FUSING LASER AND VISION DATA FOR A PERCEPTUALLY RICH ENVIRONMENT DESCRIPTION

Key words
Mobile robot, navigation, computer vision, feature extraction.

Summary
In this paper we discuss methods to increase the discriminative properties of the laser-based geometric landmarks used in simultaneous localisation and mapping by employing monocular vision data. Vertical edges extracted from images help to estimate the length of the line segments, which are only partially observed. Salient visual features, which defy simple geometric interpretation, are handled by the scale invariant feature transform method. These different types of photometric features are aggregated together with the basic 2D line segments extracted from the laser scanner data into the Perceptually Rich Segments.

Introduction
An important competence for an autonomous mobile robot is the ability to build an environment map from sensory data, while concurrently computing its own pose estimate. Most of the successful solutions to this Simultaneous Localization and Mapping (SLAM) problem employ 2D laser scanners that provide precise and reliable range measurements [3]. In spite of many improvements, data association is still a challenge in laser-based SLAM. To relocate itself, the robot should unambiguously recognise its surroundings, which in many environments is impossible relying solely on geometry and the 2D laser data, due to environment symmetries and a lack of distinguishable
geometric landmarks at the height of the laser beam plane. Recently, vision received much attention in the SLAM research, because of its ability to capture a rich description of the environment, including both the geometric and photometric information. Cameras enable the robot to recognise a large variety of visually salient photometric features, even in environments where distinctive geometric landmarks are unavailable. However, passive visual sensing, particularly with a single camera (monocular), has many limitations related to occlusions, shadows, specular reflections, and proper landmark selection [2, 4]. Some authors also exploit the idea of combining laser scanner and vision data in SLAM. Within the context of Extended Kalman Filter based SLAM (EKF-SLAM), a multi-sensor system consisting of a 2D laser scanner and a camera has been employed in [1]. The authors of [7] use a combination of the feature-based and appearance-based approaches to SLAM with both the laser and monocular vision data to solve the first-location problem in SLAM. A recent work by Newman and Ho [6] shows how monocular vision data can be used for loop closing in SLAM.

The laser scanners and vision sensors have complementary properties. Hence, the main idea behind this work is that, for autonomous navigation in unknown indoor environments, we should employ both these sensors, making proper use of the characteristics of both sensing modalities, i.e. we should use the laser data to estimate the location of the landmarks and use the vision data to make these landmarks more distinguishable in the data association process.

The baseline of the approach is our previous research on robust landmark extraction in laser-only SLAM [10]. Here, we add to the line segments reliable geometric constraints on their length. These constraints and the assumption that segments are parts of planar vertical surfaces allow us to convert the segments into semiplanes and embed into them descriptors of salient photometric features. Scale Invariant Feature Transform (SIFT) descriptors have been chosen to efficiently store the information on photometric features. A segment which is delimited by credible edges confirmed by image data, and is augmented with SIFT descriptors of the salient features detected within the area of a semiplane defined by this segment becomes a Perceptually Rich Segment – a new landmark type we introduce.

1. Improving segments with vision data

1.1. Line segments from laser scans

Reliable extraction of line segments from the noisy clusters of laser scanner points is accomplished by using the methods described in [10]. Segments are extracted from the ordered set of laser scanner points (single scan). The spatial uncertainty of a point is represented by its covariance matrix computed upon the uncertainty model of the LMS 200 sensor range measurements. A split-and-merge procedure is applied to find line segments from groups of points. The
supporting line of an extracted segment is described by the SPmodel [1] landmark \( L_{RF} \), with regard to the robot frame. The endpoints of the segment are determined by projecting the laser points onto the infinite line and trimming the line at the extreme points. Finally, the centre point and the length \( l_F \) are computed. Uncertainty of the segment length is not determined, due to the low credibility of endpoints extracted solely from the laser scan.

The line segments obtained from laser data are further structured into poly-lines. This allows upgrading of the semantic representation of the environment by representing distinctive objects [9] and facilitates a straightforward detection of corners (either convex or concave) which are points of intersections of two consecutive segments. The covariance matrix of a corner is computed by propagating uncertainty from parameters of the two crossing segments [1]. The detected corners together with the segment endpoints are then considered as candidates for vertical edges constraining the segments, if they are confirmed by the photometric information.

1.2. Vertical edges from monocular images

Vertical lines are ubiquitous visual features in many indoor environments. The first step of the detection procedure is vertical edge enhancement by using a custom Sobel filter approximating the image gradient in the horizontal direction. Then, hysteresis thresholding is applied to obtain a binary image. The resulting edges are thinned, and the horizontal position of each edge pixel having the value of '1' (the background is set to '0') is corrected upon the camera calibration results [11].

Because we look only for vertical lines in the binary images, the line fitting is reduced to a one-dimensional problem. Initially, we have applied the Hough transform to solve it. However, vertical edges detected with a gradient filter are usually represented in the binary image by pixels located in several neighbouring columns of this image. Because of that, the one-dimensional Hough transform does not work well, whenever there are many vertical lines located close each other. Applying non-maxima suppression in the parameter space eliminates false positives; however, in many cases, this prevents detection of weaker edges neighbouring the stronger ones. To overcome this problem, we have developed a vertical edge extractor based on the detection of local maxima of a one-dimensional, discrete function describing the location of vertical edges in an image. For each \( n = 1, 2, 3, \ldots, \omega \), where \( \omega \) is the image width in pixels the number of pixels \( I_e \) belonging to edges in the \( n \)-th column of the binary image \( I_b \) is computed:

\[
I_e(n) = \sum_{m=1}^{h} I_b(n, m)
\]  

where: \( h \) – the image height in pixels.
In this way, a discrete function describing the vertical edge locations is established. The edges are detected by analysing an approximation of the first derivative of \( I_e \), computed by convolving this function with the mask \([1 1 1 0 -1 -1 -1]\), which has been chosen experimentally.

The values of \( n \) at which the derivative changes its sign are the local maxima of \( I_e \), thus indicating a detected vertical edge. If a zero of \( I_e \) occurs between two discrete values of \( n \) (i.e. between two image columns), it is approximated, it allows us to compute vertical edge locations with sub-pixel resolution. This method outperforms the Hough transform, particularly when multiple vertical edges appear in compact groups.

Each vertical edge extracted from an image is represented initially by an angle related to the camera coordinate system:

\[
\phi_v = \arctg \left( \frac{x_f - x_c}{f} \right),
\]

(2)

where: \( x_f \) – the \( x \)-coordinate of the extracted edge,
\( x_c \) – the \( x \)-coordinate of the camera’s centre of distortion,
\( f \) – the focal length (all values in pixels).

Uncertainty of the vision angle \( \sigma_\phi \) is computed by propagating the uncertainty of edge detection and camera calibration parameters obtained with a standard calibration technique [11]. Although the position of the edge is computed with sub-pixel-resolution, we conservatively set the standard deviation of this measurement to the size of a pixel on the CCD matrix.

1.3. Integrating vertical edges with segments

A vertical edge extracted from the camera image is represented by its vision angle \( \phi_v \) with regard to (w.r.t.), the camera coordinate system. In order to establish correspondence between the features obtained from both sensors, the vision angles are transformed to the scanner coordinates. The \( i \)-th vision angle \( \phi_v^i \) is converted to the \( \phi_v^L \) angle relative to the laser-scanner-coordinate system by using the coordinate transformation between the camera and the scanner. This transformation was obtained by using a calibration technique proposed in [1]. Because both sensors involved are parallel to the floor plane, only the 2D translation \( t_{LC} = [x_{LC} \ y_{LC}]^T \) and rotation \( \phi_{LC} \) have been computed. Therefore, photometric features can be represented in the reference frame of the laser scanner.

Geometric information on possible vertical edges is extracted from the segment-based local map, as described in Section 1.1. These edges are represented as angles in the scanner-coordinate system. The observation angle of \( j \)-th edge extracted from the local segment-based map is given as:
\[ \phi_j = \arctg \left( \frac{x_j}{y_j} \right) \]  

(3)

Where: \([x_j \ y_j]^T\) – the edge coordinates w.r.t., the scanner reference frame.

Squared Mahalanobis distance test is performed on the observation angles of all pairs of features from both sensors to find the geometric feature corresponding to the photometric observation. This test takes into account the variances \(\sigma_\phi^2\) and \(\sigma_\sigma^2\) computed by propagating the uncertainty from the particular sensor model. If multiple geometric features satisfy the test, the one that is closest to the sensor is chosen to avoid false pairings resulting from occlusions (Fig. 1A – false match marked with arrow). A geometric vertical edge, whose observation angle \(\phi_j\) corresponds to the vision angle \(\phi^L\) representing a photometric edge is considered a confirmed vertical edge. Such a feature can be used to terminate the 2D line segments that are associated to this vertical edge in the local map (Fig. 1B – constrained segments are marked with arrows).

Fig. 1. Vertical edges extracted from scans and confirmed by image data (A,B), and photometric edges (C)

To achieve an efficient representation of segments with edges, we have adopted an extension to the SPmap framework proposed in [8]. This extension retains the basic SPmap landmark representation: a segment is defined by the uncertain location of its supporting line w.r.t., the global reference frame \(W\):

\[ L_{WF} = (\hat{x}_F \ \hat{y}_F \ C_F \ B_F) \]  

(4)

where: 
\[ \hat{x}_F = \begin{bmatrix} \hat{x}_F \ \hat{y}_F \ \hat{\phi}_F \end{bmatrix}^T \] – the estimate of the location,
\[ \hat{\phi}_F = \begin{bmatrix} d_y \ d_\phi \end{bmatrix}^T \] – the error vector,
\[ C_F \] – covariance matrix.
The row selection matrix $B_F$ selects the coordinates relevant for the uncertainty of a segment – lateral displacement and orientation. To represent a segment rather than infinite line, the landmark has the reference frame attached in its centre point and is augmented with its length $l_F$. In [8], two flags $\alpha_l$ and $\alpha_r$ are added to this representation in order to indicate if the edges constraining the segment (left and right, respectively) have been confirmed. The error vector $\hat{p}_F$ is also augmented by a variable size vector for the corrections of edges. Uncertainty reduction in the edge position is accomplished by a standard EKF algorithm operating on the SPmap landmarks and the vision angles of respective photometric features [1]. Segments constrained by confirmed vertical edges at both ends are promoted to semiplanes.

Some photometric edges do not have corresponding features in the laser-based map. Although such edges do not constrain the segments, they are useful for robot localisation and should be represented in the map. Unfortunately, information provided by monocular vision is insufficient to determine range to the detected vertical edges from a single image. Multiple views are required to estimate an edge location w.r.t., the camera coordinates. To overcome this problem, we determine from an image only the bearing information $\phi_v^i$, while the range to the vertical edge is then estimated using the laser data and the coordinate transformation between the laser scanner and the camera. We are interested only in photometric edges that are reliable enough to be used in SLAM: They have to be longer than 160 pixels and located on a planar vertical surface defined by some laser-based segment. His helps to eliminate spurious features resulting from shadows or light spots on the floor. The location of a vertical edge is estimated as the intersection point of a virtual line defined by the centre of the coordinate system and the given vision angle $\phi_v^i$ and the supporting line of the first segment crossed by this virtual line (Fig. 1C). The location uncertainty of a photometric edge is computed by propagating the covariance of the vision angle $\sigma_v^2$ and the uncertainty $C_F$ of the segment involved in the intersection.

2. Perceptually rich segments

Once the laser-based 2D segments have been converted into semiplanes representing planar vertical surfaces, they can be used as "frames" for rich photometric information, which, however, defies geometric interpretation. Such photometric information can give the segments much more distinctiveness and enable robust place recognition in SLAM. In order to efficiently embed the photometric information into the segments, we need a method for extracting distinctive features from images that can be used to perform reliable matching between different views of a scene. We also wish to detect image features that
are robust to variations in such image acquisition parameters as viewpoint, translation, scale, and illumination changes. A feature extraction method satisfying these conditions is the Scale Invariant Feature Transform [5].

The SIFT algorithm detects keypoints of interest, which are local extremes of the Difference-of-Gaussian images through location and scale space. A SIFT descriptor is created by computing the gradient magnitude and orientation at each key point in a region around this point location. Dominant orientations are determined by accumulating the sampled information into orientation histograms over 4 × 4 local image regions. Then, SIFT descriptor vectors are computed for each processed region. The descriptor is a 4 × 4 array of histograms, each with 8 orientation bins, which results in a 128-dimensional feature vector [5].

Figure 2 shows examples of SIFT descriptors detected on images taken in a typical office-like environment. The keypoints are represented by arrows, with the orientation of each arrow indicating the orientation of a descriptor and the length proportional to scale. Note, that SIFT keypoints are detected mainly at visually salient regions of images, such as the doorplates (Fig. 2A), fire extinguishers, and signboards (Fig. 2B and C).

![Fig. 2. Examples of the SIFT features detected in different locations](image)

The SIFT keypoints found in an image are projected onto the vertical semiplanes extracted from the corresponding laser scan by using projective geometry and the camera-scanner calibration data. This is done under an assumption that the photometric features represented by SIFT vectors are located on approximately flat vertical surfaces. Because the sensors provide no reliable information about the height of the vertical semiplanes, we set this height to a predefined value. A data structure consisting of a 2D segment limited in length by photometry-confirmed vertical edges and a set of SIFT descriptors located inside the rectangular area created by this segment and the edges is considered a new landmark type: Perceptually Rich Segment (PRS).

Although the PRS landmarks are much more distinctive than ordinary 2D segments, an efficient SIFT matching procedure is required to employ the photometric information for landmark matching in SLAM. In [5], images are matched by individually comparing each SIFT feature from the considered
image to a database of keypoints (from other images) and finding candidate matching features based on nearest neighbour criteria. The nearest neighbour is defined as the key point from the database with minimum Euclidean distance for the descriptor vector to the considered feature. In the case under study, there is no database. SIFT keypoints from two PRS landmarks are compared to judge if they represent a matching pair of local scenes (Fig. 3). For each vector representing a SIFT feature in the first PRS distances to all the keypoints in the second PRS are computed. Then, the distance of the closest neighbour is compared to that of the second-closest neighbour. If the difference between these distances is smaller than a given threshold, the vectors are considered as matching.

If we use only the Euclidean distance for descriptor matching, false matches occur frequently, mostly due to large changes in scale and background clutter.

In order to increase matching reliability, we use the similarity of SIFT vector orientation as an additional matching criteria. If the difference in orientation of two compared vectors is bigger than 20°, then the matching result is considered negative. This additional criteria is very useful in the robotic application under study, because the images acquired by the mobile robot with the camera attached to its body and parallel to the floor plane are free from rotation variations; but, on the other hand, they may differ significantly in scale.

3. Experimental results

In order to verify the presented approach to environment description, we have performed a number of experiments with the Labmate robot equipped with a Sick LMS 200 laser scanner and a CCD camera providing 640 × 480 B/W images.

Fig. 4 shows the results of local map building by fusing the geometric and photometric information extracted from a laser scan and a camera image taken at the same robot pose. In Fig. 4A, strong vertical edges found on the image are shown. These edges are depicted in Fig. 4B as vision angles w.r.t., the reference frame of the laser scanner overlaid on the segments extracted from the scan. The dashed lines indicate the spatial uncertainty of these angles. SIFT keypoints extracted from the image are visualised in Fig. 4C. The resulting local map consisting of “plain” segment landmarks (light grey) and PRS landmarks (dark
grey) is shown in Fig. 4D. Note that only the segments having their edges confirmed by the photometric data have been upgraded to PRS landmarks and that the SIFT keypoints embedded in the existing PRS'es clearly represent visually salient features in this scene, particularly the poster on the doors.

Fig. 4. Example of a local map built with the proposed method

Fig. 5 shows the results of a more extended experiment, in which a map of a part of the corridor has been created. This map contains segment landmarks (light grey) and PRS landmarks (dark grey). The inset images are the source camera views of the photometric features: vertical edges A, B, and SIFTs C shown in the respective parts of the map. The map was built by using the EKF-SLAM method implemented as in [10]. In this experiment, the path of the robot did not make any loop, and the amount of overlapping in the camera images acquired along this path was very small. Because of that, SIFT matching was not used, but whenever possible PRS length was used as an additional constraint for landmark matching.

Fig. 5. Fragment of a corridor mapped by a SLAM system using the PRS landmarks

Conclusions

This paper describes a work in progress. Some initial results concerning the use of photometric features to increase the discriminative properties of laser-
based line segments commonly used in indoor EKF-SLAM have been presented. The ongoing work is devoted to the full integration of the PRS landmarks within the EKF-SLAM framework, including SIFT-based landmark matching and loop detection, and to the implementation of more effective saliency detection schemes which would help to reduce the number of detected meaningless photometric features that increase the probability of false matchings in SLAM.

References


Reviewer: Barbara PUTZ
Opis otoczenia na podstawie danych z sensorów laserowych i wizyjnych

Słowa kluczowe

Robot mobilny, nawigacja, system wizyjny, ekstrakcja cech.

Streszczenie
