015) use deep neural networks to learn a feature detector robust against illumination changes. The process consists in firstly identifying good keypoint candidates in multiple training images and secondly training a regressor to predict a score map whose maxima are those points. Yi et al. (2016a) propose an end-to-end framework based on the use of a deep network architecture to detect keypoints, estimate orientation and compute descriptors. These authors also developed a work to train a CNN to estimate the canonical orientation of a local feature given an image patch centered on the feature point and extended it to several different description methods (Yi et al., 2016b). They propose siamese networks to avoid the task of finding a target orientation to learn. Furthermore, they also propose a new activation function. Concerning CNNs to obtain local features, its use has been widely proposed by several authors. This technique is specially successful for large-scale image retrieval applications. For instance, Noh et al. (2016) propose a CNN-based local feature that is trained for instance-level recognition tasks without the need of object and patch-level annotations and it is suitable to replace hand-crafted descriptors. This framework can be used in image retrieval problems, enabling more accurate feature matching and geometric verification.

Nonetheless, despite the wide use of deep convolutional networks, local viewpoint invariant features based on hand-crafted techniques still play an important role in applications such as motion and image retrieval. Recently, Lenc & Vedaldi (2018) have carried out a deep evaluation of local feature detectors by evaluating a range of state-of-the-art local feature detectors. Through this study, they concluded that machine-learning-based detectors help to improve illumination invariance, that traditional methods are still competitive and they suggest also that a significant progress regarding deep-learning-based detectors can be done.

3.2. Global-Appearance Description

The approach based on global-appearance or holistic descriptors consists on working with the image as a whole, i.e., without extracting any local information. For this type of approaches, each image is represented by a unique descriptor that contains information on its global appearance (Payá et al., 2017). Concerning mobile robotics, this description method presents advantages in dynamic and poorly structured environments, where extracting stable local features may result difficult. Additionally, due to the fact that each image is represented by a unique descriptor, global-appearance descriptors lead to simpler mapping and localization algorithms (Amorós et al., 2018; Berenguer et al., 2019; Cebollada et al., 2019a; Cebollada et al., 2019b).

A wide range of works have been proposed during the past few years to develop holistic descriptors by using AI techniques. This method is known by some authors as feature engineering and it tries to take advantage of human prior knowledge to compensate the weaknesses that may present the algorithms (Bengio et al., 2013). One of the main objectives of developing methods to learn descriptors is to achieve faster solutions to proposed AI problems. Furthermore, AI applications have proved to be able to understand the environment that surrounds the camera, thank to their capability of identifying interesting and rejecting unprofitable information from the sensory data. Remarking the global-appearance descriptors based on deep architectures, they are usually effective to train robust models and introduce two advantages in this topic: first, deep architectures promote the reuse of features and second, they lead to more abstract features in higher layers that are typically invariant to local changes. A profound review about a wide range of unsupervised feature learning techniques can be found in the work presented by Bengio et al. (2013).

Among the early techniques proposed, **PCA** was one of the first alternatives that presented robustness. PCA basically performs a linear transformation $h = f(x) = W^T x + b$ of the input $x \in \mathbb{R}^n$ and the results are d_h features that are the first components of the representation h (Kirby, 2000). Similar to PCA, Independent Component Analysis (**ICA**) performs a linear analysis to obtain distinctive features based on linear generative models with non-Gaussian independent variables. Like sparse coding, ICA and its variants have also been used to obtain nonlinear features such as in the works developed by Bell & Sejnowski (1997); Jutten & Herault (1991); Le et al. (2011).

Successful feature learning algorithms and related applications are used in many works using a variety of approaches such as **RBMs** (Restricted Boltzmann Machines). For example, Hinton et al. (2006a) propose a technique that consists in stacking pre-trained RBMs into deep belief networks (DBN), where the top layer is interpreted as an RBM and the lower layers as a directed sigmoid belief network. This work has proved to give better digit classification than discriminative learning algorithms. Salakhutdinov & Hinton (2009) propose to combine RBM parameters into DBM (Deep Boltzmann Machines) by halving the RBM weights to obtain the DBM weights and train it by approximate maximum likelihood. This way, this work shows that DBM learn good generative models and perform well on handwritten digit and visual object recognition tasks. Larochelle et al. (2009) carry out an empirical study about the use of different RBM input unit distributions. This study confirms the hypothesis that the greedy layer-wise unsupervised training strategy improves the optimization by initializing weights in a region near a good local minimum and it also brings better generalization of the input. Another important perspective on global-appearance descriptors is based on the **manifold learning**, a geometric notion whose premise is based on the concentration of high-dimensional input space in the vicinity of a manifold M of lower dimensionality. The majority of the methods based on this technique lead to a non-parametric approach based on neighbor graphs. Belkin & Niyogi (2003) propose a geometrical algorithm for representing the high-dimensional data that provides a computationally efficient reduction of dimensionality. This reduction has locality-preserving properties and a natural connection to clustering. Donoho & Grimes (2003) propose a Hessian-based locally linear embedding method for recovering the underlying parametrization of scattered data. Weinberger & Saul (2006) introduce an algorithm for unsupervised learning of image manifolds by semidefinite programming. The algorithm computes a low dimensional representation of each image so that distances between nearby images are preserved. “Accelerating t-SNE using tree-based algorithms, author=van der Maaten, L.” (2014) proposes variants of the Barnes-Hut algorithm with the t-SNE (t-distributed Stochastic Neighbor Embedding) algorithm to learn embeddings of data sets with millions of objects. More recently, some authors have proposed to use free energy functions, that is, without explicit latent variables. For instance, Ngiam et al. (2011) use a hybrid of Monte Carlo to train the free energy function. In brief, they propose using deep feedforward neural networks to model the energy landscapes that define probabilistic models. The lower layers of the model adapt the training of the higher layers, and thereby this produces better generative models. By means of this method, all the layers of the model are simultaneously and efficiently trained. Kingma & Cun (2010) propose denoising score matching. Differentiating the loss with respect to the model parameters is automated with an extended version of a double-backpropagation algorithm.

Apart from the previously mentioned techniques, it is worth remarking the use of deep neural networks, specially **CNNs**, to obtain holistic descriptors, since a number studies have proved that these networks can learn more transferable features for domain adaptation and produce successful results in a wide range of scenarios and applications (Glorot et al., 2011; Yosinski et al., 2014). For instance, Donahue et al. (2014) propose the use of features extracted from the activation of a deep convolutional network trained in a fully supervised fashion on a large, fixed set of object recognition tasks and use it for a completely different task. The works focuses on investigating the semantic clustering of deep convolutional features with respect to a variety of tasks such as scene recognition, domain adaptation, and fine-grained recognition. The study addresses an efficacy comparison relying on various network levels to define a fixed feature. Yosinski et al. (2014) carry out a deep study about how transferable features are in deep neural networks. They conclude that features obtained from the initial layers appear not to be specific to a particular dataset or task and the features become more specific as the selected layer approaches to the last one. Additionally, they conclude that initializing a network with transferred features from almost any number of layers can produce a boost to generalization. Long et al. (2015) propose a Deep Adaption Network (DAN) architecture that generalizes deep CNNs to the domain adaptation scenario. The DAN architecture learns transferable features and can scale linearly by unbiased estimate of kernel embedding. Arandjelovic et al. (2016) solve the image retrieval problem by developing a CNN and using it to obtain global-appearance descriptors. This network incorporates a new layer which is inspired by the "Vector of Locally Aggregated Descriptors" (VLAD) image representation. VLAD is commonly used in image retrieval tasks. Gordo et al. (2016) introduce a method that employs a region proposal network to learn which regions should be pooled to form the final global descriptor. This approach produces a global image representation in a single forward pass. Most recently, Xu et al. (2019) propose a transfer learning method based on a pre-trained model to transform general features into special features, which are adapted to the desired task. The pre-trained model of Faster R-CNN is used to extract the high dimensional convolution features of images.

Another technique widely used in recent years within the deep learning has been the use of **autoencoders** and variations of this tool. As outlined in subsection 2.3.4, in the autoencoder frameworks, the starting point is a feature-extracting function in a specific parameterized closed form. This function, $f(x)$, is the encoder and tackles the straightforward computation of a feature vector from an input x through $h = f(x)$. So, for each sample of the dataset $X = \{x_1, ..., x_N\}$, we define $h(i) = f(x_i)$, where $h(i)$ is the feature vector computed from x_i. For example, Vincent et al. (2010) propose a denoising autoencoder (trained to denoise corrupted versions of the inputs) to obtain a robust global-appearance descriptor, which is used successfully to solve classification problems. This work shows that the denoising autoencoders are able to learn Gabor-like edge detectors from natural image patches. The descriptor generated permits performing classification task similar than using deep belief networks. Coates et al. (2011) carried out a detailed analysis of the effect of changes in the model setup (receptive field size, number of hidden nodes, the step-size) and compare the extracted features (sparse autoencoders, sparse RBMs, K-means clustering, and Gaussian mixtures) with a whitening process. They conclude that complex algorithms can have greater representational power and simple but fast algorithms can be highly competitive. Le (2013) trains a deep sparse autoencoder on a large dataset of images to build high-level features. The experiments show that this feature detector is robust against translation, scaling and out-of-plane rotation. X. Gao & Zhang (2017) propose an approach based on the Stacked Denoising Autoencoder (SDA) to detect loops for a visual SLAM system. The descriptors are calculated by using SDA over patches of the original images. This autoencoder can lead to very complicated structures, since the learned features reflect the inner patterns of the data, whereas traditional hand-crafted feature descriptors are usually not able to show that.

Within this subsection, it is worth to mention the **Bag of local Features (BoF)** method, which can be considered as a blended method between local features and global-appearance descriptors. This method comes from Bag of Words (BoW) representation, which basically consists of a model for representing text data with machine learning algorithms. The bag-of-words model is simple to understand and implement and has seen great success in problems such as language modeling and document classification. Concerning BoF, many approaches have been developed in computer vision during the last few decades (Csurka et al., 2004; Jurie & Triggs, 2005; Lazebnik et al., 2006; J. Zhang et al., 2007). Bag of Features methods have been applied to image classification, object detection, image retrieval, and even visual localization for robots. BoF approaches consists basically in a characterization based on the use of a collection of image local features to form a vector that characterizes the input image. Despite the lack of structure or spatial information, this image representation can be good enough for many state-of-the-art applications. A detailed explanation of this technique can be found in O'Hara & Draper (2011).

Like local features and global-appearance descriptors, many works have proposed the use of AI techniques in bag-of-features frameworks. In the majority of these works, AI (specially deep learning) is used to replace previously hand-crafted features extraction methods in BoF. For example, Gong et al. (2014) introduce a multi-scale orderless pooling (MOP-CNN), which extracts CNN activations for local patches at multiple scale levels, performs orderless VLAD pooling of these activations at each level separately, and concatenates the result. The proposed method can be used as a generic feature for either supervised or unsupervised recognition tasks. Ng et al. (2015) present an approach for extracting convolutional features from different layers of the networks OxfordNet and GoogLeNet, and adopt VLAD encoding to encode features into a single vector for each image. Mohedano et al. (2016) propose an image retrieval pipeline based on encoding the convolutional features to obtain BoF by assigning each local array of activations in a convolutional layer to a visual word. Feng et al. (2017) propose CNNs for optimizing the feature extraction to perform BoF for geographical scene classification. J. Cao et al. (2017) build an effective BoF model using deep local features. They show how to use the CNN as a combination of local feature detector and extractor without the need of feeding multiple image patches to the network. Khan et al. (2018) propose two strategies to encode multi-scale information explicitly during the image encoding stage. The aim of this approach is to recognize human actions. The first approach is based on a multi-scale image representation with scale encoded with respect to the image size. The second approach, instead, encodes feature scale relative to the size of the bounding box corresponding to the person instance. Scale coding of bag of deep features is performed by applying the coding strategies to the convolutional features from the pre-trained VGG-19 network. Brendel & Bethge (2019) implement a variant of the ResNet-50 CNN that classifies images by using BoF as input and compare their method with other high-performance deep neural networks (VGG-16, ResNet-50 and DenseNet-169) to carry out the classification task with the ImageNet dataset.

From this section, the conclusions achieved are that the use of techniques based on AI present some advantages: compared to other description methods, they lead to semantic, objects and geometric forms interpretation, and once trained, the model is easy and quick to use and obtain the necessary descriptors. Nonetheless, these descriptors are not based on a closed mathematical process and the method to obtain such description is not known *a priori*; it depends on the parameters configuration during the training process. Therefore, the modelling will be sensitive to the specific training dataset and it is expected to work well under similar environments, but it may lead to less robust descriptors under different circumstances, what may limit further results.

About the performance of AI and deep learning tools in the process of extracting information from the scenes, as detailed in the previous paragraph, it typically depends on a training process which requires a high number of data vectors. Therefore, this process con be computationally expensive and require large computational resources and a long period of time to learn how to extract significant data from the set of training images. However, once the model has been trained, using this model to extract information from new images (as the robot performs a task) is a relatively fast process. Quantitative data about these processes can be found in the works by Cebollada et al. (2019c) and Cebollada et al. (2020).

4. Mobile Robotics Tasks Using Vision and AI

This section presents a review of recent works related to the mapping, localization, navigation, SLAM and exploration tasks in robotics using visual information and AI tools.

4.1. Map Building

Mapping consists basically in creating a map, model or representation of the environment using the data provided by the sensors mounted on the robot. Such models are useful to solve, subsequently, other tasks, such as localization, path planning or navigation. These problems can be solved by comparing the information provided by the sensors of the robot with the model. Thrun (2002) presented an exhaustive explanation to the robotic mapping concept. In the related literature, two main frameworks have been proposed in order to carry out this task: the metric maps, which represent the environment with geometric accuracy; and the topological maps which describe the environment as a graph containing a set of locations with the related links among them. For example, Tanzmeister et al. (2014) propose an approach that estimates a uniform, low-level, grid-based world model including dynamic and static objects. da Silva et al. (2018) propose a localization and navigation approach for mobile robots using topological maps and using a CNN to obtain descriptors from omnidirectional images. Apart from these options, arranging the information hierarchically constitutes an efficient alternative. This framework consists in creating a map which is composed of several layers with a hierarchical structure. The high-level layers contain a relatively compact amount of information, which permits a rough but quick localization. The low-level layers have usually more information and are used to refine the position. Kuipers et al. (2004) propose a hierarchical hybrid map, which consists in using a metrical approach to build local maps of small-scale space and topological maps to represent the structure of large-scale space. This approach is proposed to solve the SLAM task in an environment with multiple nested large-scale loops. Cebollada et al. (2019a) propose a study about clustering methods to carry out efficiently the data compaction of metric and topological maps based on omnidirectional images with the aim of building hierarchical maps. They have also tested the robustness of the hierarchical maps proposed under different illumination conditions to tackle the localization task (Cebollada et al., 2019b).

Concerning mapping by using visual data, the models are commonly created using either local or global features. Beyond these frameworks, the use of AI with vision systems has contributed to the emergence of new paradigms to create visual maps. Zivkovic et al. (2005) build a hierarchical model based on omnidirectional images. The characterization of the data is done through local features (SIFT) and to carry out the graph partitioning, and define the map hierarchy, a cluster algorithm is proposed. Peretroukhin et al. (2017) propose the use of Bayesian Convolutional Neural Networks (BCNN) to train and implement a sun detection model from a single RGB image to incorporate global orientation information from the sun into a visual odometry pipeline. They also propose an uncertainty associated with each prediction by using a Monte Carlo dropout scheme. Clark et al. (2017) carry out a mapping and a posterior relocalization task by means of feeding an LSTM network with holistic descriptors obtained from a CNN. The proposed model estimates the current pose within an environment departing from short sequences of monocular frames. Similarly to this work, concerning the use of the CNNs to obtain global-appearance descriptors, many authors have proposed this strategy. For instance, Iyer et al. (2018) propose a self-supervised

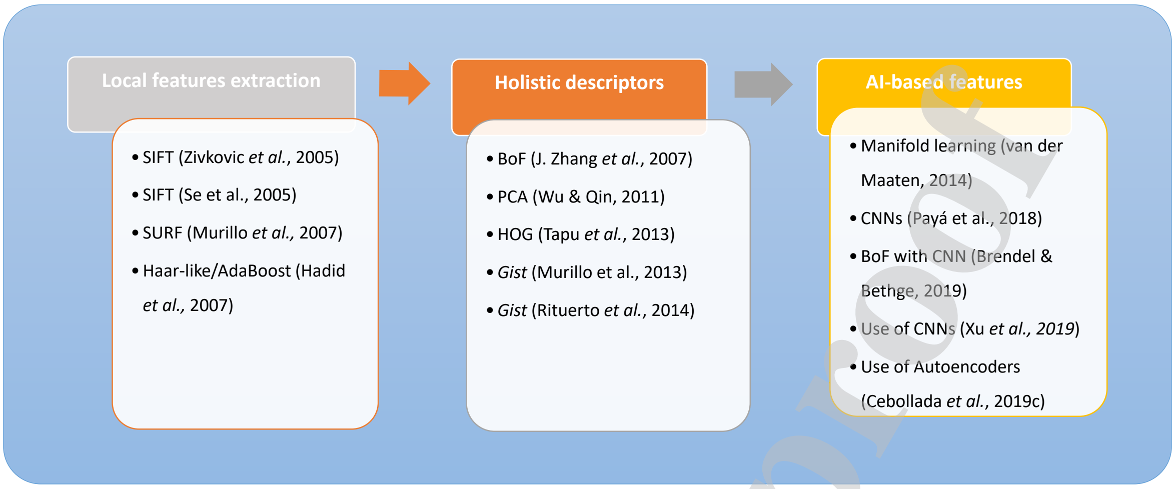
Local features extraction
• SIFT (Zivkovic *et al.*, 2005)
• SIFT (Se et al., 2005)
• SURF (Murillo *et al.*, 2007)
• Haar-like/AdaBoost (Hadid *et al.*, 2007)
Holistic descriptors
• BoF (J. Zhang *et al.*, 2007)
• PCA (Wu & Qin, 2011)
• HOG (Tapu *et al.*, 2013)
• *Gist* (Murillo et al., 2013)
• *Gist* (Rituerto *et al.*, 2014)
AI-based features
• Manifold learning (van der Maaten, 2014)
• CNNs (Payá et al., 2018)
• BoF with CNN (Brendel & Bethge, 2019)
• Use of CNNs (Xu *et al., 2019*)
• Use of Autoencoders (Cebollada *et al.*, 2019c)

Figure 4: Evolution of methods to extract relevant information from the scenes, with the purpose of solving problems in mobile robotics.

visual odometry estimation. The approach first obtains global-appearance descriptors from the fully connected layer of the VGG-11 CNN. Second, an LSTM network is used to regress pose transformations between monocular frame-pair sequences. Kopitkov & Indelman (2018) propose an approach to estimate the robot position via CNN holistic descriptors and using neural networks to learn a generative viewpoint-dependent model of CNN features given the robot pose and approximate this model by a spatially-varying Gaussian distribution. Furthermore, once developed the proposed model, it is utilized within a Bayesian framework for probabilistic inference to solve the localization problem. Sarlin et al. (2019) propose a hierarchical model using a CNN. This network simultaneously extracts local features and global descriptors that are used for accurate 6-DOF localization. Once the model is built, the coarse localization is solved by using global retrieval through a k-nearest neighbour algorithm and the holistic descriptors. The fine localization is solved through evaluating matching points from the local features.

Another widely developed strategy is the use of neural networks to model a system that is capable of estimating the position directly from the raw data. For example, Kuse et al. (2017) propose a deep residual network to model the environment representation. Naseer & Burgard (2017) develop a model that allows a 6-DOF localization using a regression neural network and a single monocular RGB image. The resulting map size is constant with respect to the size of the dataset and during the localization task, the time complexity is also constant and independent of the dataset size. Walch et al. (2017) introduce a CNN+LSTM model to estimate the pose in both indoor and outdoor environments. Raw data are introduced to the network and it is trained in such a way that the CNN layers learn suitable local features and they are then used by the LSTM layers to improve the pose estimation. In this way, the whole network learns to optimize the localization task. Brahmbhatt et al. (2018) propose a mapping model based on a regression neural network, which enables learning a data-driven map representation. Furthermore, the proposed network can be updated with unlabeled data. Payá et al. (2018) use a CNN to obtain holistic descriptors and create a hierarchical visual model with that information. Sinha et al. (2018) also propose a mapping and a subsequent localization task based on regression neural networks. The proposed method first trains a CNN that takes RGB images from a monocular camera as input and performs regression for robot pose estimation. It then incorporates the relocalization output of the CNN in an Extended Kalman Filter to tackle the localization task. Moolan-Feroze et al. (2019) propose the deployment of a model to map the environment that surrounds wind turbines. For this purpose, a CNN is trained to extract an estimate of the projection of the 3D skeleton representation departing from monocular images. After that, the localization task is solved by means of a pose graph optimization that uses the 3D representation outputs from the CNN.

4.2. Localization

As denoted in 4.1, localization is the task that tries to estimate the current position and orientation of the robot in the environment and to carry out this, a model of the environment must be available prior to start the localization. Filliat & Meyer (2003) presented an exhaustive review about the state-of-the-art strategies to carry out the localization in mobile robots. Regarding the use of vision systems together with AI, a wide range of works have been proposed in recent years. For example, Kendall et al. (2015) present a robust and real-time monocular 6-DOF relocalization system. The proposed system trains a CNN to regress the 6-DOF camera pose from a single RGB image in an end-to-end manner without additional graph optimization. Neto (2015) proposes a topological localization system based on monocular images, learning classifier systems and self-organizing maps (SOM). The whole system carries out a localization task through detecting and avoiding obstacles by means of both local and holistic features. Meng et al. (2017) address the localization issue by using methods based on Random Forests that directly estimate 3D positions with SIFT features as input. Li et al. (2018) introduce an indoor localization approach using a dual-stream regression CNN by introducing color data as well as depth data from monocular images. This system is tested under night illumination conditions and also under blur effects. As in mapping, there are also a wide range of works proposed during the last few years that propose and evaluate the use of intermediate layers from several CNNs to obtain local features or holistic descriptors. For instance, Sünderhauf et al. (2015) introduce a real-time place recognition algorithm by using different layers from CNNs to carry out the localization in large maps by integrating a variety of existing optimization techniques such as semantic search space partitioning.

Cascianelli et al. (2017) propose a strategy for mapping and posterior localization that relies on the use of a CNN to obtain local features that are robust to appearance changes. Similar to this work, Unicomb et al. (2018) also use a CNN to extract local features; in this case, they extract ground plane edges and then estimate a 6-DOF position through an EKF (Extended Kalman Filter) algorithm. Moolan-Feroze & Calway (2018) present a framework that uses CNNs to predict object feature points that are out-of-view in the input image. These feature points are then fed to estimate more robustly the pose of the robot in the environment. Holliday & Dudek (2018) propose a combination of deep-learning-based hierarchical object features and SIFT features. These points are used to perform more robust localization tasks. Regression networks have been extensively proposed to directly estimate the position within the map. For example, Sommer et al. (2017) carry out a 6-DOF localization with CNN by applying transfer learning over pre-trained CNN Google's Inception-V4. Xu et al. (2019) introduce a multi-sensor-based indoor global localization system using visual localization aided by CNN-based image retrieval with a Monte Carlo probabilistic approach. Cebollada et al. (2019c) propose the use of autoencoders and also a CNN to obtain holistic descriptors from omnidirectional images and then use them to solve the localization task in an indoor environment. Cattaneo et al. (2019) develop a regression network, which learns to localize an RGB-D image of a scene in a map built from LIDAR (Laser Imaging Detection and Ranging) data. In a similar way, Weinzaepfel et al. (2019) introduce a regression strategy based on CNN for visual localization from a single RGB image that relies on densely matching a set of objects of interest. Given a query image, the network model detects the objects, segments them and finds a dense set of 2D-2D matches between each detected object and its corresponding one in the reference image. Given these 2D-2D matches, a Perspective-n-Point problem is used to estimate the pose.

4.3. Navigation

The navigation task basically consists in solving the problem of how the robot can get to other places from its current position (Levitt & Lawton, 1990). That is, to perform a trajectory to reach a certain place. This trajectory is calculated by a path-planning system, and the robot is subsequently commanded by the control system (Barber et al., 2018). During the past few years, a wide number of works about navigation using visual sensors and AI can be found in the related literature. Maier et al. (2011) propose a machine learning classifier to address the autonomous navigation for a humanoid equipped with a monocular vision and a sparse laser data system. The classifier proposed for visual information is based on the hue and saturation values from the HSV color space with the aim of being less sensitive to illumination changes. Additionally, a classifier is also trained with texture-based information. The classifiers learn to estimate from the images which parts of the surroundings of the robot are traversable. The goal is to make the system as independent as possible of the 3D scan data. Tapu et al. (2013) introduce a visual navigation system based on HOG (Histogram of Oriented Gradients) (Dalal & Triggs, 2005) and BoW for obstacle classification. This approach basically consists in detecting objects from an image. For each object, its related HOG descriptor is introduced into the BoW retrieval framework. After that, an SVM classifier is applied to retrieve which object is. Object detection plays a vital role to carry out the navigation approach successfully. Giusti et al. (2015) presents a method to autonomously navigate through a man-made trail in the mountain for UAVs. This approach is based on a CNN classifier with the aim of operating the main direction of the trail.

S. Yang et al. (2017) propose a two-stage CNN with intermediate perception. The first stage CNN predicts the depth and surface normal from images. The second CNN predicts a path from the depth and normal maps using another CNN model. Smolyanskiy et al. (2017) propose several deep learning tools to address the autonomous navigation of Micro Aerial Vehicles (MAVs). A Recurrent Neural Network is used to estimate the view orientation and lateral offset of the MAV with respect to the trail center. In addition, another network is used to estimate depth with the aim of carrying out low-level obstacle detection. Puthussery et al. (2017) propose an autonomous navigation approach that uses the Inception v3 CNN to classify the objects detected by the camera. This classification is included within the Marker Detection Phase, which is done prior to the Robot Navigation Phase. Richter & Roy (2017) introduce a navigation system based on deep learning. On the one hand, they propose a fully connected feedforward network to model collision probability. On the other hand, they propose the use of an autoencoder to recognize when a query image is novel and requires a priority treatment. This is due to the fact that neural networks may not be efficient to provide accurate estimations when queried input data are very different from training data. Concerning the use of autoencoders for navigation, Mancini et al. (2016) introduce an object detection method that is able to detect obstacles at very long range and at a very high speed without making motion assumptions. This method is based on the use of an autoencoder which is trained with real and synthetic images and performs depth predictions. Deepika & Variyar (2017) propose a method which uses an autoencoder architecture for pixel-wise semantic segmentation of the image followed by an obstacle detection algorithm. The aim of this approach is to develop a robust vision based autonomous navigation system for self-driving cars. Walker et al. (2017) propose this tool to build a compressed representation of visual data. Images reconstructed from the compressed representation retain enough information to be used as a visual compass (an image is matched with another to recall a movement direction).

Zhao et al. (2018) present a hybrid structure with a CNN and local image features to achieve first-person vision pedestrian navigation. They also developed a novel global pooling operator which improves the results obtained by the CNN for real-time scene recognition. A SIFT-based tracking algorithm is designed for movement calculation, then the mixture of both threads perform a robust trajectory tracking. Anderson et al. (2018) present a reinforcement learning approach based on vision and natural language in a large-scale environment. The aim of this work is to provide a visually-grounded natural language navigation which is able to work properly in real buildings. Hui et al. (2018) propose an autonomous navigation approach for UAVs in outdoor environments surrounded by transmission towers and power lines. This method introduces the use of a CNN trained from end to end to address semantic segmentation and detected the power lines. Mansouri et al. (2019) present a CNN to address autonomous navigation of low-cost Micro Aerial Vehicle platforms along dark underground mine environments. The proposed CNN provides online heading rate commands for the MAV by utilising the image stream from the on-board camera, thus allowing the platform to follow a collision-free path along the tunnel axis. L. Ma et al. (2019) introduce a navigation system approach based on reinforcement learning. Moreover, they use a variational autoencoder to obtain visual features from the input images, these features are put into the network together with the target and motion information. Ruan et al. (2019) propose an approach for the navigation of mobile robots in an unknown environment using deep reinforcement learning. Through a dueling network architectures based on the double deep q network (D3QN) algorithm, the robot learns the environment and also models to navigate autonomously to the target destination with an RGB-D camera only.

4.4. Simultaneous Localization and Mapping

Additionally to the mapping and the subsequent localization task, the SLAM presents a blended alternative. This process consists in building continuously a map and updating it as the robot simultaneously estimates its position within the model. Fuentes-Pacheco et al. (2015) presented an exhaustive review about the state-of-the-art strategies to carry out SLAM. The related literature shows that not many approaches have been proposed to solve this task by using visual information and AI tools. Apart from some of the examples commented in subsections 4.1 and 4.2, which propose mapping or localization tasks with the aim of developing subsequent SLAM, there are other examples that propose complete SLAM systems. For instance, Wu & Qin (2011) propose a SLAM algorithm based on omnidirectional images. This algorithm uses incremental landmark appearance learning to provide posterior probability distribution for estimating the robot pose under a particle filtering framework. The major contribution of the work is to represent the posterior estimation of the robot pose by incremental probabilistic PCA, which can be incorporated into the particle filtering algorithm for SLAM. G. Lu et al. (2015) propose a machine learning tool known as multi-task point retrieval to develop a regression model based on 3D points local features extracted from monocular images. Garg et al. (2016) carry out an unsupervised deep convolutional network which behaves like an autoencoder, since it does not require annotated ground-truth data. This network is trained with the aim of predicting the depth map for the source image. During the training step, a pair of images (source and target) are fed into the network. Schmidt et al. (2017) introduce a CNN to produce robust local features and then use them for dense correspondence estimation and solve the SLAM task. An interesting work was developed by X. Gao & Zhang (2017) in which they carry out a loop detection by training an autoencoder that calculates local features from monocular images. Tateno et al. (2017) propose an approach which consists in two CNNs based on monocular RGB images. The first network is trained with the aim of predicting depth. CNN-predicted dense depth maps are naturally fused together with depth measurements obtained from direct monocular SLAM, based on a scheme that privileges depth prediction in image locations. The second network is trained to address semantic segmentation. Once the information is obtained from the CNNs, this is fused with more data to address the SLAM task in a highly accurate way. Mukasa et al. (2017) introduce a SLAM framework that integrates the geometrical measurements obtained from a monocular vision system with depth information predicted by means of a CNN. Tang et al. (2017) introduce the use of CNNs fed with visual information and human voice commands with the aim of solving the SLAM task with a mobile robot. Focusing on the visual information, the raw data is fed into two CNNs, the first network is used to produce accurate localization updates and the second is used to perform an object recognition task. The recognized objects are used to reinforce the mapping task. X. Zhang et al. (2017) propose a loop closure detection framework based on CNNs. In this way, the images are fed into a pre-trained CNN model to extract holistic descriptors and, after that, these descriptors are pre-processed with PCA.

More recently, Milz et al. (2018) explore the use of deep learning tools to enhance the visual SLAM. On the one hand, they propose the use of CNN to carry out the depth estimation. On the other hand, on the other hand, they propose the use of CNNs to address an end-to-end approach for learning of feature matching. This technique can learn diversity and distribution instead of just picking the top high textured features. Zhong et al. (2018) address the SLAM and object detection by using a Single Shot multi-box object Detector (SSD). The RGB-D information is introduced to the SSD and it detects moving and static objects inside the image. After this, the detected moving objects are eliminated and the rest of the data are used to build a semantic map composed of all the detected static objects in the mapping thread. Simultaneously, the dynamic objects are used to update the tracking and local mapping thread. Liang et al. (2018) propose a CNN based on 360 degrees panoramic images to carry out the visual navigation and SLAM tasks in outdoor environments. Bloesch et al. (2018) present a compact but dense representation of the scene geometry based on a deep autoencoder. This method is suitable to solve a keyframe-based monocular dense SLAM task. W. Liu et al. (2019) propose a feature-based visual SLAM. They use a CNN to obtain a more robust object location information. Y. Lu & Lu (2019) propose a SLAM approach that uses a regression CNN for pose estimation without ground truth data.

4.5. Exploration

As stated by Burgard et al. (2005) the problem of exploring an environment belongs to the fundamental problems in mobile robotics and it consists basically in covering the whole environment in a minimum amount of time. Hence, the robot must keep track of the already visited areas. This task is a blend of the navigation and mapping tasks, that is, the robot has to construct a global map in order to plan their paths and to coordinate its actions. This problem has been commonly solved by using a team of robots (Ferri et al., 2017; Michel & McIsaac, 2012; Pawgasame, 2016), since the use of multiple robots is often suggested to have several advantages over single robot systems (Y. U. Cao et al., 1995).

During the past few years, a wide number of works about exploration have been proposed by using visual sensors and AI techniques. For example, Krishnan & Krishna (2010) present a vision based exploration algorithm that invokes semantic cues for constructing a hybrid map. The approach proposes semantic labeling of the input images through a probabilistic SVM classifier that runs over a BoWs. The objective is to provide the robot with hybrid understanding of its surroundings from the lower metric characterizations to higher semantic recognition. Mukhija et al. (2012) proposes a segmentation and classification method based on visual information to support the laser to solve the obstacle avoidance task. The classification is addressed with a Gaussian Mixture Models algorithm. Craye et al. (2015) propose a method for incrementally learning a mechanism of visual saliency. The proposed system is trained and learns the visual aspect of salient elements within their context. The RGB-D data are used to train a random forest classifier, which learns to determine whether the area is salient or not. Tai & Liu (2016) introduce a reinforcement learning method to address the exploration task in a corridor environment. The learning model receives information from a CNN fed with RGB-D data. The Q-network is used in the robot controller.

More recently, Choudhury et al. (2017) propose an algorithm which trains a policy to gather information to carry out a exploration task with UAVs. When the distribution corresponds to a scene containing ladders, the learned policy executes a helical motion around parts of the observed. On the contrary, when the distribution corresponds to a scene from a construction site, the learned policy executes a large sweeping motion. The proposed algorithm is based on a classifier among other tools. C. Liu et al. (2017) introduce an end-to-end learning model based CNN that converts directly the raw visual data to steering commands. Tai et al. (2017a) propose an exploration algorithm which uses a hierarchical structure that fuses several CNN layers with decision-making process. The system is trained by taking RGB-D information as input and generates a sequence of main moving direction as output. Flaspohler et al. (2017) present an autoencoder which encodes information from visual data to carry out exploration with marine robots. H. Wang et al. (2018) propose an optimal light intensity optimization method to address efficiently the visual navigation. The proposed method is mainly based on a regression model to automatically predict optimal light intensity values for desired image quality when camera observation distances fluctuate. Nevertheless, a classification task is addressed rather than a regression. The query image is classified according to five levels of brightness. In this framework, simple features are extracted using intensity histogram and utilized as primary features to describe the distribution of image intensity. After that, the primary features are made more discriminative by an evolution process with stacked autoencoders. The evaluation of brightness is solved by introducing the holistic descriptor into a softmax classifier. Ly & Tsai (2019) propose an autonomous exploration, reconstruction, and surveillance of 3D Environments by using deep learning. They establish a gain function for each issue. After that, they use CNNs to approximate the corresponding gain function.

To conclude, Table 2 presents an outline of the approaches presented in the present section and their main characteristics: the type of task (subsections 4.1, 4.2, 4.3, 4.4, or 4.5), the type of image, the AI tool used and the kind of visual data employed (local features, holistic descriptors or raw data).

5. Conclusions

Vision sensors constitute a robust alternative to capture the necessary information to solve a variety of tasks in the field of mobile robotics. Additionally, during the past few years, systems based on AI have been used extensively to carry out the tasks more efficiently. Consequently, the amount of works that use visual sensors and AI has increased substantially and many approaches can be found to solve the mapping, localization, navigation, SLAM and exploration tasks.

The present review presents a collection of the main proposals to solve the mobile robotics tasks by means of visual information and AI tools. To this end, this work started focusing on the AI tools which are more commonly used together with vision systems. After that, the review has focused on how visual information can be described and handled by means of AI techniques. Two main options are available: local features and methods based on holistic description. Finally, the present work has focused on the study of the mapping, localization, navigation, SLAM and exploration tasks in mobile robotics.

The huge amount of works regarding these topics show how vision systems, AI techniques and mobile robotics are three very active research areas and hence, the research on them is expected to continue increasing during the following years. This work has shown that a great variety of AI and deep learning tools can be used in mobile robotics and computer vision. Such tools have provided good solutions to some specific problems in these fields, such as the extraction and labelling of relevant information from the scenes; the creation of models of the environment and the estimation of the position and orientation of the robot from raw data; and the exploration of initially unknown environments. Good solutions to these problems have been proposed in specific scenarios, depending on the motion abilities of the robot and the characteristics of the surroundings (indoors-outdoors, aerial-terrestrial-underwater, etc.). Notwithstanding that, there are still some issues that need to be addressed more robustly to enable mobile robots to move and perform their tasks more autonomously in complex, heterogeneous and changing environments and circumstances, trying to provide a complete and more integral solution to the SLAM and navigation problems, and AI shows potential to address these challenges. In this sense, finding robust and fast solutions to overcome current problems will help to improve the autonomy of mobile robots. Therefore, their range of use will also increase. Nowadays, there are some technologies which are closely related to AI and can be of interest to researchers in the fields of mobile robotics and computer vision. As detailed in the survey, one of the major issues of AI and deep learning is the computationally expensive process to train the models with a large number of samples. In this sense, some current technologies, such as cloud computing, data science and big data provide robust tools and approaches to address this issue, and therefore they may contribute to a quicker development of AI techniques in mobile robotics (Gill et al., 2019; Zhu & Zheng, 2018; Allam & Dhunny, 2019). Second, as shown throughout the review, AI techniques are contributing to a major autonomy of the mobile robots in a wider variety of environments and circumstances. This increase of autonomy plays a crucial role in the development of other relevant technologies to current society, such as IoT (Internet of Things), smart cities and industry 5.0 (S. K. Singh et al., 2020; Chui et al., 2018; Özdemir & Hekim, 2018).

Table 2: Summary of works that use AI together with vision systems to solve mapping, localization, navigation,SLAM or exploration tasks.

Reference	Task	Type of image	AI tool	Input data
Zivkovic et al. (2005)	4.1	omnidirectional	Clustering	local features
Peretroukhin et al. (2017)	4.1	stereo	BCNN, Monte Carlo	local features
Clark et al. (2017)	4.1, 4.2	monocular	CNN, LSTM	holistic descriptors
Kuse et al. (2017)	4.1	monocular	Regression CNN	raw images
Naseer & Burgard (2017)	4.1, 4.2	monocular	Regression CNN	raw images
Walch et al. (2017)	4.1, 4.2	monocular	CNN, LSTM	raw images
Iyer et al. (2018)	4.1	monocular	CNN, LSTM	holistic descriptors
Brahmbhatt et al. (2018)	4.1	monocular	Regression CNN	raw images
Sinha et al. (2018)	4.1, 4.2	monocular	Regression CNN	raw images
Kopitkov & Indelman (2018)	4.1, 4.2	monocular	CNN, Gaussian prametrization	holistic descriptors
Moolan-Feroze et al. (2019)	4.1, 4.2	monocular	CNN	raw images
Sarlin et al. (2019)	4.1, 4.2	monocular	k-nearest neighbors, CNN	local features, holistic descriptors
Kendall et al. (2015)	4.2	monocular	CNN	raw images
Neto (2015)	4.2	monocular	Classifier, Kohonen SOM	local features, holistic descriptors
Sünderhauf et al. (2015)	4.2	monocular	CNN	holistic descriptors
Li et al. (2018)	4.2	monocular	Regression CNN	raw images
Cascianelli et al. (2017)	4.1, 4.2	monocular	CNN	local features
Sommer et al. (2017)	4.2	monocular	Regression CNN	raw images
Meng et al. (2017)	4.2	monocular	Random forests	local features
Unicomb et al. (2018)	4.2	monocular	CNN	local features
Moolan-Feroze & Calway (2018)	4.2	monocular	Recurrent Neural Network	local features
Holliday & Dudek (2018)	4.2	monocular	CNN	local features
Cattaneo et al. (2019)	4.2	stereo	Regression CNN	raw RGB-D data
Cebollada et al. (2019c)	4.2	omnidirecional	Autoencoder, CNN	holistic descriptors
Weinzaepfel et al. (2019)	4.2		Regression CNN	local features
Wu & Qin (2011)	4.4	omnidirectional	Incremental landmark appearance learning	raw images
G. Lu et al. (2015)	4.4	monocular	Multi-task learning	local features
Garg et al. (2016)	4.4	panoramic	Autoencoder	raw images
Schmidt et al. (2017)	4.4	monocular	CNN	local features
X. Gao & Zhang (2017)	4.4	monocular	Autoencoder	local features
Tateno et al. (2017)	4.4	monocular	CNN	raw images
Mukasa et al. (2017)	4.4	monocular	CNN	raw images
Tang et al. (2017)	4.4	stereo	CNN	raw images
X. Zhang et al. (2017)	4.4	monocular	CNN	raw images
Milz et al. (2018)	4.4	monocular	CNN	raw images
Zhong et al. (2018)	4.4	monocular	SSD	raw RGB-D data
Liang et al. (2018)	4.3, 4.4	panoramic	CNN	raw images
Bloesch et al. (2018)	4.4	monocular	Autoencoder	raw images
W. Liu et al. (2019)	4.4	monocular	CNN	local features
Y. Lu & Lu (2019)	4.4	monocular	Recurrent CNN	raw images
Maier et al. (2011)	4.3	monocular	Classifier	HSV and texture data
Tapu et al. (2013)	4.3	monocular	SVM classifier	HOG and BoW
Giusti et al. (2015)	4.3	monocular	CNN	raw images
Mancini et al. (2016)	4.3	monocular	Autoencoder	raw images and Optical Flow
Deepika & Variyar (2017)	4.3	monocular	Autoencoder	raw images
S. Yang et al. (2017)	4.1, 4.3	monocular	CNN	raw imagees
Smolyanskiy et al. (2017)	4.3	monocular	Recurrent Neural Network	raw images
Puthussery et al. (2017)	4.3	monocular	CNN	raw RGB-D data
Richter & Roy (2017)	4.3	monocular	CNN and Autoencoder	raw images

Table 2: Summary of works that use AI together with vision systems to solve mapping, localization, navigation,SLAM or exploration tasks.

Reference	Task	Type of image	AI tool	Input data
Walker et al. (2017)	4.3	panoramic	Autoencoder	raw images
Anderson et al. (2018)	4.3	monocular	Reinforcement Learning	raw images
Zhao et al. (2018)	4.3	monocular	CNN	raw images, local features
Hui et al. (2018)	4.3	monocular	CNN	raw images
Mansouri et al. (2019)	4.3	monocular	CNN	raw images
L. Ma et al. (2019)	4.3	monocular	Reinforcement Learning, Autoencoder	raw images
Ruan et al. (2019)	4.3	monocular	Deep Reinforcement Learning	raw RGB-D data
Krishnan & Krishna (2010)	4.5	monocular	SVM classifier	BoW
Mukhija et al. (2012)	4.5	monocular	Gaussian Mixture Models classifier	raw images
Craye et al. (2015)	4.5	monocular	Random forest classifier	local features
Tai & Liu (2016)	4.5	monocular	Deep Reinforcement Learning, CNN	raw RGB-D data
Choudhury et al. (2017)	4.5	monocular	Classifier	local features
C. Liu et al. (2017)	4.5	monocular	CNN	raw images
Tai et al. (2017a)	4.5	monocular	CNN	raw RGB-D data
Flaspohler et al. (2017)	4.5	monocular	Autoencoder	holistic descriptors
H. Wang et al. (2018)	4.5	monocular	Autoencoder and Classifier	holistic descriptors
Ly & Tsai (2019)	4.5	monocular	CNN	raw RGB-D data

The following abbreviations are used in this manuscript:

AI	Artificial intelligence
BCNN	Bayesian Convolutional Neural Network
BoF	Bag of Features
BoW	Bag of Words
BRIEF	Binary Robust Independent Elementary Features
CCD	Charge-Coupled Device
CCTV	Closed-Circuit Television
CenSurE	Center Surround Extremas
CNN	Convolutional Neural Network
D3QN	Dueling Architecture based double deep Q Network
DAN	Deep Adaption Network
DARPA	Defense Advanced Research Projects Agency
DBM	Deep Boltzmann Machines
DBN	Deep Belief Networks
EEG	Electroencephalography
EKF	Extended Kalman Filter
FAST	Features From Accelerated Segment Test
FREAK	Fast Retina Keypoint
HLC	High-level controller
HOG	Histogram of Oriented Gradients
HP-CNN	Hypercube Pyramid Convolutional Neural Network
PCA	Principal Components Analysis
ICA	Independent Component Analysis
IoT	Internet of Things
IP	Interest point
LBP	Local Binary Pattern
LIDAR	Laser Imaging Detection and Ranging
LSTM	Long short-term memory
MAV	Micro Aerial Vehicles
ML	Machine Learning
MLP	Multilayer Perceptrons
MOP-CNN	Multi-Scale Orderless Pooling Convolutional Neural Network
MSE	Mean Squared Error
ORB	Oriented FAST and rotated BRIEF
PI	Proportional–integral
RBM	Restricted Boltzmann Machines
RF	Regression Forest
R-CNN	Regression Convolutional Neural Network
RNN	Recurrent Neural Network
ROI	Region Of Interest
RVM	Relevance Vector Machine
SDA	Stacked Denoising Autoencoder
SIFT	Scale-Invariant Feature Transform
SLAM	Simultaneous Localization And Mapping
SOM	Self-Organizing Maps
SSD	Single Shot multi-box object Detector
SURF	Speeded-Up Robust Features
SVM	Support Vector Machine
t-SNE	t-distributed Stochastic Neighbor Embedding
UAV	Unmanned Aerial Vehicle
VLAD	Vector of Locally Aggregated Descriptors
WSN	Wireless Sensor Network

References

Abate, A. F., Nappi, M., Riccio, D., & Sabatino, G. (2007). 2d and 3d face recognition: A survey. *Pattern recognition letters*, *28*(14), 1885–1906.

Accelerating t-SNE using tree-based algorithms, author=van der Maaten, L. (2014). *The Journal of Machine Learning Research*, *15*(1), 3221–3245.

Aguilar, W. G., Luna, M. A., Moya, J. F., Abad, V., Parra, H., & Ruiz, H. (2017). Pedestrian detection for UAVs using cascade classifiers with meanshift. In *2017 IEEE 11th international conference on semantic computing (ICSC)* (pp. 509–514).

Ahuja, K., K V, A. K., Kiran, A., Dalin, E., & Sagar, S. (2018, 06). Smart office surveillance robot using face recognition. *International Journal of Mechanical and Production Engineering Research and Development*, *8*, 725-734. doi: 10.24247/ijmperdjun201877

Allam, Z., & Dhunny, Z. A. (2019). On big data, artificial intelligence and smart cities. *Cities*, *89*, 80–91.

Amorós, F., Payá, L., Marín, J. M., & Reinoso, O. (2018). Trajectory estimation and optimization through loop closure detection, using omnidirectional imaging and global-appearance descriptors. *Expert Systems with Applications*, *102*, 273–290.

Anderson, P., Wu, Q., Teney, D., Bruce, J., J., M., Sünderhauf, N., … van den Hengel, A. (2018). Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3674–3683).

Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., & Sivic, J. (2016). NetVLAD: CNN architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5297–5307).

Arroyo, R., Alcantarilla, P. F., Bergasa, L. M., & Romera, E. (2016, Oct). Fusion and binarization of CNN features for robust topological localization across seasons. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (p. 4656-4663). doi: 10.1109/IROS.2016.7759685

Atkinson, J., & Campos, D. (2016). Improving bci-based emotion recognition by combining eeg feature selection and kernel classifiers. *Expert Systems with Applications*, *47*, 35–41.

Badue, C., Guidolini, R., Carneiro, R. V., Azevedo, P., Cardoso, V. B., Forechi, A., … others (2019). Self-driving cars: A survey. *arXiv preprint arXiv:1901.04407*.

Barber, R., Crespo, J., Gómez, C., Hernámdez, A. C., & Galli, M. (2018). Mobile Robot Navigation in Indoor Environments: Geometric, Topological, and Semantic Navigation. In *Applications of Mobile Robots*. IntechOpen.

Bay, H., Ess, A., Tuytelaars, T., & Van Gool, L. (2008). Speeded-up robust features (surf). *Computer vision and image understanding*, *110*(3), 346–359.

Belkin, M., & Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, *15*(6), 1373–1396.

Bell, A. J., & Sejnowski, T. J. (1997). The "independent components" of natural scenes are edge filters. *Vision research*, *37*(23), 3327–3338.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, *35*(8), 1798–1828.

Berenguer, Y., Payá, L., Valiente, D., Peidró, A., & Reinoso, O. (2019). Relative altitude estimation using omnidirectional imaging and holistic descriptors. *Remote Sensing*, *11*(3), 323.

Bilgili, M., & Sahin, B. (2010). Comparative analysis of regression and artificial neural network models for wind speed prediction. *Meteorology and Atmospheric Physics*, *109*(1-2), 61–72.

Bishop, C. M. (2006). Pattern recognition and machine learning springer-verlag new york. *Inc. Secaucus, NJ, USA*, *2006*.

Bloesch, M., Czarnowski, J., Clark, R., Leutenegger, S., & Davison, A. J. (2018). Codeslam—learning a compact, optimisable representation for dense visual slam. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2560–2568).

Boularias, A., Bagnell, J. A., & Stentz, A. (2015). Learning to manipulate unknown objects in clutter by reinforcement. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Brahmbhatt, S., Gu, J., Kim, K., Hays, J., & Kautz, J. (2018, June). Geometry-Aware Learning of Maps for Camera Localization. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (p. 2616-2625). doi: 10.1109/CVPR.2018.00277

Brendel, W., & Bethge, M. (2019). Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. *arXiv preprint arXiv:1904.00760*.

Burgard, W., Moors, M., Stachniss, C., & Schneider, F. E. (2005). Coordinated multi-robot exploration. *IEEE Transactions on robotics*, *21*(3), 376–386.

Calderon-Cordova, C., Ramírez, C., Barros, V., Quezada-Sarmiento, P. A., & Barba-Guamán, L. (2016). EMG signal patterns recognition based on feedforward Artificial Neural Network applied to robotic prosthesis myoelectric control. In *2016 Future Technologies Conference (FTC)* (pp. 868–875).

Calonder, M., Lepetit, V., Strecha, C., & Fua, P. (2010). Brief: Binary robust independent elementary features. *Computer Vision–ECCV 2010*, 778–792.

Cao, J., Huang, Z., & Shen, H. T. (2017). Local deep descriptors in bag-of-words for image retrieval. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017* (pp. 52–58).

Cao, Y. U., Fukunaga, A. S., Kahng, A. B., & Meng, F. (1995). Cooperative mobile robotics: Antecedents and directions. In *Proceedings 1995 IEEE/RSJ International Conference on Intelligent Robots and Systems. Human Robot Interaction and Cooperative Robots* (Vol. 1, pp. 226–234).

Carreira, F., Calado, J. M., Cardeira, C., & Oliveira, P. (2015). Enhanced PCA-based localization using depth maps with missing data. *Journal of Intelligent & Robotic Systems*, *77*(2), 341–360.

Cascianelli, S., Costante, G., Bellocchio, E., Valigi, P., Fravolini, M. L., & Ciarfuglia, T. A. (2017). Robust visual semi-semantic loop closure detection by a covisibility graph and CNN features. *Robotics and Autonomous Systems*, *92*, 53 - 65. Retrieved from http://www.sciencedirect.com/science/article/pii/S0921889016304900 doi: https://doi.org/10.1016/j.robot.2017.03.004

Cattaneo, D., Vaghi, M., Ballardini, A. L., Fontana, S., Sorrenti, D. G., & Burgard, W. (2019, Oct). CMRNet: Camera to LiDAR-Map Registration. In *2019 ieee intelligent transportation systems conference (itsc)* (p. 1283-1289). doi: 10.1109/ITSC.2019.8917470

Cebollada, S., Payá, L., Flores, M., Román, V., Peidró, A., & Reinoso, O. (2020). A deep learning tool to solve localization in mobile autonomous robotics. In *ICINCO 2020, 17th Intl. Conf. On Informatics in Control, Automation and Robotics (Online streaming, 7-9 July 2020)* (p. 232-241). Ed. INSTICC.

Cebollada, S., Payá, L., Mayol, W., & Reinoso, O. (2019a). Evaluation of clustering methods in compression of topological models and visual place recognition using global appearance descriptors. *Applied Sciences*, *9*(3), 377.

Cebollada, S., Payá, L., Román, V., & Reinoso, O. (2019b). Hierarchical localization in topological models under varying illumination using holistic visual descriptors. *IEEE Access*, *7*, 49580-49595. doi: 10.1109/ACCESS.2019.2910581

Cebollada, S., Payá, L., Valiente, D., Jiang, X., & O., R. (2019c). An evaluation between global appearance descriptors based on analytic methods and deep learning techniques for localization in autonomous mobile robots. In *ICINCO 2019, 16th International Conference on Informatics in Control, Automation and Robotics (Prague, Czech Republic, 29-31 July, 2019)* (pp. 284–291). Ed. INSTICC.

Cebollada, S., Payá, L., Juliá, M., Holloway, M., & Reinoso, O. (2018). Mapping and localization module in a mobile robot for insulating building crawl spaces. *Automation in Construction*, *87*, 248 - 262. Retrieved from http://www.sciencedirect.com/science/article/pii/S0926580517306726 doi: https://doi.org/10.1016/j.autcon.2017.11.007

Charniak, E., McDermott, D., & McDermott, D. (1985). *Introduction to artificial intelligence*. Addison-Wesley. Retrieved from https://books

.google.es/books?id=kPCAR2VEd1YC

Chaves, D., Ruiz-Sarmiento, J., Petkov, N., & Gonzalez-Jimenez, J. (2019). Integration of cnn into a robotic architecture to build semantic maps of indoor environments. In *International work-conference on artificial neural networks* (pp. 313–324).

Chollet, F. (2017). *Deep learning with python.* Manning Publications Company.

Choudhury, S., Kapoor, A., Ranade, G., & Dey, D. (2017). Learning to gather information via imitation. In *2017 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 908–915).

Chui, K. T., Lytras, M. D., & Visvizi, A. (2018). Energy sustainability in smart cities: Artificial intelligence, smart monitoring, and optimization of energy consumption. *Energies*, *11*(11), 2869.

Clark, R., Wang, S., Markham, A., Trigoni, N., & Wen, H. (2017, July). VidLoc: A Deep Spatio-Temporal Model for 6-DoF Video-Clip Relocalization. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (p. 2652-2660). doi: 10.1109/CVPR.2017.284

Coates, A., Ng, A., & Lee, H. (2011). An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics* (pp. 215–223).

Craye, C., Filliat, D., & Goudou, J. F. (2015). Exploration strategies for incremental learning of object-based visual saliency. In *2015 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)* (pp. 13–18).

Csurka, G., Dance, C., Fan, L., Willamowski, J., & Bray, C. (2004). Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV* (Vol. 1, pp. 1–2).

da Silva, S. P. P., da Nóbrega, R. V. M., Medeiros, A. G., Marinho, L. B., Almeida, J. S., & Filho, P. P. R. (2018, July). Localization of Mobile Robots with Topological Maps and Classification with Reject Option using Convolutional Neural Networks in Omnidirectional Images. In *2018 International Joint Conference on Neural Networks (IJCNN)* (p. 1-8). doi: 10.1109/IJCNN.2018.8489328

Dalal, N., & Triggs, B. (2005). Histograms of Oriented Gradients for Human Detection. In *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Diego, USA. Vol. II, pp. 886-893.*

Deepika, N., & Variyar, V. (2017). Obstacle classification and detection for vision based navigation for autonomous driving. In *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)* (pp. 2092–2097).

De Momi, E., & Ferrigno, G. (2010). Robotic and artificial intelligence for keyhole neurosurgery: the robocast project, a multi-modal autonomous path planner. *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine*, *224*(5), 715–727.

Dezfoulian, S. H., Wu, D., & Ahmad, I. S. (2013). A generalized neural network approach to mobile robot navigation and obstacle avoidance. In *Intelligent Autonomous Systems 12* (pp. 25–42). Springer.

Dhanachandra, N., Manglem, K., & Chanu, Y. J. (2015). Image segmentation using k-means clustering algorithm and subtractive clustering algorithm. *Procedia Computer Science*, *54*, 764–771.

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., & Darrell, T. (2014). Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning* (pp. 647–655).

Donoho, D., & Grimes, C. (2003). *Hessian eigenmaps: new locally linear embedding techniques for highdimensional data (Technical Report TR2003-08).* Stanford.

Faessler, M., Fontana, F., Forster, C., Mueggler, E., Pizzoli, M., & Scaramuzza, D. (2016). Autonomous, vision-based flight and live dense 3d mapping with a quadrotor micro aerial vehicle. *Journal of Field Robotics*, *33*(4), 431–450.

Fan, H., Zheng, L., Yan, C., & Yang, Y. (2018). Unsupervised person reidentification: Clustering and fine-tuning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, *14*(4), 83.

Feng, J., Liu, Y., & Wu, L. (2017). Bag of visual words model with deep spatial features for geographical scene classification. *Computational intelligence and neuroscience*, *2017*.

Ferreira, J. P., Amaral, T. G., Pires, V. F., Crisostomo, M. M., & Coimbra, P. (2004). A neural-fuzzy walking control of an autonomous biped robot. In *Proceedings world automation congress, 2004.* (Vol. 15, pp. 253–258).

Ferri, G., Munafò, A., Tesei, A., Braca, P., Meyer, F., Pelekanakis, K., … LePage, K. (2017). Cooperative robotic networks for underwater surveillance: an overview. *IET Radar, Sonar & Navigation*, *11*(12), 1740–1761.

Filliat, D., & Meyer, J. A. (2003). Map-based navigation in mobile robots:: I. a review of localization strategies. *Cognitive Systems Research*, *4*(4), 243 - 282. Retrieved from http://www.sciencedirect.com/science/article/pii/S1389041703000081 doi: https://doi.org/10.1016/S1389-0417(03)00008-1

Flaspohler, G., Roy, N., & Girdhar, Y. (2017). Feature discovery and visualization of robot mission data using convolutional autoencoders and bayesian nonparametric topic models. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 1–8).

François-Lavet, V., Henderson, P., Islam, R., Bellemare, M. G., & Pineau, J. (2018). An introduction to deep reinforcement learning. *Foundations and Trends® in Machine Learning*, *11*(3-4), 219–354.

Freund, Y., & Schapire, R. E. (1995). A desicion-theoretic generalization of on-line learning and an application to boosting. In *European conference on computational learning theory* (pp. 23–37).

Fuentes-Pacheco, J., Ruiz-Ascencio, J., & Rendón-Mancha, J. M. (2015). Visual simultaneous localization and mapping: a survey. *Artificial intelligence review*, *43*(1), 55–81.

Gadoue, S. M., Giaouris, D., & Finch, J. (2009). Artificial intelligence-based speed control of dtc induction motor drives—a comparative study. *Electric Power Systems Research*, *79*(1), 210–219.

Gao, H., Cheng, B., Wang, J., Li, K., Zhao, J., & Li, D. (2018). Object classification using cnn-based fusion of vision and lidar in autonomous vehicle environment. *IEEE Transactions on Industrial Informatics*, *14*(9), 4224–4231.

Gao, X., & Zhang, T. (2017). Unsupervised learning to detect loops using deep neural networks for visual slam system. *Autonomous robots*, *41*(1), 1–18.

Garcia-Fidalgo, E., & Ortiz, A. (2015). Vision-based topological mapping and localization methods: A survey. *Robotics and Autonomous Systems*, *64*, 1 - 20. Retrieved from http://www.sciencedirect.com/science/article/pii/S0921889014002619 doi: https://doi.org/10.1016/j.robot.2014.11.009

Garg, R., BG, V. K., Carneiro, G., & Reid, I. (2016). Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *European Conference on Computer Vision* (pp. 740–756).

Gill, S. S., Tuli, S., Xu, M., Singh, I., Singh, K. V., Lindsay, D., … others (2019). Transformative effects of IoT, Blockchain and Artificial Intelligence on cloud computing: Evolution, vision, trends and open challenges. *Internet of Things*, *8*, 100118.

Giusti, J., A.and Guzzi, Cireşan, D. C., He, F. L., Rodríguez, J. P., Fontana, F., Faessler, M., … Gambardella, L. M. (2015). A machine learning approach to visual perception of forest trails for mobile robots. *IEEE Robotics and Automation Letters*, *1*(2), 661–667.

Glorot, X., Bordes, A., & Bengio, Y. (2011). Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th international conference on machine learning (ICML-11)* (pp. 513–520).

Gong, Y., Wang, L., Guo, R., & Lazebnik, S. (2014). Multi-scale orderless pooling of deep convolutional activation features. In *European conference on computer vision* (pp. 392–407).

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Gordo, A., Almazán, J., Revaud, J., & Larlus, D. (2016). Deep image retrieval: Learning global representations for image search. In *European conference on computer vision* (pp. 241–257).

Graves, A. (2012). Supervised sequence labelling. In *Supervised sequence labelling with recurrent neural networks* (pp. 5–13). Springer.

Graves, A., Mohamed, A., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing* (pp. 6645–6649).

Gwinner, K., Jaumann, R., Hauber, E., Hoffmann, H., Heipke, C., Oberst, J., … others (2016). The High Resolution Stereo Camera (HRSC) of Mars Express and its approach to science analysis and mapping for Mars and its satellites. *Planetary and Space Science*, *126*, 93–138.

Hadid, A., Heikkila, J., Silvén, O., & Pietikainen, M. (2007). Face and eye detection for person authentication in mobile phones. In *2007 First ACM/IEEE International Conference on Distributed Smart Cameras* (pp. 101–108).

Han, D., Liu, Q., & Fan, W. (2018). A new image classification method using cnn transfer learning and web data augmentation. *Expert Systems with Applications*, *95*, 43–56.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Hinton, G. E., Osindero, S., & Teh, Y. (2006a). A fast learning algorithm for deep belief nets. *Neural computation*, *18*(7), 1527–1554.

Hinton, G. E., & Salakhutdinov, R. R. (2006b). Reducing the dimensionality of data with neural networks. *science*, *313*(5786), 504–507.

Hinton, G. E., Sejnowski, T. J., & Poggio, T. A. (1999). *Unsupervised learning: foundations of neural computation*. MIT press.

Holliday, A., & Dudek, G. (2018, Oct). Scale-Robust Localization Using General Object Landmarks. In *2018 ieee/rsj international conference on intelligent robots and systems (iros)* (p. 1688-1694). doi: 10.1109/IROS.2018.8594011

Holzer, S., Shotton, J., & Kohli, P. (2012). Learning to efficiently detect repeatable interest points in depth data. In *European conference on computer vision* (pp. 200–213).

Hui, X., Bian, J., Zhao, X., & Tan, M. (2018). Vision-based autonomous navigation approach for unmanned aerial vehicle transmission-line inspection. *International Journal of Advanced Robotic Systems*, *15*(1), 1729881417752821.

Iyer, G., Murthy, J. K., Gupta, G., Krishna, K. M., & Paull, L. (2018, June). Geometric Consistency for Self-Supervised End-to-End Visual Odometry. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (p. 380-3808). doi: 10.1109/CVPRW.2018.00064

Jafri, R., & Arabnia, H. R. (2009). A survey of face recognition techniques. *Jips*, *5*(2), 41–68.

Jain, A. K., & Li, S. Z. (2011). *Handbook of face recognition*. Springer.

Jia, Y., Li, M., An, L., & Zhang, X. (2003). Autonomous navigation of a miniature mobile robot using real-time trinocular stereo machine. In *Robotics, Intelligent Systems and Signal Processing, 2003. Proceedings. 2003 IEEE International Conference on* (Vol. 1, p. 417-421 vol.1). doi: 10.1109/RISSP.2003.1285610

Jiang, W., & Wang, W. (2017, March). Face detection and recognition for home service robots with end-to-end deep neural networks. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (p. 2232-2236). doi: 10.1109/ICASSP.2017.7952553

Jiménez, P. (2012). Survey on model-based manipulation planning of deformable objects. *Robotics and computer-integrated manufacturing*, *28*(2), 154–163.

Jurie, F., & Triggs, B. (2005). Creating efficient codebooks for visual recognition. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1* (Vol. 1, pp. 604–610).

Jutten, C., & Herault, J. (1991). Blind separation of sources, part i: An adaptive algorithm based on neuromimetic architecture. *Signal processing*, *24*(1), 1–10.

Kahn, G., Villaflor, A., Ding, B., Abbeel, P., & Levine, S. (2018). Self-supervised deep reinforcement learning with generalized computation graphs for robot navigation. In *2018 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 1–8).

Kanezaki, A., Matsushita, Y., & Nishida, Y. (2018). Rotationnet: Joint object categorization and pose estimation using multiviews from unsupervised viewpoints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 5010–5019).

Kayaer, K., & Yıldırım, T. (2003). Medical diagnosis on Pima Indian diabetes using general regression neural networks. In *Proceedings of the international conference on artificial neural networks and neural information processing (ICANN/ICONIP)* (Vol. 181, p. 184).

Kendall, A., Grimes, M., & Cipolla, R. (2015, Dec). PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization. In *2015 IEEE International Conference on Computer Vision (ICCV)* (p. 2938-2946). doi: 10.1109/ICCV.2015.336

Khan, F. S., Van De Weijer, J., Anwer, R. M., Bagdanov, A. D., Felsberg, M., & Laaksonen, J. (2018). Scale coding bag of deep features for human attribute and action recognition. *Machine Vision and Applications*, *29*(1), 55–71.

Kim, K. (2005). Intelligent immigration control system by using passport recognition and face verification. In *International Symposium on Neural Networks* (pp. 147–156).

Kingma, D. P., & Cun, Y. L. (2010). Regularized estimation of image statistics by score matching. In *Advances in neural information processing systems* (pp. 1126–1134).

Kirby, M. (2000). *Geometric data analysis: an empirical approach to dimensionality reduction and the study of patterns*. John Wiley & Sons, Inc.

Kohavi, R., & Quinlan, J. R. (2002). Data mining tasks and methods: Classification: decision-tree discovery. In *Handbook of data mining and knowledge discovery* (pp. 267–276).

Kopitkov, D., & Indelman, V. (2018, Oct). Bayesian Information Recovery from CNN for Probabilistic Inference. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (p. 7795-7802). doi: 10.1109/IROS.2018.8594506

Korrapati, H., & Mezouar, Y. (2017). Multi-resolution map building and loop closure with omnidirectional images. *Autonomous Robots*, *41*(4), 967–987.

Korytkowski, M., Rutkowski, L., & Scherer, R. (2016). Fast image classification by boosting fuzzy classifiers. *Information Sciences*, *327*, 175–182.

Krishnan, A. K., & Krishna, K. M. (2010). A visual exploration algorithm using semantic cues that constructs image based hybrid maps. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems* (pp. 1316–1321).

Krittanawong, C., Zhang, H., Wang, Z., Aydar, M., & Kitai, T. (2017). Artificial intelligence in precision cardiovascular medicine. *Journal of the American College of Cardiology*, *69*(21), 2657–2664.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Kuipers, B., Modayil, J., Beeson, P., MacMahon, M., & Savelli, F. (2004). Local metrical and global topological maps in the hybrid spatial semantic hierarchy. In *IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA'04. 2004* (Vol. 5, pp. 4845–4851).

Kumar, R., Aggarwal, R., & Sharma, J. (2015). Comparison of regression and artificial neural network models for estimation of global solar radiations. *Renewable and Sustainable Energy Reviews*, *52*, 1294–1299.

Kunii, Y., Kovacs, G., & Hoshi, N. (2017). Mobile robot navigation in natural environments using robust object tracking. In *2017 IEEE 26th international symposium on industrial electronics (ISIE)* (pp. 1747–1752).

Kuse, M., Jaiswal, S. P., & Shen, S. (2017, Sep.). Deep-mapnets : A residual network for 3D environment representation. In *2017 IEEE International Conference on Image Processing (ICIP)* (p. 2652-2656). doi: 10.1109/ICIP.2017.8296763

Kuutti, S., Fallah, S., Katsaros, K., Dianati, M., Mccullough, F., & Mouzakitis, A. (2018). A survey of the state-of-the-art localization techniques and their potentials for autonomous vehicle applications. *IEEE Internet of Things Journal*, *5*(2), 829–846.

Larochelle, H., Bengio, Y., Louradour, J., & Lamblin, P. (2009). Exploring strategies for training deep neural networks. *Journal of machine learning research*, *10*(Jan), 1–40.

Lazebnik, S., Schmid, C., & Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *2006 IEEE Computer Society Conference on Computer Vision and Pat-*

tern Recognition (CVPR'06) (Vol. 2, pp. 2169–2178).

Le, Q. V. (2013). Building high-level features using large scale unsupervised learning. In *2013 IEEE international conference on acoustics, speech and signal processing* (pp. 8595–8598).

Le, Q. V., Zou, W. Y., Yeung, S. Y., & Ng, A. Y. (2011). Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In *CVPR 2011* (pp. 3361–3368).

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

Lenc, K., & Vedaldi, A. (2016). Learning covariant feature detectors. In *European Conference on Computer Vision* (pp. 100–117).

Lenc, K., & Vedaldi, A. (2018). Large scale evaluation of local image feature detectors on homography datasets. *arXiv preprint arXiv:1807.07939*.

Levitt, T. S., & Lawton, D. T. (1990). Qualitative navigation for mobile robots. *Artificial intelligence*, *44*(3), 305–360.

Li, R., Liu, Q., Gui, J., Gu, D., & Hu, H. (2018, April). Indoor relocalization in challenging environments with dual-stream convolutional neural networks. *IEEE Transactions on Automation Science and Engineering*, *15*(2), 651-662. doi: 10.1109/TASE.2017.2664920

Li, T., Chang, X., Wu, Z., Li, J., Shao, G., Deng, X., … others (2017). Autonomous collision-free navigation of microvehicles in complex and dynamically changing environments. *ACS nano*, *11*(9), 9268–9275.

Liang, C., Tie, Y., Qi, L., & Bi, C. (2018). Deep ViDAR: CNN based 360° panoramic video system for outdoor robot visual navigation and SLAM. In *Tenth International Conference on Digital Image Processing (ICDIP 2018)* (Vol. 10806, p. 1080663).

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., … Wierstra, D. (2015). Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*.

Liu, C., Zheng, B., Wang, C., Zhao, Y., Fu, S., & Li, H. (2017). CNN-based vision model for obstacle avoidance of mobile robot. In *MATEC Web of Conferences* (Vol. 139, p. 00007).

Liu, W., Mo, Y., & Jiao, J. (2019). An efficient edge-feature constraint visual SLAM. In *Proceedings of the International Conference on Artificial Intelligence, Information Processing and Cloud Computing* (pp. 1–7).

Loncomilla, P., Ruiz-del Solar, J., & Martínez, L. (2016). Object recognition using local invariant features for robotic applications: A survey. *Pattern Recognition*, *60*, 499–514.

Long, M., Cao, Y., Wang, J., & Jordan, M. I. (2015). Learning transferable features with deep adaptation networks. *arXiv preprint arXiv:1502.02791*.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, *60*(2), 91–110. Retrieved from http://dx.doi.org/10.1023/B:VISI.0000029664.99615.94 doi: 10.1023/B:VISI.0000029664.99615.94

Lu, G., Yan, Y., Ren, L., Song, J., Sebe, N., & Kambhamettu, C. (2015, Dec). Localize Me Anywhere, Anytime: A Multi-task Point-Retrieval Approach. In *2015 IEEE International Conference on Computer Vision (ICCV)* (p. 2434-2442). doi: 10.1109/ICCV.2015.280

Lu, Y., & Lu, G. (2019, Sep.). Deep Unsupervised Learning for Simultaneous Visual Odometry and Depth Estimation. In *2019 IEEE International Conference on Image Processing (ICIP)* (p. 2571-2575). doi: 10.1109/ICIP.2019.8803247

Ly, L., & Tsai, Y. R. (2019). Autonomous exploration, reconstruction, and surveillance of 3d environments aided by deep learning. In *2019 International Conference on Robotics and Automation (ICRA)* (pp. 5467–5473).

Ma, L., Chen, J., & Liu, Y. (2019). Using RGB Image as Visual Input for Mapless Robot Navigation. *arXiv preprint arXiv:1903.09927*.

Ma, X., Wang, H., & Geng, J. (2016). Spectral–spatial classification of hyperspectral image based on deep auto-encoder. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *9*(9), 4073–4085.

Maier, D., Bennewitz, M., & Stachniss, C. (2011). Self-supervised obstacle detection for humanoid navigation using monocular vision and sparse laser data. In *2011 IEEE International Conference on Robotics and Automation* (pp. 1263–1269).

Mancini, M., Bulò, S. R., Ricci, E., & Caputo, B. (2017). Learning deep nbnn representations for robust place categorization. *IEEE Robotics and Automation Letters*, *2*(3), 1794–1801.

Mancini, M., Costante, G., Valigi, P., & Ciarfuglia, T. A. (2016). Fast robust monocular depth estimation for obstacle detection with fully convolutional networks. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 4296–4303).

Mansouri, S. S., Karvelis, P., Kanellakis, C., Kominiak, D., & Nikolakopoulos, G. (2019). Vision-based mav navigation in underground mine using convolutional neural network. In *IECON 2019-45th Annual Conference of the IEEE Industrial Electronics Society* (Vol. 1, pp. 750–755).

Matas, J., James, S., & Davison, A. J. (2018). Sim-to-real reinforcement learning for deformable object manipulation. *arXiv preprint arXiv:1806.07851*.

Mehryar, M., Rostamizadeh, A., & Talwalkar, A. (2012). Foundations of machine learning. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *17*.

Menegatti, E., Pretto, A., Scarpa, A., & Pagello, E. (2006, June). Omnidirectional vision scan matching for robot localization in dynamic environments. *IEEE Transactions on Robotics*, *22*(3), 523-535. doi: 10.1109/TRO.2006.875495

Meng, L., Chen, J., Tung, F., Little, J. J., Valentin, J., & de Silva, C. W. (2017, Sep.). Backtracking regression forests for accurate camera relocalization. In *2017 ieee/rsj international conference on intelligent robots and systems (iros)* (p. 6886-6893). doi: 10.1109/IROS.2017.8206611

Messina, R., & Louradour, J. (2015). Segmentation-free handwritten Chinese text recognition with LSTM-RNN. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)* (pp. 171–175).

Michel, D., & McIsaac, K. (2012). New path planning scheme for complete coverage of mapped areas by single and multiple robots. In *2012 IEEE International Conference on Mechatronics and Automation* (pp. 1233–1240).

Michelson, R. C. (2000). Autonomous navigation. *Access Science*. doi: 10.1036/1097-8542.YB000130

Milz, S., Arbeiter, G., Witt, C., Abdallah, B., & Yogamani, S. (2018). Visual slam for automated driving: Exploring the applications of deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 247–257).

Minaei-Bidgoli, B., Kashy, D. A., Kortemeyer, G., & Punch, W. F. (2003). Predicting student performance: an application of data mining methods with an educational web-based system. In *33rd Annual Frontiers in Education, 2003. FIE 2003* (Vol. 1, pp. T2A–13).

Mishkin, D., Radenovic, F., & Matas, J. (2018). Repeatability is not enough: Learning affine regions via discriminability. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 284–300).

Mitchell, T. M. (1997). *Machine learning*. McGraw-hill New York.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., & Riedmiller, M. (2013). Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*.

Mohedano, E., McGuinness, K., O'Connor, N. E., Salvador, A., Marques, F., & Giro-i Nieto, X. (2016). Bags of local convolutional features for scalable instance search. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval* (pp. 327–331).

Moolan-Feroze, O., & Calway, A. (2018). Predicting Out-of-View Feature Points for Model-Based Camera Pose Estimation. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (p. 82-88). doi: 10.1109/IROS.2018.8594297

Moolan-Feroze, O., Karachalios, K., Nikolaidis, D. N., & Calway, A. (2019). Improving drone localisation around wind turbines using monocular model-based tracking. In *2019 International Conference on Robotics and Automation (ICRA)* (p. 7713-7719). doi: 10.1109/ICRA.2019.8794156

Mostajabi, M., Yadollahpour, P., & Shakhnarovich, G. (2015). Feedforward semantic segmentation with zoom-out features. In *Proceedings of the*

IEEE conference on computer vision and pattern recognition (pp. 3376–3385).

Muhammad, N., Fofi, D., & Ainouz, S. (2009). Current state of the art of vision based SLAM. In *Image Processing: Machine Vision Applications II* (Vol. 7251, p. 72510F).

Mukasa, T., Xu, J., & Stenger, B. (2017). 3D scene mesh from CNN depth predictions and sparse monocular SLAM. In *Proceedings of the IEEE International Conference on Computer Vision Workshops* (pp. 921–928).

Mukherjee, D., Wu, Q. J., & Wang, G. (2015). A comparative experimental study of image feature detectors and descriptors. *Machine Vision and Applications*, *26*(4), 443–466.

Mukhija, P., Tourani, S., & Krishna, K. M. (2012). Outdoor intersection detection for autonomous exploration. In *2012 15th International IEEE Conference on Intelligent Transportation Systems* (pp. 218–223).

Murillo, A., Guerrero, J., & Sagues, C. (2007, april). SURF features for efficient robot localization with omnidirectional images. In *Robotics and Automation, 2007 IEEE International Conference on* (p. 3901 -3907). doi: 10.1109/ROBOT.2007.364077

Murillo, A. C., Singh, G., Kosecká, J., & Guerrero, J. J. (2013). Localization in urban environments using a panoramic gist descriptor. *IEEE Transactions on Robotics*, *29*(1), 146–160.

Narudin, F. A., Feizollah, A., Anuar, N. B., & Gani, A. (2016). Evaluation of machine learning classifiers for mobile malware detection. *Soft Computing*, *20*(1), 343–357.

Naseer, T., & Burgard, W. (2017, Sep.). Deep regression for monocular camera-based 6-DoF global localization in outdoor environments. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (p. 1525-1530). doi: 10.1109/IROS.2017.8205957

Neto, A. M. (2015, June). Short-term visual mapping and robot localization based on learning classifier systems and self-organizing maps. In *2015 IEEE Intelligent Vehicles Symposium (IV)* (p. 235-240). doi: 10.1109/IVS.2015.7225692

Ng, J. Y. H., Yang, F., & Davis, L. S. (2015). Exploiting local features from deep networks for image retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 53–61).

Ngiam, J., Chen, Z., Koh, P. W., & Ng, A. Y. (2011). Learning deep energy models.

Noh, H., Araujo, A., Sim, J., & Han, B. (2016). Image retrieval with deep local features and attention-based keypoints. *CoRR*.

O'Hara, S., & Draper, B. A. (2011). Introduction to the bag of features paradigm for image classification and retrieval. *arXiv preprint arXiv:1101.3354*.

Okuyama, K., Kawasaki, T., & Kroumov, V. (2011). Localization and position correction for mobile robot using artificial visual landmarks. In *Advanced Mechatronic Systems (ICAMechS), 2011 International Conference on* (p. 414-418).

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: a holistic representation of the spatial envelope. In *International journal of computer vision, vol. 42(3): 145-175.*

Otte, S., Weiss, C., Scherer, T., & Zell, A. (2016). Recurrent neural networks for fast and robust vibration-based ground classification on mobile robots. In *2016 ieee international conference on robotics and automation (icra)* (pp. 5603–5608).

Özdemir, V., & Hekim, N. (2018). Birth of industry 5.0: Making sense of big data with artificial intelligence,"the internet of things" and next-generation technology policy. *Omics: a journal of integrative biology*, *22*(1), 65–76.

Pak, M., & Kim, S. (2017). A review of deep learning in image recognition. In *2017 4th international conference on computer applications and information processing technology (CAIPT)* (pp. 1–3).

Parker, L. E. (2000). Current state of the art in distributed autonomous mobile robotics. In *Distributed Autonomous Robotic Systems 4* (pp. 3–12). Springer.

Parra, C., Cebollada, S., Payá, L., Holloway, M., & Reinoso, O. (2020). A Novel Method to Estimate the Position of a Mobile Robot in Underfloor Environments Using RGB-D Point Clouds. *IEEE Access*, *8*, 9084–9101.

Pawgasame, W. (2016). A survey in adaptive hybrid wireless sensor network for military operations. In *2016 Second Asian Conference on Defence Technology (ACDT)* (pp. 78–83).

Payá, L., Gil, A., & Reinoso, O. (2017). A State-of-the-Art Review on Mapping and Localization of Mobile Robots Using Omnidirectional Vision Sensors. *Journal of Sensors*, *2017*.

Payá, L., Peidró, A., Amorós, F., Valiente, D., & Reinoso, O. (2018). Modeling Environments Hierarchically with Omnidirectional Imaging and Global-Appearance Descriptors. *Remote Sensing*, *10*(4), 522.

Payá, L., Reinoso, O., Berenguer, Y., & Úbeda, D. (2016). Using omnidirectional vision to create a model of the environment: A comparative evaluation of global-appearance descriptors. *Journal of Sensors*, *2016*.

Peretroukhin, V., Clement, L., & Kelly, J. (2017, May). Reducing drift in visual odometry by inferring sun direction using a Bayesian Convolutional Neural Network. In *2017 IEEE International Conference on Robotics and Automation (ICRA)* (p. 2035-2042). doi: 10.1109/ICRA.2017.7989235

Pfeiffer, J., Broscheit, S., Gemulla, R., & Göschl, M. (2018). A neural autoencoder approach for document ranking and query refinement in pharmacogenomic information retrieval..

Polvara, R., Sharma, S., Wan, J., Manning, A., & Sutton, R. (2018). Obstacle avoidance approaches for autonomous navigation of unmanned surface vehicles. *Journal of Navigation*, *71*(1), 241–256. doi: 10.1017/S0373463317000753

Puthussery, A. R., Haradi, K. P., Erol, B. A., Benavidez, P., Rad, P., & Jamshidi, M. (2017). A deep vision landmark framework for robot navigation. In *2017 12th system of systems engineering conference (SoSE)* (pp. 1–6).

Rabbath, M., Sandhaus, P., & Boll, S. (2012). Analysing facebook features to support event detection for photo-based facebook applications. In *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval* (p. 11).

Rahman, M. S., Park, Y., & Kim, K. D. (2012). RSS-based indoor localization algorithm for wireless sensor network using generalized regression neural network. *Arabian journal for science and engineering*, *37*(4), 1043–1053.

Rahmatizadeh, R., Abolghasemi, P., Behal, A., & Bölöni, L. (2016). Learning real manipulation tasks from virtual demonstrations using LSTM. *arXiv preprint arXiv:1603.03833*.

Ray, S. (2018, Aug). *History of AI, in Towards Data Science.* https://towardsdatascience.com/history-of-ai-484a86fc16ef.

Reinoso, O., & Payá, L. (2020a). Special Issue on Mobile Robots Navigation. *Applied Sciences*, *10*(4). Retrieved from https://www.mdpi.com/2076-3417/10/4/1317 doi: 10.3390/app10041317

Reinoso, O., & Payá, L. (2020b). Special Issue on Visual Sensors. *Sensors*, *20*(3). Retrieved from https://www.mdpi.com/1424-8220/20/3/910 doi: 10.3390/s20030910

Richter, C., & Roy, N. (2017). Safe visual navigation via deep learning and novelty detection.

Rituerto, A., Murillo, A. C., & Guerrero, J. (2014). Semantic labeling for indoor topological mapping using a wearable catadioptric system. *Robotics and Autonomous Systems*, *62*(5), 685–695.

Romdhani, S., Torr, P., Scholkopf, B., & Blake, A. (2001). Computationally efficient face detection. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001* (Vol. 2, pp. 695–700).

Rosten, E., & Drummond, T. (2006). Machine learning for high-speed corner detection. In *European conference on computer vision* (pp. 430–443).

Rosten, E., Porter, R., & Drummond, T. (2008). Faster and better: A machine learning approach to corner detection. *IEEE transactions on pattern analysis and machine intelligence*, *32*(1), 105–119.

Ruan, X., Ren, D., Zhu, X., & Huang, J. (2019). Mobile robot navigation based on deep reinforcement learning. In *2019 Chinese control and decision conference (CCDC)* (pp. 6174–6178).

Sak, H., Senior, A., & Beaufays, F. (2014). Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In *Fifteenth annual conference of the international speech communication association.*

Salakhutdinov, R., & Hinton, G. (2009). Deep boltzmann machines. In

Artificial intelligence and statistics (pp. 448–455).

Sarlin, P., Cadena, C., Siegwart, R., & Dymczyk, M. (2019, June). From Coarse to Fine: Robust Hierarchical Localization at Large Scale. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (p. 12708-12717). doi: 10.1109/CVPR.2019.01300

Schalkoff, R. J. (1990). *Artificial intelligence: an engineering approach*. McGraw-Hill New York.

Schleichert, H. (1970). Marvin minsky (ed.)," semantic information processing"(book review). *Theory and Decision*, *1*(2), 222.

Schmidt, T., Newcombe, R., & Fox, D. (2017, April). Self-supervised visual descriptor learning for dense correspondence. *IEEE Robotics and Automation Letters*, *2*(2), 420-427. doi: 10.1109/LRA.2016.2634089

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 815–823).

Se, S., Lowe, D. G., & Little, J. J. (2005). Vision-based global localization and mapping for mobile robots. *IEEE Transactions on robotics*, *21*(3), 364–375.

Sergeant, J., Sünderhauf, N., Milford, M., & Upcroft, B. (2015). Multimodal deep autoencoders for control of a mobile robot. In *Proc. of Australasian Conf. for Robotics and Automation (ACRA)*.

Sharma, P., Liu, H., Wang, H., & Zhang, S. (2017). Securing wireless communications of connected vehicles with artificial intelligence. In *2017 IEEE international symposium on technologies for homeland security (HST)* (pp. 1–7).

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Singh, M. K., & Parhi, D. R. (2011). Path optimisation of a mobile robot using an artificial neural network controller. *International Journal of Systems Science*, *42*(1), 107–120.

Singh, S. K., Rathore, S., & Park, J. H. (2020). Blockiotintelligence: A blockchain-enabled intelligent iot architecture with artificial intelligence. *Future Generation Computer Systems*, *110*, 721–743.

Sinha, H., Patrikar, J., Dhekane, E. G., Pandey, G., & Kothari, M. (2018, Aug). Convolutional neural network based sensors for mobile robot relocalization. In *2018 23rd international conference on methods models in automation robotics (mmar)* (p. 774-779). doi: 10.1109/MMAR.2018.8485921

Smith, C., Karayiannidis, Y., Nalpantidis, L., Gratal, X., Qi, P., Dimarogonas, D. V., & Kragic, D. (2012). Dual arm manipulation—A survey. *Robotics and Autonomous systems*, *60*(10), 1340–1353.

Smolyanskiy, N., Kamenev, A., Smith, J., & Birchfield, S. (2017). Toward low-flying autonomous MAV trail navigation using deep neural networks for environmental awareness. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 4241–4247).

Šochman, J., & Matas, J. (2009). Learning fast emulators of binary decision processes. *International Journal of Computer Vision*, *83*(2), 149–163.

Sommer, K., Kim, K., Kim, Y., & Jo, S. (2017, June). Towards accurate kidnap resolution through deep learning. In *2017 14th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI)* (p. 502-506). doi: 10.1109/URAI.2017.7992654

Stachniss, C., & Burgard, W. (2003). Exploring unknown environments with mobile robots using coverage maps. In *IJCAI* (Vol. 2003, pp. 1127–1134).

Strecha, C., Bronstein, A., Bronstein, M., & Fua, P. (2011). Ldahash: Improved matching with smaller descriptors. *IEEE transactions on pattern analysis and machine intelligence*, *34*(1), 66–78.

Su, Z., Zhou, X., Cheng, T., Zhang, H., Xu, B., & Chen, W. (2017). Global localization of a mobile robot using lidar and visual features. In *2017 IEEE International Conference on Robotics and Biomimetics (ROBIO)* (pp. 2377–2383).

Sun, L., Yan, Z., Mellado, S. M., Hanheide, M., & Duckett, T. (2018). 3DOF pedestrian trajectory prediction learned from long-term autonomous mobile robot deployment data. In *2018 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 1–7).

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., … Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1–9).

Sünderhauf, N., Shirazi, S., Dayoub, F., Upcroft, B., & Milford, M. (2015, Sep.). On the performance of ConvNet features for place recognition. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (p. 4297-4304). doi: 10.1109/IROS.2015.7353986

Tai, L., Li, S., & Liu, M. (2017a). Autonomous exploration of mobile robots through deep neural networks. *International Journal of Advanced Robotic Systems*, *14*(4), 1729881417703571.

Tai, L., & Liu, M. (2016). Mobile robots exploration through cnn-based reinforcement learning. *Robotics and biomimetics*, *3*(1), 24.

Tai, L., Paolo, G., & Liu, M. (2017b). Virtual-to-real deep reinforcement learning: Continuous control of mobile robots for mapless navigation. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 31–36).

Tang, J., Ren, Y., & Liu, S. (2017). Real-time robot localization, vision, and speech recognition on nvidia jetson tx1. *arXiv preprint arXiv:1705.10945*.

Tanzmeister, G., Thomas, J., Wollherr, D., & Buss, M. (2014). Grid-based mapping and tracking in dynamic environments using a uniform evidential environment representation. In *2014 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 6090–6095).

Tapu, R., Mocanu, B., & Zaharia, T. (2013). A computer vision system that ensure the autonomous navigation of blind people. In *2013 E-Health and Bioengineering Conference (EHB)* (pp. 1–4).

Tardif, J. P., Pavlidis, Y., & Daniilidis, K. (2008, Sept). Monocular visual odometry in urban environments using an omnidirectional camera. In *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems* (p. 2531-2538). doi: 10.1109/IROS.2008.4651205

Tateno, K., Tombari, F., Laina, I., & Navab, N. (2017). Cnn-slam: Real-time dense monocular slam with learned depth prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 6243–6252).

Theodoridis, S. (2015). *Machine learning: a Bayesian and optimization perspective*. Academic Press.

Theodoridis, S., & Koutroumbas, K. (1999). Pattern recognition and neural networks. In *Advanced Course on Artificial Intelligence* (pp. 169–195).

Thrun, S. (2002). Robotic mapping: A survey. *Exploring artificial intelligence in the new millennium*, *1*(1-35), 1.

Trujillo, L., & Olague, G. (2006). Synthesis of interest point detectors through genetic programming. In *Proceedings of the 8th annual conference on genetic and evolutionary computation* (pp. 887–894).

Unicomb, J., Ranasinghe, R., Dantanarayana, L., & Dissanayake, G. (2018, Oct). A Monocular Indoor Localiser Based on an Extended Kalman Filter and Edge Images from a Convolutional Neural Network. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (p. 1-9). doi: 10.1109/IROS.2018.8594337

Valiente, D., Payá, L., Jiménez, L. M., Sebastián, J. M., & Reinoso, O. (2018). Visual information fusion through bayesian inference for adaptive probability-oriented feature matching. *Sensors*, *18*(7), 2041.

Verdie, Y., Yi, K., Fua, P., & Lepetit, V. (2015). TILDE: a temporally invariant learned detector. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 5279–5288).

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., & Manzagol, P. A. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, *11*(Dec), 3371–3408.

Viola, P., & Jones, M. J. (2004). Robust real-time face detection. *International journal of computer vision*, *57*(2), 137–154.

Vyborny, C. J., & Giger, M. L. (1994). Computer vision and artificial intelligence in mammography. *AJR. American journal of roentgenology*, *162*(3), 699–708.

Wachs, J. P., Kölsch, M., Stern, H., & Edan, Y. (2011). Vision-based hand-gesture applications. *Communications of the ACM*, *54*(2), 60–71.

Walch, F., Hazirbas, C., Leal-Taixé, L., Sattler, T., Hilsenbeck, S., & Cremers, D. (2017, Oct). Image-Based Localization Using LSTMs for Structured Feature Correlation. In *2017 IEEE International Conference on Computer Vision (ICCV)* (p. 627-637). doi: 10.1109/ICCV.2017.75

Walker, C., Graham, P., & Philippides, A. (2017). Using deep autoencoders to investigate image matching in visual navigation. In *Conference on biomimetic and biohybrid systems* (pp. 465–474).

Wang, C., Pelillo, M., & Siddiqi, K. (2019). Dominant set clustering and pooling for multi-view 3d object recognition. *arXiv preprint arXiv:1906.01592*.

Wang, H., Yang, W., Huang, W., Lin, Z., & Tang, Y. (2018). Multi-feature Fusion for Deep Reinforcement Learning: Sequential Control of Mobile Robots. In *International Conference on Neural Information Processing* (pp. 303–315).

Wang, K., Wang, W., & Zhuang, Y. (2007). Appearance-based map learning for mobile robot by using generalized regression neural network. In *International Symposium on Neural Networks* (pp. 834–842).

Wang, Y., Bao, T., Ding, C., & Zhu, M. (2017). Face recognition in real-world surveillance videos with deep learning method. In *2017 2nd International Conference on Image, Vision and Computing (ICIVC)* (pp. 239–243).

Wei, X., Xie, C. W., Wu, J., & Shen, C. (2018). Mask-CNN: Localizing parts and selecting descriptors for fine-grained bird species categorization. *Pattern Recognition*, *76*, 704 - 714. Retrieved from http://www.sciencedirect.com/science/article/pii/S0031320317303990 doi: https://doi.org/10.1016/j.patcog.2017.10.002

Weinberger, K. Q., & Saul, L. K. (2006). Unsupervised learning of image manifolds by semidefinite programming. *International journal of computer vision*, *70*(1), 77–90.

Weinzaepfel, P., Csurka, G., Cabon, Y., & Humenberger, M. (2019, June). Visual Localization by Learning Objects-Of-Interest Dense Match Regression. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (p. 5627-5636). doi: 10.1109/CVPR.2019.00578

Wong, P., Tam, L., Li, K., & Vong, C. (2010). Engine idle-speed system modelling and control optimization using artificial intelligence. *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, *224*(1), 55–72.

Wozniak, P., Afrisal, H., Esparza, R. G., & Kwolek, B. (2018). Scene recognition for indoor localization of mobile robots using deep CNN. In *International Conference on Computer Vision and Graphics* (pp. 137–147).

Wu, H., & Qin, S. Y. (2011). An approach to robot slam based on incremental appearance learning with omnidirectional vision. *International journal of systems science*, *42*(3), 407–427.

Xing, F., Xie, Y., & Yang, L. (2015). An automatic learning-based framework for robust nucleus segmentation. *IEEE transactions on medical imaging*, *35*(2), 550–566.

Xu, S., Chou, W., & Dong, H. (2019). A Robust Indoor Localization System Integrating Visual Localization Aided by CNN-Based Image Retrieval with Monte Carlo Localization. *Sensors*, *19*(2), 249.

Yang, M., Kriegman, D. J., & Ahuja, N. (2002). Detecting faces in images: A survey. *IEEE Transactions on pattern analysis and machine intelligence*, *24*(1), 34–58.

Yang, S., Konam, S., Ma, C., Rosenthal, S., Veloso, M., & Scherer, S. (2017). Obstacle avoidance through deep networks based intermediate perception. *arXiv preprint arXiv:1704.08759*.

Yang, Y., Li, Y., Fermuller, C., & Aloimonos, Y. (2015). Robot learning manipulation action plans by "watching" unconstrained videos from the world wide web. In *Twenty-Ninth AAAI Conference on Artificial Intelligence.*

Yi, K. M., Trulls, E., Lepetit, V., & Fua, P. (2016a). Lift: Learned invariant feature transform. In *European Conference on Computer Vision* (pp. 467–483).

Yi, K. M., Verdie, Y., Fua, P., & Lepetit, V. (2016b). Learning to assign orientations to feature points. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 107–116).

Yong-guo, Z., Wei, C., & Guang-liang, L. (2012). The navigation of mobile robot based on stereo vision. In *Intelligent Computation Technology and Automation (ICICTA), 2012 Fifth International Conference on* (pp. 670–673).

Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? In *Advances in neural information processing systems* (pp. 3320–3328).

Zaki, H. F., Shafait, F., & Mian, A. (2019). Viewpoint invariant semantic object and scene categorization with RGB-D sensors. *Autonomous Robots*, *43*(4), 1005–1022.

Zhang, B., Huang, W., Gong, L., Li, J., Zhao, C., Liu, C., & Huang, D. (2015). Computer vision detection of defective apples using automatic lightness correction and weighted RVM classifier. *Journal of Food Engineering*, *146*, 143–151.

Zhang, J., Marszałek, M., Lazebnik, S., & Schmid, C. (2007). Local features and kernels for classification of texture and object categories: A comprehensive study. *International journal of computer vision*, *73*(2), 213–238.

Zhang, J., Springenberg, J. T., Boedecker, J., & Burgard, W. (2017). Deep reinforcement learning with successor features for navigation across similar environments. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 2371–2378).

Zhang, X., Su, Y., & Zhu, X. (2017). Loop closure detection for visual slam systems using convolutional neural network. In *2017 23rd international conference on automation and computing (icac)* (pp. 1–6).

Zhao, Q., Zhang, B., Lyu, S., Zhang, H., Sun, D., Li, G., & Feng, W. (2018). A CNN-SIFT hybrid pedestrian navigation method based on first-person vision. *Remote Sensing*, *10*(8), 1229.

Zheng, L., Yang, Y., & Tian, Q. (2017). SIFT meets CNN: A decade survey of instance retrieval. *IEEE transactions on pattern analysis and machine intelligence*, *40*(5), 1224–1244.

Zhong, F., Wang, S., Zhang, Z., & Wang, Y. (2018). Detect-slam: Making object detection and slam mutually beneficial. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)* (pp. 1001–1010).

Zhu, L., & Zheng, W. J. (2018). Informatics, data science, and artificial intelligence. *Jama*, *320*(11), 1103–1104.

Zhu, Y., Mottaghi, R., Kolve, E., Lim, J. J., Gupta, A., Fei-Fei, L., & Farhadi, A. (2017, May). Target-driven visual navigation in indoor scenes using deep reinforcement learning. In *2017 IEEE International Conference on Robotics and Automation (ICRA)* (p. 3357-3364). doi: 10.1109/ICRA.2017.7989381

Zhu, Z., Liang, D., Zhang, S., Huang, X., Li, B., & Hu, S. (2016). Traffic-sign detection and classification in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2110–2118).

Zivkovic, Z., Bakker, B., & Krose, B. (2005, Aug). Hierarchical map building using visual landmarks and geometric constraints. In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems* (p. 2480-2485). doi: 10.1109/IROS.2005.1544951

*Highlights (for review)

- Review of relevant works in mobile robotics by AI techniques and visual information.
- The review concentrates mostly on Deep Learning tools for mobile robotics.
- Relevant AI methods to describe the visual information (local and holistic).
- AI frameworks to solve mapping, localization, SLAM, navigation and exploration.

***Declaration of Interest Statement**

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: