Visual Simultaneous Localization and Mapping: A Survey

Article *in* Artificial Intelligence Review · November 2015

DOI: 10.1007/s10462-012-9365-8

3 authors:

Jorge Fuentes-Pacheco
Centro Nacional de Investigación y Desarroll…
2 PUBLICATIONS **106** CITATIONS

SEE PROFILE

Jose Ruiz Ascencio
Centro Nacional de Investigación y Desarroll…
15 PUBLICATIONS **119** CITATIONS

SEE PROFILE

J. M. Rendon-Mancha
Universidad Autónoma del Estado de Morelos
18 PUBLICATIONS **148** CITATIONS

SEE PROFILE

Artif Intell Rev
DOI 10.1007/s10462-012-9365-8

Visual simultaneous localization and mapping: a survey

Jorge Fuentes-Pacheco · José Ruiz-Ascencio · Juan Manuel Rendón-Mancha

Abstract Visual SLAM (simultaneous localization and mapping) refers to the problem of using images, as the only source of external information, in order to establish the position of a robot, a vehicle, or a moving camera in an environment, and at the same time, construct a representation of the explored zone. SLAM is an essential task for the autonomy of a robot. Nowadays, the problem of SLAM is considered solved when range sensors such as lasers or sonar are used to built 2D maps of small static environments. However SLAM for dynamic, complex and large scale environments, using vision as the sole external sensor, is an active area of research. The computer vision techniques employed in visual SLAM, such as detection, description and matching of salient features, image recognition and retrieval, among others, are still susceptible of improvement. The objective of this article is to provide new researchers in the field of visual SLAM a brief and comprehensible review of the state-of-the-art.

Keywords Visual SLAM · Salient feature selection · Image matching · Data association · Topological and metric maps

1 Introduction

The problem of autonomous navigation of mobile robots is divided into three main areas: localization, mapping and path planning (Cyrill 2009). Localization consists in determining in an exact manner the current pose of the robot in an environment. Mapping integrates the

J. Fuentes-Pacheco · J. Ruiz-Ascencio (✉)
Centro Nacional de Investigación y Desarrollo Tecnológico, Cuernavaca, Morelos, México
e-mail: josera@cenidet.edu.mx

J. Fuentes-Pacheco
e-mail: jorge_fuentes@cenidet.edu.mx

J. M. Rendón-Mancha
Universidad Autónoma del Estado de Morelos, Cuernavaca, Morelos, México
e-mail: rendon@uaem.mx

Published online: 13 November 2012

partial observations of the surroundings into a single consistent model and path planning determines the best route in the map to navigate through the environment.

Initially, mapping and localization were studied independently, later it was recognized that they are dependent. This means that, for being precisely localized in an environment, a correct map is necessary, but in order to construct a good map it is necessary to be properly localized when elements are added to the map. Currently, this problem is known as *Simultaneous Localization and Mapping* (SLAM). When cameras are employed as the only exteroceptive sensor, it is called *visual SLAM*. The terms *vision-based SLAM* (Se et al. 2005; Lemaire et al. 2007) or *vSLAM* (Solà 2007) are also used. In this article the term visual SLAM is used because it is the best known. Visual SLAM systems can be complemented with information from proprioceptive sensors, with the aim of increasing accuracy and robustness. This approach is known as *visual-inertial SLAM* (Jones and Soatto 2011). However, when vision is used as the only system of perception (without making use of information extracted from the robot odometry or inertial sensors) it can be called *vision-only SLAM* (Paz et al. 2008; Davison et al. 2007) or *camera-only SLAM* (Milford and Wyeth 2008).

Many visual SLAM systems fail when work under the following conditions: in external environments, in dynamic environments, in environments with too many or very few salient features, in large scale environments, during erratic movements of the camera and when partial or total occlusions of the sensor occur. A key to a successful visual SLAM system is the ability to operate correctly despite these difficulties.

Important applications of SLAM are oriented towards automatic car piloting on unrehearsed off-road terrains (Thrun et al. 2005a); rescue tasks for high-risk or difficult-navigation environments (Thrun 2003; Piniés et al. 2006); planetary, aerial, terrestrial and oceanic exploration (Olson et al. 2007; Artieda et al. 2009; Steder et al. 2008; Johnson et al. 2010); augmented reality applications where virtual objects are included in real-world scenes (Chekhlov et al. 2007; Klein and Murray 2007); visual surveillance systems (Mei et al. 2011); medicine (Auat et al. 2010; Grasa et al. 2011), and so forth.

In this article a detailed study of visual SLAM is presented, as well as the most recent contributions and diverse current problems. Previously, Durrant and Bailey presented a tutorial divided into two parts that summarizes the SLAM problem (Durrant and Bailey 2006; Bailey and Durrant 2006). The latter tutorial describes works that are centered on the use of laser range-finder sensors, building 2D maps under a probabilistic approach. Similarly, Thrun and Leonard (2008) presented an introduction to the SLAM problem, analyzed three paradigms of solution (the first is based on the Extended Kalman Filter, and the other two use optimization techniques based on graphs and particle filters) and proposed a taxonomy of the problem. Nevertheless, the above-mentioned articles are not focused on methods using vision as the only external sensor. On the other hand, Kragic and Vincze (2009) present a review of computer vision for robotics in a general context, considering the visual SLAM problem but not in detail as it is intended in this article.

This article is structured in the following way: Sect. 2 describes the SLAM problem in general. In Sect. 3, the use of cameras as the only external sensor is discussed and the weak points of such systems are mentioned. Section 4 describes the type of salient features that can be extracted and the descriptors used to achieve invariance to various transformations that the images may suffer. Section 5 deals with image matching and the data association problem. Section 6 gives a detailed review of the different methods to solve the visual SLAM problem and weaknesses and strengths of each one are discussed. The different ways of representing the observed world are described on Sect. 7. Section 8 provides conclusions and potential problems for further investigations. Final section presents bibliographic references.

2 Simultaneous localization and mapping

During the period of 1985–1990, Chatila and Laumond (1985) and Smith et al. (1990) proposed carrying out mapping and localization in a concurrent manner. Sometime later, this problem received the name of SLAM (simultaneous localization and mapping). The reader may refer to the tutorial of Durrant and Bailey (2006), Bailey and Durrant (2006) for a detailed description of the history of the SLAM problem. In some publications of Newman et al. (2002) and Andrade and Sanfeliu (2002) it is also known as CML (Concurrent Mapping and Localization). SLAM or CML is the process whereby an entity (robot, vehicle or even a central processing unit with sensor devices carried by a person) has the capacity for building a global map of the visited environment and, at the same time, utilizing this map to deduce its own location at any moment.

In order to build a map from the environment, the entity must possess sensors that allow it to perceive and obtain measurements of the elements from the surrounding world. These sensors are classified into *exteroceptive* and *proprioceptive*. Among the exteroceptive sensors it is possible to find: sonar (Tardós et al. 2002; Ribas et al. 2008), range lasers (Nüchter et al. 2007; Thrun et al. 2006), cameras (Se et al. 2005; Lemaire et al. 2007; Davison 2003; Bogdan et al. 2009) and global positioning systems (GPS) (Thrun et al. 2005a). All of these sensors are noisy and have limited range capabilities. In addition, only local views of the environment can be obtained using the first three aforementioned sensors. Laser sensors and sonar allow precise and very dense information of the environment structure. Nevertheless, they have the following problems: not useful in highly cluttered environments or for recognizing objects; both are expensive, heavy and consist of large pieces of equipment, making their use difficult for airborne robots or humanoids. On the other hand, a GPS sensor does not work well in narrow streets (urban canyons), under water, on other planets, and occasionally is not available indoors.

Proprioceptive sensors allow the entity to obtain measurements like velocity, position change and acceleration. Some examples are: encoders, accelerometers and gyroscopes. These allow obtaining an incremental estimate of the entity's movements by means of a *dead-reckoning* navigation method (also known as *deduced-reckoning*), but due to their inherent noise they are not sufficient to have an accurate estimation of the entity's position all the time, since errors are cumulative.

As has been demonstrated in some investigations (Castellanos et al. 2001; Majumder et al. 2005; Nützi et al. 2010), to maintain an accurate and robust estimation of the robot position it is needed to use the *fusion of information* from multiple sensors of perception. However, the addition of sensors increases the cost, weight and power requirements of a system; therefore, it is important to investigate how an entity may locate itself and create a map with only cameras.

3 Cameras as the only exteroceptive sensors

In the last 10 years, published articles reflect a clear tendency for using vision as the only external sensorial perception system to solve the problem of SLAM (Paz et al. 2008; Davison et al. 2007; Klein and Murray 2007; Sáez and Escolano 2006; Piniés and Tardós 2008). The main reason for this tendency is attributed to the capability for a system based on cameras to obtain range information, and also retrieving the environment's appearance, color and texture, giving a robot the possibility of integrating other high-level tasks like detection and recognition of people and places. Furthermore, cameras are less expensive, lighter and have

lower power consumption. Unfortunately, there might be errors in the data due to the following reasons: insufficient camera resolution, lighting changes, surfaces with lack of texture, blurred images due to fast movements, among other factors.

The first works on visual navigation were based on a binocular stereo configuration (Se et al. 2002; Olson et al. 2003). However,-in many cases it is difficult to have a device with binocular or trinocular stereo cameras due to their high costs. An alternative is to use a pair of monocular cameras (for example webcams), which leads to consider different aspects such as: (a) the camera synchronization through the use of hardware or software, (b) the different responses of each CCD sensor to color and luminance, and (c) the mechanical alignment according to the geometry scheme chosen (parallel or convergent axes).

Works also exist that make use of multi-camera rigs with or without overlapping between the views (Kaess and Dellaert 2010; Carrera et al. 2011) and cameras with special lens such as wide-angle (Davison et al. 2004) or omnidirectional (Scaramuzza and Siegwart 2008) with the goal of increasing visual range and thus decrease, to some extent, the cumulative error of pose estimation. Recently, RGB-D (color images and depth maps) sensors have been used to map indoor environments (Huang et al. 2011), proving to be a promising alternative for SLAM applications.

Independently of the configuration used, cameras have to be calibrated (manually off-line or automatically on-line). *Calibration* estimates *intrinsic* and *extrinsic parameters*, the first depend on the camera's geometry (focal length and principal point), while the second depend on the camera's position in space (rotation and translation with respect to some coordinate system). The necessary parameters are usually estimated from a set of images that contain multiple views of a checkerboard calibration pattern, to relate the image's coordinates with the real-world coordinates (Hartley and Zisserman 2003). Many tools exist to execute the process of calibration, some of them are: the calibration functions of OpenCV (2009) (based on the Zhang algorithm (Zhang 2000)), Camera Calibration Toolbox for Matlab (Bouguet 2010), Tsai Camera Calibration Software (Willson 1995), OCamCalib Toolbox for omnidirectional cameras (Scaramuzza 2011), and Multi-Camera Self-Calibration to calibrate several cameras (at least 3) (Svoboda 2011).

If the camera calibration is performed off-line, then it is assumed that the intrinsic properties of the camera will not change during the entire period of operation of the SLAM system. This is the most popular option, since it reduces the number of parameters calculated online. Nevertheless, the intrinsic camera information may change due to some environmental factors of the environment, such as humidity or temperature. Furthermore, a robot that works in real world conditions can be hit or damaged, which could invalidate the previously acquired calibration (Koch et al. 2010).

Stereo configurations (binocular, trinocular or multiple cameras with their fields of vision partially overlapped) offer the advantage of being able to easily and accurately calculate the real 3D positions of the landmarks contained in the scene, by means of *triangulation* (Hartley and Sturm 1997), which is information of great utility in the visual SLAM problem. The works of Konolige and Agrawal (2008), Konolige et al. (2009), Mei et al. (2009) represent the most current and effective binocular stereo SLAM systems. When localization and mapping is being done with a single camera, the map will suffer from a scale ambiguity problem (Nistér 2004; Strasdat et al. 2010a). To obtain 3D information from a single camera, two cases exist depending on the a priori knowledge of the camera. These are: (a) with the knowledge of the intrinsic parameters only; with this alternative the environment structure and the extrinsic parameters are recovered with an undetermined scale-factor. Scale is determined if the real distance between two points in space is known; and (b) where only correspondences are known; in this case, the reconstruction is made up to a projective transformation.

The idea of utilizing one camera has become popular since the emergence of *single camera SLAM* or *MonoSLAM* (Davison 2003). This is probably also because it is now easier to access a single camera than a stereo pair, through cell phones, personal digital assistants or personal computers. This monocular approach offers a very simple, flexible and economic solution in terms of hardware and processing times.

Monocular SLAM is a particular case of *bearing-only SLAM*. The latter is a partially observable problem, where sensors do not provide sufficient information from a simple observation to determine the depth of a landmark. This causes a landmark initialization problem, where solutions can be divided into two categories: *delayed* and *undelayed* (Lemaire et al. 2007; Vidal et al. 2007). A salient feature tracking across multiple observations has to be performed to obtain tridimensional information from a single camera.

Even though many contributions have been made to visual SLAM, there are still many problems. The solutions proposed for the visual SLAM problem are reviewed in Sect. 6. Many visual SLAM systems suffer from large accumulated errors while the environment is being explored (or fail completely in visually complex environments), which leads to inconsistent estimates of robot position and totally incongruous maps. Three primary reasons exist:

(1) First, generally it is assumed that camera movement is smooth and that there will be consistency in the appearance of salient features (Davison 2003; Nistér et al. 2004), but in general this is not true. The above assumptions are highly related to the selection of the salient feature detector and of the matching technique used. This originates an inaccuracy in camera position when capturing images with little texture or that are blurred due to rapid movements of the sensor (e.g. due to vibration or quick direction changes) (Pupilli and Calway 2006). These phenomena are typical when the camera is carried by a person, humanoid robots, and quad-rotor helicopters, among others. One way of alleviating this problem to some extent is by the use of *keyframes* (see "Appendix I") (Mouragnon et al. 2006; Klein and Murray 2008). Alternatively, Pretto et al. (2007) and Mei and Reid (2008) analyze the problem of visual tracking in real time over blurred image sequences due to an out-of-focus camera.
(2) Second, most of researchers assume that the environments to explore are static and that they only contain stationary and rigid elements; the majority of the environments contain people and objects in motion. If this is not considered, the moving elements will originate false matches and consequently will generate unpredictable errors in all the system. The first approaches to this problem are proposed by Wang et al. (2007); Wangsiripitak and Murray (2009); Migliore et al. (2009), as well as Lin and Wang (2010).
(3) Third, the world is visually repetitive. There are many similar textures, such as the repeated architectural elements, foliage and walls of brick or stone. Also some objects such as traffic signals appear repeatedly within an urban outdoor environment. This makes it difficult to recognize a previously explored area and also to do SLAM on large extensions of land.

4 Salient feature selection

We will make a difference between *salient features* and *landmarks*, since in some articles they are treated indistinctly. According to Frintrop and Jensfelt (2008), a landmark is a region in the real world described by 3D position and appearance information. On the other hand, a salient feature is a region of the image described by its 2D position (on the image) and an

appearance. In this survey, the term salient feature is used as a generalization that can include points, regions, or even edge segments which are extracted from images.

The salient features that are easiest to locate, are those produced by artificial landmarks (Frintrop and Jensfelt 2008). These landmarks are added intentionally to the environment with the purpose of serving as an aid for navigation, e.g. squares or circles situated on the floor or walls. These landmarks have the advantage that their appearance is known in advance, making them easy to detect at any time. However, the environment has to be prepared by a person before the system is initialized. Natural landmarks are those that exist habitually in the environment (Se et al. 2002). For indoor environments it is common to use as natural landmarks the corners of doors or windows. In outdoor environments, tree trunks (Asmar 2006), regions (Matas et al. 2002), or interest points (Lowe 2004) are used. An *interest point* is an image pixel with such a neighborhood that is easy to distinguish from other points using a given detector.

One good-quality feature has the following properties: it must be notable (easy to extract), precise (it may be measured with precision) and invariant to rotation, translation, scale and illumination changes (Lemaire et al. 2007). Therefore, a good-quality landmark has a similar appearance from different viewpoints in 3D space. The salient feature extraction process is composed of two phases: *detection* and *description*. The detection consists in processing the image to obtain a number of salient elements. The description consists in building a feature vector based on visual appearance in the image. The invariance of the descriptor to changes in position and orientation will permit to improve the image matching and data association processes (described in Sect. 5).

4.1 Detectors

In the majority of SLAM systems based on vision, natural features present everywhere have been used, such as corners, interest points, edge segments and regions. The selection of the type of features to be used will depend largely on the environment in which the robot is going to work.

There is a large number of salient feature detectors. Some examples are: Harris corners detector (Harris and Stephens 1988), Harris-Laplace and Hessian-Laplace points detectors, as well as their respective affine invariants versions Harris-Affine and Hessian-Affine (Mikolajcczyk and Schmid 2002); Difference of Gaussians (DoG) used on SIFT (Scale Invariant Feature Transform) (Lowe 2004); Maximally Stable Extremal Regions (MSERs) (Matas 2002), FAST (Features from Accelerated Segment Test) (Rosten and Drummond 2006) and the Fast-Hessian used on SURF (Speeded Up Robust Features) (Bay et al. 2006). Mikolajczyk et al. (2005) made an evaluation of the performance of these algorithms with respect to viewpoint, zoom, rotation, out-of-focus, JPEG compression and lighting changes. The Hessian-Affine and MSER detectors had the best performance, MSER was the most robust with respect to viewpoint and lighting changes, and the Hessian-Affine was the best in presence of out-of-focus features and JPEG compression. In (Tuytelaars and Mikolajczyk 2008) these detectors and some others are classified, taking into consideration their repeatability, precision, robustness, efficiency and invariant characteristics.

The majority of visual SLAM systems use corners as landmarks due to their invariant features and their wide study in the computer vision context. However, Eade and Drummond (2006a) propose to use edge segments called edgelets in a real-time MonoSLAM system, allowing the construction of maps with high levels of geometrical information. The authors demonstrated that edges are good features for tracking and SLAM, due to their invariance to

lighting, orientation and scale changes. The use of edges as features looks promising, since edges are little affected by blurring caused by the sudden movements of the camera (Klein and Murray 2008). Nonetheless, edges have the limitation of not being easy to extract and match. On the other hand, Gee et al. (2008) and Martinez and Calway (2010) investigate the fusion of features (i.e. points, lines and planar structures) in a single map, with the purpose of increasing the precision of SLAM systems and creating a better representation of the environment.

4.2 Descriptors

One of the most commonly used descriptors for object recognition is the histogram-type SIFT descriptor, proposed by Lowe (2004), which is based on the spatial distribution of local features in the neighborhood of the salient point, obtaining a vector of 128 components. Ke and Sukthankar (2004) propose a modification to SIFT called PCA-SIFT, whose main idea is to obtain a descriptor as distinctive and robust as SIFT but with a vector of less components. The reduction is accomplished by means of the Principal Component Analysis technique. The histogram-type descriptors have the property of being invariant to translation, rotation, and scale, and partially invariant to lighting and viewpoint changes. An exhaustive evaluation of several description algorithms and a proposal for an extension of the SIFT descriptor (Gradient Location-Orientation Histogram-GLOH) may be found in (Moreels and Perona 2005) and (Mikolajcczyk and Schmid 2005), respectively.

In (Gil et al. 2009) appears a comparative study of different local description algorithms focused on the visual SLAM problem. The evaluation is based on the number of correct and incorrect matches found through video sequences with significant changes in scale, viewpoint and lighting. In this work it is demonstrated that SURF descriptor is superior to SIFT descriptor in terms of robustness and computation time. Further on, the authors manifest that SIFT does not demonstrate great stability, which means that many of the landmarks detected from a particular position of the camera disappear when moving it slightly. Currently there are many variants that improve the performance of the SIFT algorithm, for example: ASIFT, which incorporates invariance to affine transformations (Morel and Yu 2009), BRIEF (Binary Robust Independent Elementary Features) (Calonder et al. 2010); ORB, a fast binary descriptor based on BRIEF but rotation invariant and resistant to noise (Rublee et al. 2011); PIRF (Position-Invariant Robust Feature) (Kawewong et al. 2010) and GPU-SIFT, an implementation of SIFT on a GPU (Graphics Processing Unit) in order to make processing in parallel and in real time (Sinha et al. 2006).

5 The image matching and data association problems

In stereo correspondence, the *image matching* consists in searching for each element in one image, its correspondent in the other image. The matching techniques can be divided into two categories: *short baseline* and *long baseline*. These techniques are necessary during the stage of salient features tracking and loop closure detection in visual SLAM. In the area of robot navigation, the *data association* consists in relating the sensor's measurements with the elements already inside the robot's map (Neira and Tardós 2001). This problem also involves determining if the measurements are spurious or belong to elements not contained in the map. An efficient image matching and a correct data association are essential to a successful navigation. The errors will rapidly lead to incorrect maps.

5.1 Short baseline matching

Baseline is the line separating the optical centers of two cameras used to capture a pair of images. When the difference between the images taken from different viewpoint is small, the corresponding point will have almost the same position and appearance in both images, reducing the complexity of the problem. In this case, the point is characterized simply by the intensity values of a set of sampled pixels from a rectangular window (also known as *patch*) that is centered over the salient feature. The intensity values of the pixels are compared by means of correlation measures like cross correlation, sum of squared differences and sum of absolute differences, among others. In (Ciganek and Siebert 2009) there is a list of formulae to determine the similarity between two patches. Articles (Konolige and Agrawal 2008; Nistér et al. 2004) manifest that the measure of normalized crossed correlation (NCC) is the one exhibiting the best results. Normalization makes this method invariant to uniform changes in brightness. In (Davison 2003) and (Molton et al. 2004) an homography is calculated to deform the patch and make the correspondences with NCC invariant to viewpoints, which allows more camera movement freedom. Unfortunately, the correspondence with NCC is susceptible to false positives and false negatives. In an image region with repeated texture, two o more points inside a search region can get a strong response to NCC.

For short baseline correspondences, it is important to take into account the dimensions of the patch as well as the dimensions of the search region, otherwise errors will appear (Nistér et al. 2004). For example, patches that are too little are good for speed, but tend to generate false correspondences and too large patches require more processing time. It is recommended to use patches of approximately 9×9 or 11×11 pixels and place the patch over a corner, since in such a region the gradient of the image has two or more dominant directions and, consequently, facilitates the process of correspondence. The use of descriptors is unnecessary for frame to frame short baseline matching, but if tracking fails and the camera is lost, then it become a great help.

A disadvantage of short baseline is that depth computation is very sensitive to noise, e.g. error measurements of image's coordinates, due to the reduced distance between different viewpoints. However, it is possible to make a precise tracking of corresponding features through video sequences. Yilmaz et al. (2006), Cannons (2008) and Lepetit and Fua (2005) present a study of the state-of-the-art of techniques to perform tracking based on features, contours or regions.

5.2 Long baseline matching

When working with long baselines, images present big changes in scale or perspective, which originates that a point in an image moves to any place in the other image. This creates a difficult correspondence problem, see Sect. 3 of Brown et al. (2003). Data from the image in the neighborhood of a point are distorted by changes in viewpoint and lighting, and the correlation measures will not give good results.

The easiest way to find correspondences is to compare all the features of an image versus all the features of some other image (approach known as "brute force"). Unfortunately, this process grows in a quadratic manner for the number of extracted features, which is impractical for many applications that must work in real time.

In recent years, there has been considerable progress in the development of matching algorithms for long baseline that are invariant to several transformations of the image. Many of these algorithms obtain a descriptor for each detected feature, calculate dissimi-

larity measures between descriptors and use data structures to perform the search of pairs quickly and efficiently.

There are several dissimilarity measures, such as Euclidean distance, Manhattan distance, Chi-Square distance, among others. The data structures can be balanced binary trees called kd-trees (Beis and Lowe 1997; Silpa and Hartley 2008) or hash tables (Grauman and Darrell 2007). There are also criteria for deciding when two features must be associated (Mikolajcczyk and Schmid 2005). Some examples are: (a) distance threshold: two features are relevant if the distance between their descriptors is below a threshold; (b) nearest neighbor: A and B are related if descriptor B is the closest neighbor of the descriptor A and if the distance between them is below a threshold, and (c) nearest neighbor distance ratio: this approach is similar to nearest neighbor except that the threshold is applied to the ratio of distances of the current pixel to the first and second nearest neighbor. By using the first criterion described above, a feature of the first image can be paired with several features of the second image and vice versa. There are different techniques to disambiguate these candidate matches, for example: by means of relaxation techniques (Zhang et al. 1994) or considering collections of points (Dufournaud et al. 2004).

A *geometric constraint* can be used to speed up matching process. The epipolar constraint establishes that: a necessary condition for x and x' to be corresponding points, is that the point x' have to be on the epipolar line of x (Hartley and Zisserman 2003). In this way the search for matches is restricted to a single line instead of the whole image. Some details may be found in Tuytelaars and Van-Gool (2004); Zhang and Kosecka (2006) and Matas et al. (2002).

Other researches like (Lepetit and Fua 2006; Grauman 2010; Kulis et al. 2009; Özuysal et al. 2010) use learning strategies to determine similarity between features. This re-formulates the problem of correspondence as a problem of classification, which seems very promising. In the specific case of SLAM applications in real time, this could not look quite adequate since it is necessary to make a constant training on-line. Nonetheless, (Hinterstoisser et al. 2009; Taylor and Drummond 2009) have proposed faster methods for achieving on-line learning, which could be utilized in the future for SLAM applications.

Aguilar et al. (2009), Li et al. (2010) and Gu et al. (2010) propose a different image correspondence approach, where points neighborhood relationships are represented by means of a graph. Correspondent graphs are those that are the same or similar in both images. In the same way, Sanromá et al. (2010) propose an iterative matching algorithm based on graphs, which is used to retrieve the pose of a mobile robot. Unfortunately, these researches are still limited because they do not work in real time and cannot handle temporary occlusions.

Using high-quality descriptors or even different kinds of similarity measures do not ensure to avoid false correspondences. If these correspondences are used inside a SLAM system, important errors will be generated for camera pose and map estimation. Therefore, it is necessary to use robust estimators as RANSAC (Random Sample Consensus), PROSAC (Progressive Sample Consensus), among others, can automatically handle false correspondences. A comparative analysis of these estimators can be found in (Raguram et al. 2008). The main difference between them is the way they evaluate the quality of a model. Robust estimators are commonly used to estimate model parameters from data containing atypical values. RANSAC estimates a global relationship adapting data, and at the same time classifies data under inliers (data which is consistent with the relationship) and outliers (not consistent with the relationship). Due to the ability of tolerating a large amount of outliers, this algorithm is a popular option to solve a large variety of estimation problems.

An alternative to RANSAC is presented by Chli and Davison (2008, 2009), which propose a Bayesian technique for frame to frame correspondence called active matching. Active matching performs a search only in parts of the image where it is most likely to find true positives, reducing the number of outliers and the number of image processing operations for processing images. This algorithm uses a guided search based on the principle of Shannon Information Theory. Active matching presents good results facing rapid camera movements. The limitation of this technique is its poor scalability when the number of features increases. To solve this problem, Handa et al. (2010) propose an extension allowing managing hundreds of features in real-time, without losing precision in the correspondence.

A way to measure the performance of matching algorithms is by means of the *Receiver Operating Characteristic curve*, or ROC curve (Fawcett 2006). This is a graphical representation involving the computation of true positives, false positives, false negatives and true negatives, plus positive predictive value and accuracy. Where true positives are the number of correct matches, false negatives are matches that were not correctly detected, false positives are matches that are incorrect and true negatives are non-matches that were correctly rejected. In some papers of the information retrieval literature (Majumder et al. 2005), the following two metrics are used: *precision* (number of correct matches divided by the total number of found correspondences) and *recall* (number of correct matches divided by the total number of expected correspondences).

5.3 Data association in visual SLAM

The data association problem in visual SLAM is supported by means of Visual Place Recognition techniques. Data association has particular cases, as: *loop closure detection*, *kidnapped robot* (or camera), and *multi-session* and *cooperative mapping*; which are described in the following lines:

5.3.1 Loop closure detection

Loop closure detection consists in recognizing a place that has already been visited in a cyclical excursion of arbitrary length (Ho and Newman 2007; Clemente et al. 2007; Mei et al. 2010). This problem has been one of the greatest impediments to perform large scale SLAM and recover from critical errors. From this problem arises another one called *perceptual aliasing* (Angeli et al. 2008; Cummins and Newman 2008); where two different places from the surrounding are recognized as the same. This represents a problem even when using cameras as sensors due to the repetitive features of the environment, e.g. hallways, similar architectural elements or zones with a large quantity of bushes. A good loop closure detection method must not return any false positive and must obtain a minimum of false negatives.

According to Williams et al. (2009) detection methods for loop closures in visual SLAM can be divided into three categories: (1) *map to map*; (2) *image to image*; and (3) *image to map*. Categories differ mainly about where the association data are taken from (metric map space or image space). However the ideal would be to build a system that combines the advantages of all three categories. Loop closure detection is an important problem for any SLAM system, and taking into account that cameras have become a very common sensor for robotic applications, many researchers focus on vision methods to solve it.

Ho and Newman (2007) propose to use a similarity matrix to code the relationships of resemblance between all the possible pairs in captured images. They demonstrate by means of a single value decomposition that it is possible to detect loop closures, despite of the pres-

ence of repetitive and visually ambiguous images. Eade and Drummond (2008) present a unified method to recover from tracking failures and detect loop closures in the problem of monocular visual SLAM in real time. They also propose a system called GraphSLAM where each node stores landmarks and maintains estimations of the transformations relating nodes. In order to detect failures or loop closures, they model appearance as a *Bag of Visual Words* (BoVW) to find the nodes that have a similar appearance in the current video image (see "Appendix II"). Angeli et al. (2008) present a method to detect loop closures under a scheme of Bayessian filtering and a method of incremental BoVW, where the probability to belong to a visited scene is computed for each acquired image. Cummins and Newman (2008) propose a probabilistic framework to recognize places, which uses only image appearance data. Through the learning of a generative model of appearance, they demonstrate that not only it is possible to compute the resemblance of two observations, but also the probability that they belong to the same place; and, thus, they calculate a probability distribution function (*pdf*) of the observed position. Finally, Mei et al. (2010) propose a new topometric representation of the world, based on co-visibility, which allows to simplify data association and improve the performance of recognition based on appearance.

All the loop closure works described above, aim to achieve a precision of 100%.This is due to the fact that a single false positive can cause irremediable failures during the creation of the map. In the context of SLAM, false positives are graver than false negatives (Magnusson et al. 2009). False negatives reduce recall percentage but have no impact on precision percentage. Thus, in order to determine the efficiency of a loop closure detector, the recall rate should be as high as possible, with a precision of 100%.

5.3.2 Kidnapped robot

In the problem of the kidnapped robot, robot pose in the map is determined without previous information of its whereabouts. This case can occur if the robot is put back into an already mapped zone, without the knowledge of its displacement while it is being transported to that place, or when robot performs blind movements due to occlusions, temporary sensor malfunction, or fast camera movements (Eade and Drummond 2008; Chekhlov et al. 2008; Williams et al. 2007).

Chekhlov et al. (2008) propose a system capable of tolerating the uncertainty about camera pose and recover from minor tracking failures generated by continuous erratic movement or by occlusions. The work consists in generating a descriptor (based on SIFT) at multiple resolutions to provide robustness in the data association task. In addition, it uses an index based on low-order coefficients of the Haar wavelet. Williams et al. (2007) present a re-localization module that monitors the SLAM system, detects tracking failures, determines the camera pose in the map landmarks framework and resumes tracking as soon as conditions have improved. Re-localization is performed by a landmark recognition algorithm using the randomized trees classifier technique proposed by Lepetit and Fua (2006) and trained online through a feature harvesting technique. In this way a high recovery rate and a rapid recognition time are obtained. To find the camera pose, candidate poses are generated from correspondences between the current frame and landmarks on the map. There is a selection of sets of three potential matches, then, all the consistent poses with these sets are calculated by a three-point algorithm. These poses are evaluated seeking consensus among the other correspondences in the image found by RANSAC. If a pose with a large consensus is found, that pose is assumed to be correct.

5.3.3 Multi-session and cooperative mapping

The multi-session and cooperative mapping consists in align two or more partial maps of the environment collected by a robot in different periods of operation or by several robots at the same time (*visual cooperative SLAM*) (Ho and Newman 2007; Gil et al. 2010; Vidal et al. 2011).

In the past, the problem of associating measurements with landmarks on the map was solved through algorithms such as Nearest Neighbor, Sequential Compatibility Nearest Neighbor and Joint Compatibility Branch and Bound (Neira and Tardós 2001). However, these techniques are similar because they work only if a good initial guess of the robot in the map is available (Cummins and Newman 2008).

6 Solutions to the visual SLAM problem

The techniques used to solve the visual SLAM problem can be divided into three main groups: (a) classic ones, based on probabilistic filters, with which the system maintains a probabilistic representation of both the pose of the robot and the location of the landmarks in the environment, (b) the techniques employing Structure from Motion (SfM) in an incremental (causal) manner, and finally (c) the techniques inspired by biology. In the following sections some details of each of these techniques are described.

6.1 Probabilistic filters

Most SLAM solutions reported to date are based on probabilistic techniques. Some of these are: the Extended Kalman Filter (EKF), Factored Solution to SLAM (FastSLAM), Maximum Likelihood (ML) and Expectancy Maximization (EM) (Thrun et al. 2005b). The first two techniques listed above are the most commonly used because they offer the best results when jointly minimize uncertainties of the entity and the map. These approaches are successful on a small scale, but have a limited capability to navigate in large environments or to add information to loop closure.

A methodology for building maps in an incremental (causal) way, was first presented in the work of Smith et al. (1990). Smith et al. (1990) introduced the concept of stochastic map and developed a precise solution to the SLAM problem using the Extended Kalman Filter. The EKF-based approach to SLAM is characterized by a state vector composed of the location of the entity and some map elements, estimated recursively from the nonlinear models of observation and transition. The uncertainty is represented by probability density functions (*pdfs*). It is supposed that the recursive propagation of the mean and covariance of these *pdfs* are close to the optimal solution. The EKF has the disadvantage of being particularly sensitive to bad associations, one incorrect measurement can lead to the divergence of the entire filter. The complexity of EKF is quadratic with respect to the number of landmarks on the map, being difficult to maintain large maps. In the literature there are different methods to reduce this complexity through techniques such as: Atlas Framework (Bosse et al. 2003), Compressed Extended Kalman Filter (CEKF) (Guivant 2002), Sparse Extended Information Filter (SEIF) (Thrun et al. 2002), Divide and Conquer in $O(n)$ given by Paz et al. (2008) or Conditionally Independent Submaps (CI-Submaps) developed by Piniés and Tardós (2008).

FastSLAM was proposed by Montemerlo et al. (2002) and later improved in (Montemerlo 2003). This method maintains an entity pose distribution as a set of Rao-Blackwellized particles, where each particle represents a trajectory of the entity, maintains its own map using

the EKF, has an hypothesis on the association of data (multiple hypotheses) and survives with a probability. The algorithm consists of a particle generation process and a re-sampling process, to avoid the degeneration of the particles over time. The computational cost of this solution is logarithmic, $O(plogn)$, where p is the number of particles used and n is the number of landmarks on the map. Their main problem is that there is no way to determine the number of particles required to accurately represent the position of the entity. That is, many particles require a lot of memory and computing time, but few particles lead to inaccurate results.

Davison (2003) was the first to present a real-time monocular probabilistic system, which he called MonoSLAM. This technique of SLAM, perform simultaneously 3D metric mapping of points and location at 30 frames per second, using only a digital firewire (IEEE-1394) camera. It considers the complete camera movement (6gdl): position (x, y, z) and orientation (pitch, yaw and roll). Davison's work has the limitation of only working in confined and indoor spaces, since it employs the EKF to estimate data.

MonoSLAM system uses a motion model with constant linear and angular velocities. This represents an inconvenient due to the inability of the model to properly dealing with sudden movements, limiting camera mobility. Therefore, the distance that the salient features can be moved between frames is very small, in order to ensure tracking (otherwise, it could turn out to be very expensive, since a large region to search for features is proposed).

To face erratic movement of the camera with MonoSLAM, Gee et al. (2008) developed an optimized version, capable of operating at 200 Hz using an extended motion model that takes into account acceleration, and linear and angular velocties; however, its performance in real time is limited to only a few seconds, because the map size and the computational cost grow extremely fast.

To increase the number of maintained landmarks on the map, Eade and Drummond (2006b) used a particle filter technique inspired by the method proposed by Montemerlo et al. (2002), FastSLAM. The method of Eade and Drummond is able to track up to 30 features per video frame and maintain dense maps of thousands of landmarks. Clemente et al. (2007) propose an alternative to use the MonoSLAM in large outdoor environments. This approach is based on a hierarchical mapping technique and a robust data association algorithm based on Geometric Constraints Branch and Bound (GCBB) capable of performing large loops closure (250 m approx.).

As mentioned above, one problem in the monocular visual SLAM is the initialization of the landmarks, because their depth cannot be calculated from a single observation. For this, Davison (2003) uses a delayed initialization technique, while Montiel et al. (2006) propose a technique called inverse depth parametrization, which performs an undelayed landmark initialization in an EKF-SLAM system from the first moment they are detected.

6.2 Structure from motion

Structure from Motion (SfM) techniques allow to compute 3D structure of the scene and camera position from a set of images (Pollefeys et al. 2004). SfM has its origins in photogrammetry and computer vision. The standard procedure (carried out off-line) is to extract salient features of incoming images, to match them and perform a non-linear optimization called *Bundle Adjustment* (BA) to minimize the re-projection error (Triggs et al. 1999; Engels et al. 2006).

SfM allows a high precision in the location of the cameras but does not necessarily intend to create consistent maps. Despite this, several proposals have been made using SFM to locate with precision while creating a good representation of the environment.

One method to solve the problem of SfM incrementally is the *visual odometry* published by Nistér et al. (2004). Visual odometry consist in determine simultaneously the camera pose for each video frame and the position of features in 3D world, using only images in a causal way and in real time. Mouragnon et al. (2006, 2009) uses a visual odometry similar to Nister's proposal, but adding a technique called Local Bundle Adjustment, reporting trajectories up to 500 m. The visual odometry allow to work with thousands of features per frame, while probabilistic techniques handle only few features.

Klein and Murray (2007) present a monocular method called Parallel Tracking and Mapping (PTaM). It uses an approach based on keyframes (see "Appendix I") with two parallel processing threads. The first thread of execution perform the task of robustly tracking a lot of features, while the other one produces a 3D point map aided by BA techniques. PTaM system presents tracking failures in presence of similar textures and moving objects.

In (Konolige and Agrawal 2008; Konolige et al. 2009) the authors use a technique called FrameSLAM and View-Based Maps, respectively. The two methods are based on making a representation of the map as a "skeleton" consisting of a non-linear constraint graph between frames (rather than individual 3D features). The authors use a stereo device mounted on a wheeled robot. Their results show a good performance on long trajectories (approximately 10 km) under changing conditions such as passing through an urban environment.

Recently Strasdat et al. (2010b) have recognized that in order to increase accuracy of the position of a monocular SLAM system it is recommended to increase the number of features (essential property of SfM) rather than the number of frames; as well as, that Bundle Adjustment optimization techniques are better than filters. However, they manifest that the filter might be beneficial in situations of high uncertainty. The ideal SLAM system would exploit the benefits of both SfM techniques and probabilistic filters.

6.3 Bio-inspired models

Milford et al. (2004) use models of the hippocampus (responsible for spatial memory) of rodents to create a location and mapping system called RatSLAM. RatSLAM can generate consistent and stable representations of complex environments using a single camera. The experiments carried out in (Milford and Wyeth 2008; Glover et al. 2010) shows a good performance in real-time tasks in both indoor and outdoor environments. In addition it has the ability to close more than 51 loops of up to 5 km in length and at different hours of day. In (Milford 2008) a larger study of RatSLAM and other biological and navigation systems of bees, ants, primates and humans is presented.

Collett (2010) examines the behavior of ants in desert to analyze how they are guided by visual landmarks and not pheromone trails. Although this research focuses on understanding how ants navigate using visual information, the author states that the proposed solution would be viable and easy to implement in a robot.

7 Representation of the observed world

Mapping is nowadays a very active research area. Free and occupied environment spaces (obstacles) are represented on maps by means of a geometric representation. There are different types of maps reported in the literature, broadly divided in *metric* and *topological* maps.

Metric maps capture the geometric properties of the environment, whereas topological maps describe the connectivity between different locations.

In the metric maps category it can be considered the occupancy grid maps (Gutmann et al. 2008) and landmark-based maps (Klein and Murray 2007; Se et al. 2002; Sáez and Escolano 2006; Mouragnon et al. 2006). Grid maps model free and occupied space by means of a discretization of the environment in form of cells, which may contain 2D, 2.5D o 3D information. Landmark-based maps identify and keep the location 3D of certain salient features in the environment. Thrun (2002) performs a detailed study on the topic of robotic mapping using probabilistic techniques in indoor environments.

With representation through landmarks, only isolated landmarks from the structure of the environment are captured, minimizing thus, the memory resources and computation costs. Due to the foregoing, these types of maps are not ideal for obstacle avoidance or path planning, since the lack of a landmark in a place does not imply that the space is free. However, when the determination of the pose of the entity is more important than the map, these representations are the most suitable.

Topological maps represent the environment as a list of significant places that are connected by arcs (similar to a graph) (Fraundorfer et al. 2007; Eade and Drummond 2008; Konolige et al. 2009; Botterill et al. 2010). A representation of the world based on graphs simplifies the problem of mapping large extensions. However, it is necessary to perform a global optimization of the map to reduce local error (Frese et al. 2005; Olson et al. 2006). A tutorial to formulate the SLAM problem by means of graphs can be consulted in (Grisetti et al. 2010). Other relevant schemes based on graphs are the following: Konolige and Agrawal (2008), Konolige et al. (2009) built a sequence of relative poses between frames, which can recover from critical errors. They show results over 10 km trajectories using stereo vision, although it requires positions generated by an IMU (Inertial Measurement Unit) sensor when an occlusion of the cameras occurs. The authors state that their scheme is applicable to monocular SLAM, even though it is not demonstrated. Another alternative is presented by Mei et al. (2009), which manages to maintain a constant complexity in time to optimize local sub-maps consisting of the closest nodes using a technique called relative bundle adjustment. They generate a trajectory of approximately 2 km, through stereo cameras.

One limitation of the topological representation is the lack of metric information, thus it is impossible to use the map for the purpose of guiding a robot. Consequently, Bazeille and Filliat (2010); Angeli et al. (2009) and Konolige et al. (2011) propose strategies for mixing metric and topological information in a single consistent model.

Currently, the most promising environment representations are based on graphs. But there are still a number of challenges to be overcome, as the ability to edit the graph when detecting wrong estimations of the position, or the generation of global maps of very large dimensions (*lifelong mapping*).

Several datasets containing real image sequences for the evaluation of visual SLAM systems are described in "Appendix III".

The key characteristics of some visual SLAM systems reviewed in this paper are summarized in Table 1. Specifically, we report: (1) the author name and its respective reference, (2) the type of sensing device used, (3) the core of the visual SLAM solution, (4) the kind of environment representation, (5) details of the feature extraction process, (6) the ability and robustness of the system to operate under a variety of conditions: moving objects, abrupt movements and large environments, and also to perform loop closures, and (7) the type of environment used to test the performance of the system.

Table 1 Summary of some systems reviewed

Author	Type of sensing device	Core of the solution	Type of map	Feature extraction	
				Detector	Descriptor
Davison (2003)	Monocular camera	MonoSLAM (EKF)	Metric	Shi and Tomasi operator	Image patches
Nistér et al. (2004)	Stereo or monocular cameras	Visual Odometry	Metric	Harris corners	Image Patches
Sáez and Escolano (2006)	Stereo camera	Global Entropy Minimization Algorithm	Metric	Nitzberg operator	Image patches
Mouragnon et al. (2006)	Monocular camera	Visual odometry +Local bundle adjustment	Metric	Harris corners	Image patches
Klein and Murray (2007)	Monocular camera	Parallel Tracking and Mapping (Visual odometry + Bundle adjustment)	Metric	Fast-10	Image patches
Ho and Newman (2007)	Monocular camera and laser	Delayed state formulation	Metric	Harris affine regions	128D SIFT
Clemente et al. (2007)	Monocular camera	Hierarchical map + EKF	Metric	Shi and Tomasi operator	Image patches
Lemaire et al. (2007)	Stereo or monocular cameras	EKF	Metric	Harris corners	Image patches
Milford (2008)	Monocular camera	RatSLAM (models of the rodent hippocampus)	Topological	Appearance-based matching	
Scaramuzza and Siegwart (2008)	Omnidirectional camera	Visual odometry	Metric	SIFT (difference of gaussians)	Image patches
Eade and Drummond (2008)	Monocular camera	GraphSLAM	Topological	Scale space extrema detector	16D SIFT
Paz et al. (2008)	Stereo camera	Conditionally independent divide and conquer (EKF)	Metric	Shi and Tomasi operator	Image patches

Table 1 continued

Author	Type of sensing device	Core of the solution	Type of map	Feature extraction	
				Detector	Descriptor
Angeli et al. (2008)	Monocular camera	EKF	Topological +Metric	SIFT (difference of gaussians)	128D SIFT + Local hue histograms
Cummins and Newman (2008)	Monocular Camera mounted on a *pan-tilt*	Fast Appearance Based Mapping (FAB-MAP)	Topological	Harris-Affine	U-SURF 128D
Piniés and Tardós (2008)	Monocular camera	Conditionally independent local maps (EKF)	Metric	Harris corners	Image patches
Konolige et al. (2009)	Stereo camera +IMU	Visual odometry + Sparse bundle adjustment	Topological	Fast	Random tree signatures
Williams (2009)	Monocular camera	Hierarchical map + EKF + Visual odometry	Metric	Fast	Image patches+ 16D SIFT
Kaess and Dellaert (2010)	Multi-camera rig	Expectation maximization + Standard bundle adjustment	Metric	Harris corners	Image Patches
Botterill et al. (2010)	Monocular camera	Odometría visual + Bag of words	Topological	Fast	Image patches
Mei et al. (2010)	Stereo camera	Visual odometry +Relative bundle adjustment+ FAB-MAP	Topological + Metric	Fast	128D SIFT

Table 1 continued

Author	Type of sensing device	Core of the solution	Type of map	Cope with				Type of environment
				Moving objects?	Loop closure events?	The kidnapped robot problem?	Large-scale mapping?	
Davison (2003)	Monocular camera	MonoSLAM (EKF)	Metric	No	No	No	No	Indoor
Nistér et al. (2004)	Stereo or monocular cameras	Visual Odometry	Metric	No	No	No	No	Outdoor
Sáez and Escolano (2006)	Stereo camera	Global Entropy Minimization Algorithm	Metric	No	No	No	No	Outdoor/Indoor
Mouragnon et al. (2006)	Monocular camera	Visual odometry +Local bundle adjustment	Metric	Yes	No	No	Yes	Outdoor/indoor
Klein and Murray (2007)	Monocular camera	Parallel Tracking and Mapping (Visual odometry + Bundle adjustment)	Metric	No	No	No	No	Indoor
Ho and Newman (2007)	Monocular camera and laser	Delayed state formulation	Metric	No	Yes	No	Yes	Outdoor/indoor
Clemente et al. (2007)	Monocular camera	Hierarchical map +EKF	Metric	Yes	Yes	No	Yes	Outdoor
Lemaire et al. (2007)	Stereo or monocular cameras	EKF	Metric	No	Yes	No	No	Outdoor
Milford (2008)	Monocular camera	RatSLAM (models of the rodent hippocampus)	Topological	Yes	Yes	Yes	Yes	Outdoor
Scaramuzza and Siegwart (2008)	Omnidirectional camera	Visual odometry	Metric	No	No	No	Yes	Outdoor
Eade and Drummond (2008)	Monocular camera	GraphSLAM	Topological	Yes	Yes	Yes	Yes	Outdoor/indoor

Table 1 continued

Author	Type of sensing device	Core of the solution	Type of map	Cope with				Type of environment
				Moving objects?	Loop closure events?	The kidnapped robot problem?	Large-scale mapping?	
Paz et al. (2008)	Stereo camera	Conditionally independent divide and conquer (EKF)	Metric	No	No	No	Yes	Outdoor/indoor
Angeli et al. (2008)	Monocular camera	EKF	Topological + Metric	No	Yes	No	No	Indoor
Cummins and Newman (2008)	Monocular Camera mounted on a *pan-tilt*	Fast Appearance Based Mapping (FAB-MAP)	Topological	Yes	Yes	No	Yes	Outdoor
Piniés and Tardós (2008)	Monocular camera	Conditionally independent local maps (EKF)	Metric	Yes	Yes	No	Yes	Outdoor
Konolige et al. (2009)	Stereo camera +IMU	Visual odometry + Sparse bundle adjustment	Topological	Yes	Yes	Yes	Yes	Outdoor
Williams (2009)	Monocular camera	Hierarchical map + EKF + Visual odometry	Metric	Yes	Yes	Yes	Yes	Outdoor
Kaess and Dellaert (2010)	Multi-camera rig	Expectation maximization + Standard bundle adjustment	Metric	No	Yes	No	No	Outdoor
Botterill et al. (2010)	Monocular camera	Odometría visual + Bag of words	Topological	Yes	Yes	Yes	Yes	Outdoor/indoor
Mei et al. (2010)	Stereo camera	Visual odometry + Relative bundle adjustment+ FAB-MAP	Topological + Metric	Yes	Yes	Yes	Yes	Outdoor

8 Conclusions

This work verifies that there is a great concern to solve the SLAM problem using vision as the only exteroceptive sensor. This is due mainly to the fact that a camera is an ideal sensor, since it is light, passive, has low-energy consumption, and captures abundant and distinctive information of a scene. However, the use of vision requires reliable algorithms with good performance and consistent under variable light conditions, occlusions or changes in appearance of the environment due to moving people or objects, the apparition of featureless regions, transitions between day and night or any other unforeseen situation. Therefore, SLAM systems using vision as the only sensor are still a challenging and promising research area.

Image matching and the data association are still open research areas in the fields of computer vision and robotic vision respectively. The detector and the descriptor chosen directly affect the performance of the system to track the salient features, recognize areas previously seen, build a consistent model of the environment, and work in real time. Particular to data association is the need for navigation in the long term, in spite of a growing data base and changing and extremely loopy environments. The acceptance of a bad association will cause serious errors in the entire SLAM system, meaning that both the computation of location and map construction will be inconsistent. Therefore, it is important to propose new strategies to reduce the rate of false positives.

Appearance based methods have been very popular for solving data association problem in visual SLAM. The most common technique in this category is the BoVW, due to its speed to find similar images. However, the BoVW is affected by the phenomenon of perceptual aliasing. Likewise, this technique has not been yet thoroughly tested to detect images with large variations of viewpoint or scale, which are transformations that often occur during the loop closure detection, the kidnapped robot problem and multi-session and cooperative mapping. Also, it does not take into account the spatial distribution between the detected features and 3D geometric information, which could be useful when establishing associations.

Although there have been several proposals to build lifelong maps, this issue remains a topic of interest, as well as the ability to build maps in spite of all the problems caused by working in real world environments.

To date, there are no standards for evaluating and comparing the general efficiency and effectiveness of a complete visual SLAM system. Nonetheless, there are several indicators that may characterize their performance, such as the degree of human intervention, accuracy of location, map consistency, real time operation and the control of computational cost that arises with the growth of the map, among others.

Acknowledgments This paper has been made possible thanks to the generous support from the following institutions which we are pleased to acknowledge: CONACYT (Consejo Nacional de Ciencia y Tecnología) and CENIDET (Centro Nacional de Investigación y Desarrollo Tecnológico).

Appendix I: Keyframes

A keyframe is a video frame that is different enough from its predecessor in the sequence, to represent a new location. Keyframes are also used to estimate efficiently the pose of the camera and reduce the redundancy of information. The easiest way to classify a video frame as a keyframe is to compare a video frame with respect to another taken earlier, selecting those that maximize both the distance at which they were captured and the number of feature

matches that exist between them. In (Zhang et al. 2010) a comparative study of different techniques to detect keyframes oriented to the visual SLAM problem is presented.

Appendix II: Bag of visual words (BoVW)

Recently, most contributions to solve data association in visual SLAM use BoVW (Sivic and Zisserman 2003) and its improved version called Vocabulary tree (Nistér and Stewenius 2006). The BoVW has seen a great success in the area of information retrieval (Manning et al. 2008) and content-based image retrieval developed by the computer vision community, due to its speed in finding similar images. However, this technique is not completely precise because it detects several false positives. To solve this problem to some extent, spatial information is normally introduced in the last phase of retrieval, conducting a post-verification taking into account the epipolar constraint (Angeli et al. 2008) or, recently, by means of Conditional Random Fields (Calonder et al. 2010). This verification allows rejecting those recovered images that are not geometrically consistent with the image of reference.

The classic model of BoVW describes images as a set of local features called visual words and the full set of these words is known as visual vocabulary. Many BoVW schemes generate an off-line vocabulary by means of a K-means clustering (but any other can be used) of descriptors from a large corpus of training images (Ho and Newman 2007; Cummins and Newman 2008). An alternative and more effective approach is to dynamically construct the vocabulary from the features that are found as the environment is explored. Such a scheme is described by Angeli et al. (2008) and Botterill et al. (2010).

Some visual words are more useful than others to identify if two images show the same place. The most common scheme to assign each word a specific weight is the TF-IDF. It combines the importance of the words in the image (TF- Term Frequency) and the importance of the words in the collection (IDF- Inverse Document Frequency). In addition, there are other schemes, which are divided into local (Squared TF, Frequency logarithm, Binary, BM25 TF, among others) and global (Probabilistic IDF, Squared IDF, etc.) (Tirilly et al. 2010). An inverted index is used to speed up queries, which organizes the entire set of visual words representing images. An inverted index is structured as a book index. It has one entry for each word of the image collection, followed by a list of all the images in which the word is present.

Appendix III: Datasets to test visual SLAM systems

Some public datasets available to test the visual SLAM systems are: (a) New College and City Centre Datasets (outdoor) (Cummins 2008), used by Cummins and Newman (2008); (b) The New College Vision and Laser Data Set (outdoor) (Smith 2012), captured by Smith et al. (2009) (c) Bovisa (outdoor) and Bicocca (indoor) Datasets of Rawseeds project (Rawseeds 2012), captured by Ceriani et al. (2009); (d) The Cheddar Gorge Data Set (outdoor), captured by Simpson et al. (2012) and RGB-D datasets (indoor) (Sturm 2012) (Sturm et al. 2011).

References

Aguilar W, Frauel Y, Escolano F et al (2009) A robust graph matching for non-rigid registration. Image Vis Comput 27(7):897–910

Andrade J, Sanfeliu A (2002) Concurrent map building and localization with landmark validation. In: Proceedings of the 16th IAPR international conference on pattern recognition, vol 2, pp 693–696
Angeli A, Doncieux S, Filliat D (2008) Real time visual loop closure detection. In: Proceedings of the IEEE international conference on robotics and automation
Angeli A, Doncieux S, Meyer J (2009) Visual topological SLAM and global localization. In: Proceedings of the IEEE international conference on robotics and automation, pp 4300–4305
Artieda J, Sebastian J, Campoy P et al (2009) Visual 3-D SLAM from UAVs. J Intell Robot Syst 55(4):299–321
Asmar D (2006) Vision-inertial SLAM using natural features in outdoor environments. Dissertation, University of Waterloo, Canada
Auat C, Lopez N, Soria C, et al (2010) SLAM algorithm applied to robotics assistance for navigation in unknown environments. J Neuroeng Rehabil. doi:10.1186/1743-0003-7-10
Bailey T, Durrant H (2006) Simultaneous localization and mapping (SLAM): Part II. IEEE Robot Autom Mag 13(3):108–117
Bay H, Tuytelaars T, Van L (2006) SURF: speeded up robust features. In: Proceedings of the European conference on computer vision
Bazeille S, Filliat D (2010) Combining odometry and visual loop-closure detection for consistent topo-metrical mapping. RAIRO Int J Oper Res 44(4):365–377
Beis J, Lowe D (1997) Shape indexing using approximate nearest neighbour search in high-dimensional spaces. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1000–1006
Bogdan R, Sundaresan A, Morisset B et al (2009) Leaving flatland: efficient real-time three-dimensional perception and motion planning. J Field Robot. Special Issue on Three-Dimensional Mapping 26(10):841–862
Bosse M, Newman P, Leonard J, et al (2003) An atlas framework for scalable mapping. In: Proceedings of the IEEE international conference on robotics and automation, pp 1899–1906
Botterill T, Mills S, Green R (2010) Bag-of-words-driven single camera simultaneous localisation and mapping. J Field Robot 28(2):204–226
Bouguet (2010) Camera calibration toolbox for matlab. http://www.vision.caltech.edu/bouguetj/calib_doc/. Accesed 06 March 2012.
Brown MZ, Burschka D, Hager G (2003) Advances in computational stereo. IEEE Trans Pattern Anal Mach Intell 25(8):993–1008
Cadena C, Gálvez-López D, Ramos F, et al (2010) Robust place recognition with stereo cameras. In: Proceedings of the IEEE international conference on intelligent robots and systems, pp 5182–5189
Calonder M, Lepetit V, et al (2010) BRIEF: binary robust independent elementary features. In: Proceedings of the European conference on computer vision
Cannons K (2008) A review of visual tracking. Technical report CSE-2008-07, York University, Department of Computer Science and Engineering
Carrera G, Angeli A, Andrew D (2011) SLAM-based automatic extrinsic calibration of a multi-camera rig. In: Proceedings of the IEEE international conference on robotics and automation
Castellanos J, Tardós JD, Neira J (2001) Multisensor fusion for simultaneous localization and map building. IEEE Trans Robot Autom 17(6):908–914
Ceriani S, Fontana G, Giusti A et al (2009) Rawseeds ground truth collection systems for indoor self-localization and mapping. J Auton Robots 27(4):353–371
Chatila R, Laumond J (1985) Position referencing and consistent world modeling for mobile robots. In: Proceedings of the IEEE international conference on robotics and automation,vol 2, pp 138–145
Chekhlov D, Mayol W, Calway A (2007) Ninja on a plane: automatic discovery of physical planes for augmented reality using visual SLAM. In: Proceedings of the 6th IEEE and ACM international symposium on mixed and augmented reality, pp 1–4
Chekhlov D, Mayol W, Calway A (2008) Appearance based indexing for relocalisation in real-time visual SLAM. In: Proceedings of the British machine vision conference, pp 363–372
Chli M, Davison A (2008) Active matching. In: Proceedings of the European conference on computer vision: part I. doi:10.1007/978-3-540-88682-2_7
Chli M, Davison A (2009) Active matching for visual tracking. Robot Autonom Syst 57(12):1173–1187
Ciganek B, Siebert J (2009) An introduction to 3D computer vision techniques and algorithms. Wiley, New York, pp 194–195
Clemente L, Davison A, Reid I, et al (2007) Mapping large loops with a single hand-held camera. In: Proceedings of robotics: science and systems conference
Collett M (2010) How desert ants use a visual landmark for guidance along a habitual route. In: Psychol Cogni Sci 107(25):11638–11643

Cummins (2008) New college and city centre dataset. http://www.robots.ox.ac.uk/~mobile/IJRR_2008_Dataset. Accesed 06 March 2012
Cummins M, Newman P (2008) FAB-MAP: probabilistic localization and mapping in the space of appearance. Int J Robot Res 27(6):647–665
Cyrill S (2009) Robotic mapping and exploration. Springer Tracts in Advanced Robotics, vol 55, ISBN: 978-3-642-01096-5
Davison A (2003) Real-time simultaneous localisation and mapping with a single camera. In: Proceedings of the IEEE international conference on computer vision, vol2, pp 1403–1410
Davison A, González Y, Kita N (2004) Real-time 3D SLAM with wide-angle vision. In: 5th IFAC/EURON symposium on intelligent autonomous vehicles
Davison A, Reid I, Molton N (2007) MonoSLAM: real-time single camera SLAM. IEEE Trans Pattern Anal Mach Intell 29(6):1052–1067
Dufournaud Y, Schmid C, Horaud R (2004) Image matching with scale adjustment. Comput Vis Image Underst 93(2):175–194
Durrant H, Bailey T (2006) Simultaneous localization and mapping (SLAM): part I the essential algorithms. IEEE Robot Autom Mag 13(2):99–110
Eade E, Drummond T (2006a) Edge landmarks in monocular SLAM. In: Proceedings of the British machine vision conference
Eade E, Drummond T (2006b) Scalable monocular SLAM. In: Proceedings of the IEEE conference on computer vision and pattern recognition, vol 1, pp 469–476
Eade E, Drummond T (2008) Unified loop closing and recovery for real time monocular SLAM. In Proceedings of the British Machine vision conference
Engels C, Stewénius H, Nistér D (2006) Bundle adjustment rules. In: Photogrammetric computer vision
Fawcett T (2006) An introduction to ROC analysis. Pattern Recogn Lett 27(8): 861–874
Fraundorfer F, Engels C, Nister C (2007) Topological mapping, localization and navigation using image collections. In: Proceedings of the IEEE international conference on intelligent robots and systems, pp 3872–3877
Frese U, Larsson P, Duckett T (2005) A multilevel relaxation algorithm for simultaneous localization and mapping. IEEE Trans Robot, pp 196–207, ISSN 1552-3098
Frintrop S, Jensfelt P (2008) Attentional landmarks and active gaze control for visual SLAM. IEEE Trans Robot 24(5):1054–1065
Gee A, Chekhlov D, Calway A, Mayol W (2008) Discovering higher level structure in visual SLAM. IEEE Trans Robot 24(5):980–990
Gemeiner P, Davison A, Vincze M (2008) Improving localization robustness in monocular SLAM using a high-speed camera. In: Proceedings of robotics: science and systems IV
Gil A, Martínez O, Ballesta M, Reinoso O (2009) A comparative evaluation of interest point detectors and local descriptors for visual SLAM. Mach Vis Appl 21(6):905–920
Gil A, Reinoso O, Ballesta M, Juliá M (2010) Multi-robot visual SLAM using a rao-blackwellized particle filter. Robot Autonom Syst 58(1):68–80
Glover A, Maddern W, Milford M, et al (2010) FAB-MAP + RatSLAM: appearance-based slam for multiple times of day. In: Proceedings of the IEEE international conference on robotics and automation
Grasa O, Civera J, Montiel J (2011) EKF monocular SLAM with relocalization for laparoscopic sequences. In: Proceedings of the IEEE international conference on robotics and automation, pp 4816–4821
Grauman K (2010) Efficiently searching for similar images. Commun ACM 53(6):84–94
Grauman K, Darrell T (2007) Pyramid match hashing: sub-linear time indexing over partial correspondences. In: Proceedings of the IEEE conference on computer vision and pattern recognition
Grisetti G, Kümmerle R, Stachniss C, Burgard W (2010) A tutorial on graph-based SLAM. IEEE Trans Intell Transp Syst Mag 2(4):31–43
Gu S, Zheng Y, Tomasi C (2010) Critical nets and beta-stable features for image matching. In: Proceedings of the European conference on computer vision, pp 663–676
Guivant J (2002) Efficient simultaneous localization and mapping in large environments. Dissertation, University of Sydney, Australia
Gutmann J, Fukuchi M, Fujita M (2008) 3D Perception and environment map generation for humanoid robot. Int J Robot Res 27(10):1117–1134
Handa A, Chli M, Strasdat H, Davison A (2010) Scalable active matching. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1546–1533
Harris C, Stephens M (1988) A combined corner and edge detector. In: Proceedings of the fourth alvey vision conference, pp 147–151
Hartley R, Sturm P (1997) Triangulation. Comput Vis Image Underst 68(2): 146–157

Hartley R, Zisserman A (2003) Multiple view geometry in computer vision, 2nd edn. Cambridge, ISBN: 0521540518
Hinterstoisser S, Kutter O, Navab N, et al (2009) Real-time learning of accurate patch rectification. In: Proceedings of the IEEE conference on computer vision and pattern recognition
Ho K, Newman P (2007) Detecting loop closure with scene sequences. Int J Comput Vis 74(3):261–286
Huang A, Bachrach A, Henry P, et al (2011) Visual odometry and mapping for autonomous flight using rgb-d camera. International symposium on robotics research
Johnson M, Pizarro O, Williams S, Mahon I (2010) Generation and visualization of large-scale three-dimensional reconstructions from underwater robotic surveys. J Field Robot 27(1):21–51
Jones E, Soatto S (2011) Visual-inertial navigation, mapping and localization: a scalable real-time causal approach. Int J Robot Res 30(4):407–430
Kaess M, Dellaert F (2010) Probabilistic structure matching for visual SLAM with a multi-camera rig. Comput Vis Image Underst 114:286–296
Kawewong A, Tangruamsub S, Hasegawa O (2010) Position-invariant robust features for long-term recognition of dynamic outdoor scenes. IEICE Trans Inform Syst 9:2587–2601
Ke Y, Sukthankar R (2004) PCA-SIFT: a more distinctive representation for local image descriptors. In: Proceedings of the IEEE conference on computer vision and pattern recognition, vol 2, pp 506–513
Klein G, Murray D (2007) Parallel tracking and mapping for small AR workspaces. In: Proceedings of the 6th IEEE and ACM international symposium on mixed and augmented reality
Klein G, Murray D (2008) Improving the agility of keyframe-based SLAM. In: Proceedings of the European conference on computer vision, pp 802–815
Koch O, Walter M, Huang A, Teller S (2010) Ground robot navigation using uncalibrated cameras. In: Proceedings of the IEEE international conference on robotics and automation, pp 2423–2430
Konolige K, Agrawal M (2008) FrameSLAM: from bundle adjustment to real-time visual mapping. IEEE Trans Robot 24(5):1066–1077
Konolige K, Bowman J, Chen J (2009) View-based maps, In: Proceedings of robotics: science and systems
Konolige K, Marder-Eppstein E, Marthi B (2011) Navigation in Hybrid metric- topological maps. In: Proceedings of the IEEE international conference on robotics and automation
Kragic D, Vincze M (2009) Vision for robotics. Found Trends Robot 1(1):1–78, ISBN: 978-1-60198-260-5
Kulis B, Jain P, Grauman K (2009) Fast similarity search for learned metrics. IEEE Trans Pattern Anal Mach Intell 31(12):2143–2157
Lemaire T, Berger C, Jung I et al (2007) Vision-based SLAM: stereo and monocular approaches. Int J Comput Vis 74(3):343–364
Lepetit V, Fua P (2005) Monocular model-based 3D tracking of rigid objects. Found Trends Comput Graph Comput Vis 1(1):1–89
Lepetit V, Fua P (2006) Keypoint recognition using randomized trees. IEEE Trans Pattern Anal Mach Intell 28(9): 1465–1479
Li H, Kimi E, Huang X, He L (2010) Object matching with a locally affine-invariant constraint In: Proceedings of the International conference on pattern recognition, pp 1641–1648
Lin K, Wang C (2010) Stereo-based simultaneous localization, mapping and moving object tracking. In: Proceedings of the IEEE international conference on intelligent robots and systems, pp 3975–3980
Lowe D (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2):91–110
Magnusson M, Andreasson H, et al (2009) Automatic appearance-based loop detection from 3D laser data using the normal distribution transform. J Field Robot. Three Dimensional Mapping Part 2, 26(12): 892–914
Manning C, Schütze H, Raghavan P (2008) Introduction to information retrieval, Cambridge University Press, Cambridge, ISBN: 0521865719
Majumder S, Scheding S, Durrant H (2005) Sensor fusion and map building for underwater navigation. In: Proceedings of Australian conference on robotics and automation
Martinez J, Calway (2010) A unifying planar and point mapping in monocular SLAM. In: Proceedings of the British machine vision conference, pp 1–11
Matas J, Chum O, et al (2002) Robust wide baseline stereo from maximally stable extremal regions. In: Proceedings of the British machine vision conference vol 22, no. 10, pp 761–767
Mei C, Reid I (2008) Modeling and generating complex motion blur for real-time tracking. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1–8
Mei C, Sibley G, Cummins M, et al (2009) A constant-time efficient stereo SLAM system. In: Proceedings of the British machine vision conference
Mei C, Sibley G, Cummins M et al (2010) RSLAM: a system for large-scale mapping in constant-time using stereo. Int J Comput Vision 94(2):1–17

Mei C, Sommerlade E, Sibley C, et al (2011) Hidden view synthesis using real-time visual SLAM for simplifying video surveillance analysis. In: Proceedings of the IEEE International conference on robotics and automation, vol 8, pp 4240–4245
Migliore D, Rigamonti R, Marzorati D, et al (2009) Use a single camera for simultaneous localization and mapping with mobile object tracking in dynamic environments. In: ICRA workshop on safe navigation in open and dynamic environments: application to autonomous vehicles
Mikolajcczyk K, Schmid C (2002) An affine invariant interest point detector. In: Proceedings of the European conference on computer vision, pp 128–142
Mikolajcczyk K, Schmid C (2005) A performance evaluation of local descriptors. IEEE Trans Pattern Anal Mach Intell 27(10):1615–1630
Mikolajczyk K, Tuytelaars T, Schmid S et al (2005) A comparision of affine region detectors. Int J Comput Vis 65:43–72
Milford M (2008) Robot navigation from nature: simultaneous, localisation, mapping, and path planning based on hippocampal models, vol. 41. Springer Tracts in Advanced Robotics, ISBN: 3540775196
Milford M, Wyeth G (2008) Mapping a suburb with a single camera using a biologically inspired SLAM system. IEEE Trans Robot 24(5):1038–1053
Milford M, Wyeth G, Prasser D (2004) RatSLAM: a hippocampal model for simultaneous localization and mapping. In: Proceeding of the IEEE international conference on robotics and automation, vol 1, pp 403–408
Molton N, Davison A, Reid I (2004) Locally planar patch features for real-time structure from motion. In: Proceedings of the British machine vision conference
Montemerlo M (2003) FastSLAM: a factored solution to the simultaneous localization and mapping problem with unknown data association, Dissertation, Carnegie Mellon University, USA
Montemerlo M, Thrun S, Koller D, et al (2002) FastSLAM: a factored solution to the simultaneous localization and mapping problem. In: Proceedings of the AAAI national conference on artificial intelligence, pp 593–598
Montiel J, Civera J, Davison A (2006) Unified inverse depth parametrization for monocular SLAM. In: Proceedings of robotics: science and systems
Morel J, Yu G (2009) ASIFT: a new framework for fully affine invariant image comparison. SIAM J Imaging Sci 2(2): 438–469
Moreels P, Perona P (2005) Evaluation of features detectors and descriptors based on 3D objects. In: Proceedings of the IEEE international conference on computer vision, pp 800–807
Mouragnon E, Dhome M, Dekeyser F, et al (2006) Monocular vision based SLAM for mobile robots. In: Proceedings of the international conference on pattern recognition, pp 1027–1031
Mouragnon E, Lhuillier M, Dhome M, et al (2009) Generic and real time structure from motion using local bundle asjustment. Image Vis Comput, pp 1178–1193, ISSN: 0262-8856
Neira J, Tardós JD (2001) Data association in stochastic mapping using the joint compatibility test. In: Proceedings of the IEEE international conference on robotics and automation 17(6): 890–897
Newman P, Leonard J, Neira J, Tardós J (2002) Explore and return: experimental validation of real time concurrent mapping and localization. In: Proceedings of the IEEE international conference on robotics and automation, vol 2, pp 1802–1809
Nistér D (2004) An efficient solution to the five-point relative pose problem. IEEE Trans Pattern Anal Mach Intell 26(6):756–770
Nistér D, Stewenius H (2006) Scalable recognition with a vocabulary tree. In: Proceedings of the IEEE conference on computer vision and pattern recognition, vol 2, pp 2161–2168
Nistér D, Naroditsky O, Bergen J (2004) Visual odometry. In: Proceedings of the IEEE conference on computer vision and pattern recognition vol 1, pp 652–659
Nüchter A, Lingemann K, Hertzberg J et al (2007) 6D SLAM—3D mapping outdoor environments. J Field Robot 24(8):699–722
Nützi G, Weiss S, Scaramuzza D, Siegwart R (2010) Fusion of IMU and vision for absolute scale estimation in monocular SLAM. J Intell Robot Syst. doi:10.1007/s10846-010-9490-z
Olson C, Matthies L, Schoppers M, Maimone M (2003) Rover navigation using stereo ego-motion. Robot Autonom Syst 43(4):215–229
Olson E, Leonard J, Teller S (2006) Fast iterative optimization of pose graphs with poor initial estimates. In: Proceedings of the IEEE international conference on robotics and automation, pp 2262-2269
Olson C, Matthies L, Wright J et al (2007) Visual terrain mapping for mars exploration. Comput Vis Image Underst 105(1):73–85
OpenCV (2009) OpenCV: Camera calibration and 3D reconstruction. http://opencv.willowgarage.com/documentation/camera_calibration_and_3d_reconstruction.html Accesed 06 March 2012

Özuysal M, Calonder M, Lepetit V, Fua P (2010) Fast keypoint recognition using random ferns. IEEE Trans Pattern Anal Mach Intell 32(3):448–461
Paz L, Piniés P, Tardós JD, Neira J (2008) Large-scale 6DOF SLAM with stereo-in-hand. IEEE Trans Robot 24(5):946–957
Piniés P, Tardós JD, Neira J (2006) Localization of avalanche victims using robocentric SLAM. In: Proceedings of the IEEE international conference on intelligent robots and systems. pp 3074–3079
Piniés P, Tardós JD (2008) Large scale SLAM building conditionally independent local maps: application to monocular vision. IEEE Trans Robot 24(5):1094–1106
Pollefeys M, Van L, Vergauwen M et al (2004) Visual modeling with a hand-held camera. Int J Comput Vis 59(3):207–232
Pretto A, Menegatti E, Pagello E (2007) Reliable features matching for humanoid robots. In: IEEE-RAS international conference on humanoid robots, pp 532–538
Pupilli M, Calway A (2006) Real-time visual SLAM with resilience to erratic motion. In: Proceedings of the IEEE conference on computer vision and pattern recognition, vol 1, pp 1244–1249
Raguram R, Frahm J, Pollefeys M (2008) A comparative analysis of RANSAC techniques leading to adaptive real-time random sample consensus. In: Proceedings of the European conference on computer vision, pp 500–513
Rawseeds (2012) Bovisa and bicocca datasets. http://www.rawseeds.org/rs/datasets. Accesed 06 March 2012
Ribas D, Ridao P, Tardós JD et al (2008) Underwater SLAM in man-made structured environments. J Field Robot 25(11):898–921
Rosten E, Drummond T (2006) Machine learning for high-speed corner detection. In: Proceedings of the European conference on computer vision, pp 430–443
Rublee E, Rabaud V, Konolige K, Bradski G (2011) ORB: an efficient alternative to SIFT or SURF. In: Proceedings of the IEEE international conference on computer vision
Scaramuzza (2011) OcamCalib toolbox: omnidirectional camera and calibration toolbox for matlab. https://sites.google.com/site/scarabotix/ocamcalib-toolbox. Accesed 06 March 2012
Scaramuzza D, Siegwart R (2008) Appearance guided monocular omnidirectional visual odometry for outdoor ground vehicles. IEEE Trans Robot 24(5): 1015–1026
Se S, Lowe D, Little J (2002) Mobile robot localization and mapping with uncertainty using scale-invariant visual landmarks. Int J Robot Res 21(8):735–758
Se S, Lowe D, Little J (2005) Vision- based global localization and mapping for mobile robots. IEEE Trans Robot 21(3):364–375
Sáez J, Escolano F (2006) 6DOF entropy minimization SLAM. In: Proceedings of the IEEE international conference on robotics and automation, pp 1548–1555
Sanromá G, Alquézar R, Serratosa F (2010) Graph matching using SIFT descriptors—an application to pose recovery of a mobile robot. In: 13th joint IAPR international workshop on structural, syntactic and statistical pattern recognition, pp 254–263
Silpa C, Hartley R (2008) Optimised KD-trees for fast image descriptor matching. In: Proceedings of the IEEE conference on computer vision and pattern recognition
Sinha S, Frahm J, Pollefeys M, Genc Y (2006) GPU-based video feature tracking and matching. In: Workshop on edge computing using new commodity architectures
Sivic J, Zisserman A (2003) Video google: a text retrieval approach to object matching in videos. In: Proceedings of the IEEE international conference on computer vision
Simpson R, Cullip J, Revell J (2012) The cheddar gorge data set. http://www.openslam.org/misc/BAE_RSJCJR_2011.pdf. Accesed 06 March 2012
Smith (2012) The new college vision and laser data set. http://www.robots.ox.ac.uk/NewCollegeData/. Accesed 06 March 2012
Smith R, Self M, Cheeseman P (1990) Estimating uncertain spatial relationships in robotics. In: Autonomous robot vehicles. Springer, New York, pp 167–193, ISBN:0-387-97240-4
Smith M, Baldwin I, Churchill W et al (2009) The new college vision and laser data set. Int J Robot Res 28(5):595–599
Solà J (2007) Multi-camera VSLAM: from former information losses to self-calibration. In: Proceedings of the IEEE international conference on intelligent robots and systems, workshop on visual SLAM
Steder B, Grisetti G, Stachniss C et al (2008) Visual SLAM for flying vehicles. IEEE Trans Robot 24(5):1088–1093
Sturm (2012) RGB-D dataset and benchmark. http://cvpr.in.tum.de/data/datasets/rgbd-dataset. Accesed 06 March 2012
Sturm J, Magnenat S, et al (2011) Towards a benchmark for RGB-D SLAM evaluation. In: Proceedings of the RGB-D workshop on advanced reasoning with depth cameras at robotics: science and systems conference

Strasdat H, Montiel J, Davison A (2010a) Scale drift-aware large scale monocular SLAM. In Proceeding of robotics: science and systems
Strasdat H, Montiel J, Davison A (2010b) Real-time monocular SLAM: why filter?. In: Proceedings of the IEEE international conference on robotics and automation
Svoboda (2011) Multi-camera self-calibration. http://cmp.felk.cvut.cz/~svoboda/SelfCal/index.html Accesed 06 March 2012
Tardós JD, Neira J, Newman P et al (2002) Robust mapping and localization in indoor environments using sonar data. Int J Robot Res 21:311–330
Taylor S, Drummond T (2009) Multiple target localization at over 100 FPS. In: Proceedings of the British machine vision conference
Thrun S (2002) Robotic mapping: a survey. Exploring artificial intelligence in the new millennium, ISBN:1-55860-811-7
Thrun S (2003) A system for volumetric robotic mapping of abandoned mines. In: Proceedings of the IEEE international conference on robotics and automation, vol 3, pp 4270–4275
Thrun S, Leonard J (2008) Simultaneous localization and mapping. Springer Handbook of Robotics; Siciliano, Khatib Editors, ISBN: 978-3-540-23957-4, pp 871–886
Thrun S, Koller D, Ghahramani Z, et al (2002) Simultaneous mapping and localization with sparse extended information filters: theory and initial results. Technical Report CMU-CS-02-112, Carnegie Mellon
Thrun S, Montemerlo M, Dahlkamp H et al (2005a) Stanley: the robot that won the DARPA grand challenge. J Field Robot 23(9):661–692
Thrun S, Burgard W, Fox D, (2005b) Probabilistic Robotics. The MIT Press, New York, ISBN: 0262201623
Thrun S, Montemerlo M, Aron A (2006) Probabilistic terrain analysis for high speed desert driving. In: Proceedings of robotics: science and systems
Tirilly P, Claveau V, Gros P (2010) Distances and weighting schemes for bag of visual words image retrieval. In: Proceedings of the international conference on multimedia information retrieval, pp 323–333
Triggs B, Mclauchlan P, Hartley R, Fitzgibbon A (1999) Bundle adjustment—a modern synthesis. In: Proceedings of the international workshop on vision algorithms: theory and practice, pp 298–375
Tuytelaars T, Mikolajczyk K (2008) Local invariant feature detectors: a survey. Found Trends Comput Graph Vis
Tuytelaars T, Van-Gool L (2004) Matching widely separated views based on affine invariant regions. Int J Comput Vis 59(1):61–85
Vidal T, Bryson M, Sukkarieh S, et al (2007) On the observability of bearing-only SLAM. In: Proceedings of the IEEE international conference on robotics and automation, pp 4114–4119
Vidal T, Berger C, Sola J, Lacroix S (2011) Large scale multiple robot visual mapping with heterogeneous landmarks in semi-structured terrain. Robot Autonom Syst, pp 654–674
Wang C, Thorpe Ch, Thrun S et al (2007) Simultaneous localization, mapping and moving object tracking. Int J Robot Res 26(9):889–916
Wangsiripitak S, Murray D (2009) Avoiding moving outliers in visual SLAM by tracking moving objects. In: Proceedings of the IEEE international conference on robotics and automation, pp 375–380
Williams B (2009) Simultaneous localisation and mapping using a single camera. PhD, thesis, Oxford University, England
Williams B, Klein G, Reid I (2007) Real-time SLAM relocalisation. In: Proceedings of the IEEE international conference on computer vision
Williams B, Cummins M, Neira J, Newman P, Reid I, Tardós JD (2009) A comparision of loop closing techniques in monocular SLAM. Robot Autonom Syst 57(12):1188–1197
Willson (1995) Tsai camera calibration software. http://www.cs.cmu.edu/~rgw/TsaiCode.html. Accesed 06 March 2012
Yilmaz A, Javed O, Shah M (2006) Object tracking: a survey. ACM Comput Surv 38(4): 1–45
Zhang Z (2000) A flexible new technique for camera calibration. IEEE Trans Pattern Anal Mach Intell 22(11):1330–1334
Zhang W, Kosecka J (2006) Image based localization in urban environments. In: Proceedings of the third international symposium on 3d data processing, visualization, and transmission
Zhang Z, Deriche R, Faugeras O, Luong Q (1994) A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. J Artif Intell. Special volume on Computer Vision 78(1): 87–119
Zhang H, Li B, Yang D (2010) Keyframe detection for appearance-based visual SLAM. In: Proceedings of the IEEE international conference on intelligent robots and systems, pp 2071–2076