2nd International Conference on Autonomous Robots and Agents
December 13-15, 2004 Palmerston North, New Zealand

Using 3D Visual Landmarks to Solve the Correspondence Problem in Simultaneous Localisation and Mapping

[1]Michael J. Cree, [2]Margaret E. Jefferies, and [2]Jesse T. Baker
[1]Department of Physics and Electronic Engineering
University of Waikato
[2]Department of Computer Science
University of Waikato
cree@phys.waikato.ac.nz, {mjeff, jtb6}@cs.waikato.ac.nz

Abstract

We present an approach which uses 3D visual landmarks for solving the correspondence problem in Simultaneous Localisation and Mapping (SLAM). The 3D landmarks are computed from camera views of the robot's local space. Using multiple 2D views, identified landmarks are projected, with their correct location and orientation into 3D world space by scene reconstruction. As the robot moves around the local space, extracted landmarks are integrated into a scene representation for the local space which comprises the 3D landmarks. The landmarks for a local space's scene are compared against the landmarks for previously constructed scenes to determine when the robot is revisiting a place it has been to before.

Key Words: robot mapping, visual landmarks, correspondence problem

1. Introduction

In this paper we describe the visual landmark approach we are using to solve the correspondence problem in Robot Mapping. The challenge is to recognise that parts of the environment viewed from different vantage points *correspond* to the same physical space – the correspondence problem. This is regarded as one of the hard problems in Simultaneous Localisation and Mapping (SLAM), where it is often called cycle or loop closing. The robot traverses a cycle in its environment and must recognise that it has returned to a place it has already visited.

The problem is encountered in both topological and absolute metric maps. For absolute metric maps, current localisation methods provide consistent enough local maps but residual error accumulates over large distances. By the time a large cycle is encountered the map will contain significant inconsistencies. Current approaches use some form of probability evaluation to estimate the most likely pose (the robot's x-y location and its heading direction) of the robot given its current observations and the current state of its map [1-3]. Detecting the cycle allows the map to be aligned correctly but also means that the error has to be corrected backwards through the map.

Most topological approaches to spatial mapping partition the environment in some way and link these partitions as they are experienced to form a topological map [4-6]. The advantage of this approach is that global consistency is not an issue because the error cannot grow unbounded, as in absolute metric maps. Consistency is not a problem within the partitions as they are usually around the size of a local environment. State of the art localisation methods are good enough for local environments.

Our approach to mapping the robot's environment extends the cognitive mapping approach of [4]. Yeap and Jefferies [4] developed a Computational Theory of Cognitive Maps which is based on empirical evidence of how humans and animals remember their spatial environment [7]. Yeap and Jefferies' topological map of metric local space descriptions has been implemented on a Pioneer DX mobile robot with minor adaptions to handle input from a laser range sensor. The local space is the space which appears to enclose the robot and is termed the Absolute Space Representation (ASR) to reflect each representation having its own independent frame of reference. The use of the term "absolute" in this sense is confined to an ASR (see [4] for an in depth description of how ASRs are computed).

In this paper we will show how 3D visual landmarks can be used to recognise ASRs that have been visited previously (the correspondence problem). The initial description of the ASR is constructed from laser range data. 2D landmark representations are constructed directly from this representation and are also used as part of our solution to the correspondence problem and the related perceptual aliasing problem. This aspect of our work is described in another paper [8]. While 2D landmarks are computationally less expensive to extract and work well in environments that are rich in features, they sometimes do not contain sufficient context to distinguish one ASR from another. For example, one 90^o corner or doorway looks very much like any other. 2D landmarks are primarily derived from structural elements on the ASR boundary such as walls and exits in

(a) (b) (c)

Fig. 1 Finding a corresponding corner in consecutive views (a) the corner in one image (b) candidate corners in the other image (c) The corresponding corner in the other image

an indoors environment and other objects which block the robot's line of "sight". 3D visual landmarks, on the other hand, are also likely to be interesting objects within the ASR such as furniture in a typical office environment. Combining evidence from 2D and 3D landmarks will give us even better estimates as to when two ASRs belong to the same physical space. However this is outside the scope of this paper. Here we examine how 3D visual landmarks, on their own, can distinguish ASRs.

The initial 2D description of the ASR provides a representation of its extent and a frame of reference within which the 3D visual landmarks can be described. The visual landmarks computed are the distinctive "faces" of objects in the scene located in 3D space. Most of the landmarks will be above the line of sight of the laser range scanner which provides the data from which the ASR is computed. Furthermore, the landmarks need not be inside the local space, merely visible from within it.

The landmarks are constructed from sequential camera views of the local space. Using multiple views, identified landmarks are projected, with their correct location and orientation into 3D world space by scene reconstruction. As the robot moves around the local space, extracted landmarks are integrated into the ASR's "scene" representation which comprises the 3D landmarks, ie. the faces of the objects. The landmarks for an ASR scene are compared against the landmark scenes for previously constructed ASRs to determine when the robot is revisiting a place it has been to before.

2. Using 3D visual landmarks to recognise ASRs in a topological map

There are three main components to the visual landmark recognition. First the coordinate projections required to transform the 2D views to full 3D data are computed. These are determined by recognising strong matching corner features in a pair of images taken at different view points, usually with the robot having moved forward into the scene between the two images. Stereographic reconstruction is used to project the corners into 3D world space. Second, distinctive landmarks of uniform colour and texture are located in the views and using the known projections are projected into their correct location and orientation in 3D world space. Last, recorded distinctive landmarks in previously visited ASRs are tested against those detected in the current ASR for matches.

For the moment the projection of the 2D camera views into 3D world space has been kept separate from the segmentation and projection of distinctive landmarks in the camera view. This reductionist approach has been taken, at this stage, to provide a clear systematic route to implementation of the visual landmark recognition system. Ultimately one would envision developing an approach that calculates the scene reconstruction projection as part of the segmentation of distinctive landmarks, thus providing for better efficiency. We will now discuss each of the three main components of the system in detail.

2.1 Projection from 2D camera views to 3D coordinate space

To calculate the scene reconstruction projections required for projecting 2D camera views into 3D world space, matching points in two camera views are identified. The corners of intensity disparities which lie on landmark boundaries in camera views provide the necessary matching points. The robot's odometry provides the relative position of the two camera views. Fig. 1 shows the corresponding corners in two views. Pollefeys' [9] method with one modification is used for corner detection and projection.

The Harris corner detector [10] is employed to identify and extract corners. It proceeds by applying the Prewitt edge detector in the horizontal and vertical directions, and constructs the smoothed squared image derivatives:

$$l_x = G(dx^2)$$

$$l_y = G(dy^2)$$

$$l_{xy} = G(dx \cdot dy)$$

where *dx* and *dy* are the results of the Prewitt operator, in the horizontal and vertical directions respectively, and G is a smoothing operator that consists of a convo-

lution by a 5x5 pixel kernel of a circularly symmetric Gaussian of $\sigma = 1$. The corner intensity measure [18]

$$c = \frac{l_x l_y - l_{xy}^2}{l_x + l_y}$$

is calculated. The local maxima in c are identified; these correspond to significant corners in the camera view.

The corner detector is applied to two consecutive views between which the robot has moved a small distance. To identify matching corners a small neighbourhood is extracted about each detected corner from the camera view. We take *u* to be the pixel values of the neighbourhood of the first corner and *v* to be the pixel values of the neighbourhood of the second corner. The number of pixels in the neighbourhood is N. The means $\bar{u}$ and $\bar{v}$ are calculated for each corner, as is the cross-correlation χ between the two corners. Corners in the first view are compared to those of the second view by calculating the similarity value.

$$s = \chi(1 - |\bar{u} - \bar{v}|)$$

This similarity value is a modification and improvement to that specified by [9]. High values of *s* indicate potentially matching corners. Fig. 1(a) shows a corner to be matched. Fig. 1(b) shows a slightly different view of the same room as in Fig. 1(a) and the corners detected as possible matches. The correct matching corner is detected along with several false positives. By using the translation of the robot determined from odometry, a prediction of the likely position in the second view of a corner detected in the first view is made and this is used to reject false matches with a high value of *s*. In this manner unique corner matches between the two images are identified. Consider the known movement of the robot in going from the pose associated with Fig. 1(a) to the pose associated with Fig. 1(b). This movement suggests that the corner in Fig. 1(a) must move towards the centre of the image in the second view (Fig. 1(b)). Only one corner in Fig. 1(b) satisfies this requirement. Thus the correct corner has been found.

A perspective projection camera model is assumed [11]. With a corner matched between two views, and knowledge of the relative position of the camera of the two views, one can uniquely determine the location of the corner in 3D world space. Each matching corner is projected into 3D world space giving a set of corresponding points in the two camera views that are fully located in 3D world space. This constitutes the determination of the 3D scene reconstruction projection.

2.2 Landmark construction

The second part of the problem is to identify distinctive landmarks in camera views that can be used for recognising previously visited ASRs. We treat distinctive landmarks as being contiguous regions of relatively constant colour and texture in camera views. A straight-forward method of segmentation, satisfactory for the structured indoor environments the robot is currently being tested in, has been devised to identify the landmarks. We use a size threshold to ensure that small objects, which are more likely to move frequently, are not chosen as landmarks.

First, the magnitude of the gradient is calculated via the Sobel operator over each of the three colour components (red, green and blue) of a camera view. An edge image, constructed by taking the pixelwise maximum of the three Sobel gradient images, is then thresholded to indicate the boundaries of similar-colour regions. The boundary image is inverted (so that edge is background and connected regions are foreground) and each region is identified and uniquely labelled [12].

Two methods are used to project the identified regions into 3D world space. If the region is in a position of the camera view that intersects the laser ranger data (a single horizontal range scan at a fixed height in the world) then the range data is used to locate the region in 3D world space. Should the region not intersect the available range data then the corners of the region are compared to the corners detected as part of the Harris corner detector and information from the previously detected corners are used to project distinctive regions into 3D world coordinates. Fig. 2 shows the landmarks computed for the room depicted.

2.3 Matching landmark configurations

Having detected and localised distinctive landmarks in world coordinates it remains to compare the landmarks of the current ASR with those of previous ASRs for a match. The colour and the shape of the landmark are used for matching. The colour of landmarks are compared by calculating the Euclidean distance in RGB colour space between the two landmarks. The shape of the landmarks are compared by the histogram correlation method described below.

A landmark from a camera view is projected into 3D space then reprojected on to a 2D plane parallel to the plane of the landmark, thus giving the front on view down the normal to the landmark. The boundary of the reprojected landmark is decomposed into straight line segments and the length and angle of each segment is calculated and inserted into a histogram in which the lines are sorted by orientation along the x-axis of the histogram and the lengths of the lines of the same binned orientation are summed to give the frequency axis of the histogram. For shape comparison the correlation of the histograms of reprojected landmarks gives the alignment of pairs of landmarks (essentially a rotation of one of the landmarks until they match). The root mean square (RMS) error is then computed. The RMS is normalised by the energy of the two histograms to give a value between 0 and 1. This constitutes the shape disparity value. The disparity values for the colour of

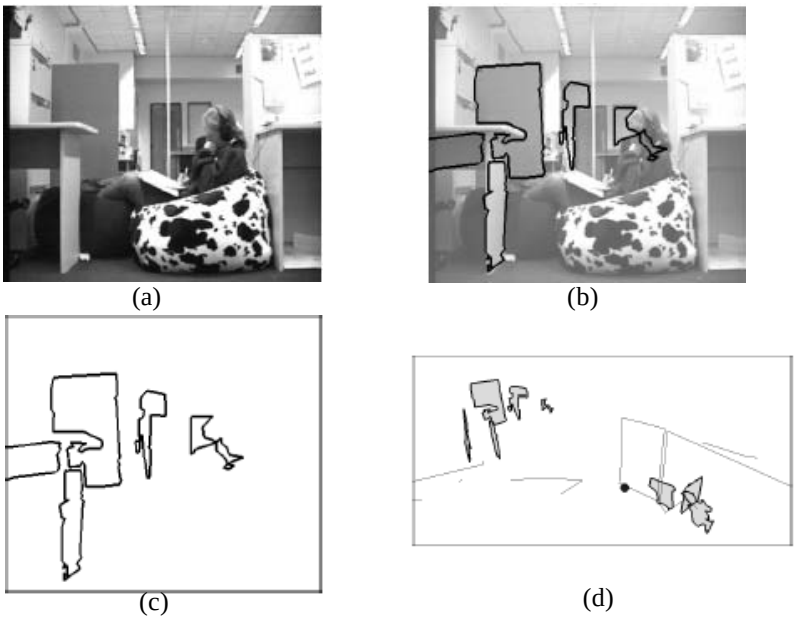

Fig. 2 The landmarks in an ASR scene (a) a view of the room (b) some landmarks overlaid the view (c) the landmarks for the view (d) the landmarks projected into the 3D scene of the ASR.

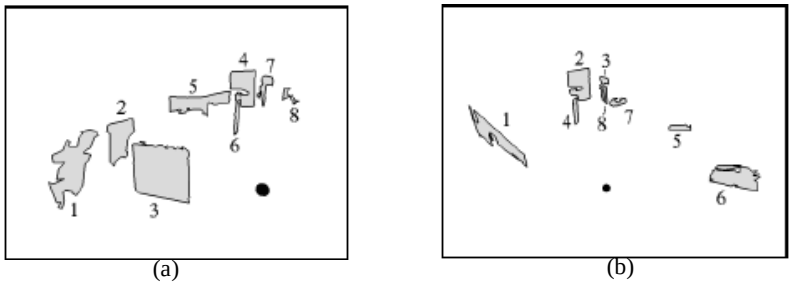

Fig. 3 . The landmark scene representation for two different encounters with the same environment. In (a) and (b) the landmarks are numbered as they are encountered in each separate encounter.

landmark pairs are computed from the Euclidean distance between the means of the pairs in RGB colour space. Fig. 3 shows the 3D landmark scene representation for two different visits to the room depicted in Fig. 2. The comparison of the landmark pairings is shown in Tables 1and 2. The most distinguishing feature is that the match of landmark 4 of Fig. 3 (a) and landmark 2 of Fig. 3 (b) has the smallest shape disparity value by far and the smallest colour disparity. Looking at Fig. 3 (a) and (b), it is obvious that these two landmarks have the most distinctive shape and are really the only two objects that should match. The fact that there are a number of low colour disparity values in the table is not of great concern; it is the combined evidence of both the colour and shape disparities that give confidence of this match. It is also to be noted that only one good matching pair of landmarks is needed. This provides the transformation from one view's coordinate system to the other, hence a combined matching of all landmarks could now be performed in 3D space. This is an area we are currently investigating.

3. Related Work

Most approaches which use vision to recognise locations in the robot's environment have used artificial landmarks [13] or easily detected features such as can be found in ceilings [14]. Recently some approaches have appeared which compute naturally occurring landmarks or features in the robot's surroundings. While not specifically engineered for robot localisation Lowe's [15] approach to matching different views of an object or a scene could well be applied in the robot mapping domain. Lowe matches Scale Invariant Feature Transform (SIFT) features, an approach which transforms image data into scale-invariant coordinates relative to local features. A database of these features is compiled from a set of reference images. Matching of new views is achieved by comparing each feature in the new image against the database of features and finding the best candidate match. Lamon et al. [16] also store a database of features, but in this case they are stored as groupings called fingerprints which characterise a location in the robot's environment. The features are ordered in the fingerprint as they appear in the robot's

immediate surroundings. A new fingerprint is computed for each new view and matched against existing fingerprints in the database. Kosecka and Li [17] represent individual locations in the environment by a set of characteristic views and the SIFT features which are extracted from these views. Hidden Markov Localization is applied to the characteristic views and the SIFT features to determine the robot's location..

Table 1: Comparison of colour disparities
(*a* refers to objects in Fig. 3 (a) and *b* to objects in Fig. 3 (b)

landmark	1.a	2.a	3.a	4.a	5.a	6.a	7.a	8.a
1.b	**0.019**	0.045	0.307	0.188	**0.020**	0.356	0.202	0.130
2.b	0.161	0.222	0.133	**0.012**	0.197	0.183	0.052	0.052
3.b	0.252	0.314	0.048	0.096	0.289	0.088	0.068	0.141
4.b	0.313	0.375	**0.025**	0.146	0.350	**0.027**	0.131	0.200
5.b	0.350	0.412	0.063	0.184	0.387	**0.019**	0.168	0.237
6.b	0.076	**0.014**	0.365	0.247	0.039	0.414	0.260	0.189
7.b	0.073	**0.018**	0.362	0.244	0.038	0.410	0.256	0.186
8.b	0.249	0.311	0.041	0.085	0.286	0.089	0.068	0.137

Table 2: Comparison of shape disparity
(*a* refers to objects in Fig. 3 (a) and *b* to objects in Fig. 3 (b)

landmark	1.a	2.a	3.a	4.a	5.a	6.a	7.a	8.a
1.b	0.241	0.286	0.226	0.112	0.052	0.133	0.155	0.254
2.b	0.127	0.137	0.097	**0.017**	0.089	0.050	0.067	0.120
3.b	0.100	0.100	0.065	0.054	0.138	0.056	0.061	0.087
4.b	0.135	0.135	0.089	0.048	0.123	0.054	0.065	0.125
5.b	0.065	0.076	0.054	0.115	0.210	0.097	0.078	0.073
6.b	0.140	0.148	0.112	0.059	0.083	0.068	0.074	0.134
7.b	0.056	0.072	0.064	0.093	0.170	0.081	0.070	0.055
8.b	0.119	0.101	0.063	0.100	0.194	0.085	0.076	0.099

4. Discussion

We have described a procedure to identify and match distinctive landmarks in differing views of the same local space. It has a number of important features, including invariance to camera view and pose. The invariance to pose was achieved by projecting landmarks in camera views into 3D world coordinates and then reprojecting back into an invariant 2D representations.This enabled efficient 2D histogram correlation shape recognition to be used.

Further testing is required to establish whether this ability to match distinctive landmarks remains when there are many differing ASRs. Can one ASR be identified from a recalled database of landmarks for a number of ASRs?

There are number of improvements to the algorithm that we are considering. The segmentation algorithm could be made more robust to changes in lighting and shading effects by using a colour space that separates colour information from intensity. The colour disparity may be better calculated in a perceptually uniform colour space, such as CIELab, rather than RGB colour space. The establishment of the world 3D coordinate system is currently done separately from the landmark segmentation process. A simpler and more efficient approach would be to integrate the two processes.

We have not yet implemented a mechanism for reasoning with configurations of landmarks. One good match provides the coordinate transformation that will allow the process to begin. In reality, due to variations in the robot's path and occlusions some landmarks could be missing from the configuration. In static environments it is possible to recover the complete set of landmarks by ensuring all of the local space is visited. Howeve, in dynamic environments particular landmarks may be occluded at different times and may disappear altogether. Reasoning with landmark configurations will allow

a match with some degree of confidence even when some landmarks are missing. Given the difficulties in computing reliable shape information across different viewpoints of an object, colour and relative location of landmarks should provide sufficient information for matching configurations. The issue then is the underlying uncertainty which results from occluded and partially occluded landmarks.

5. Conclusion

We have shown how 3D landmarks, the faces of objects in a scene, are computed from camera views of the local space. Using multiple 2D views, identified landmarks are projected, with their correct location and orientation into 3D world space by scene reconstruction. The landmarks for an ASR scene are compared against the landmark scenes for previously constructed ASRs to determine when the robot is revisiting a place it has been to before. The procedure is capable of making strong matches between distinctive landmarks. A successful match is dependent on at least one pair of corresponding landmarks that have been reliably extracted. Removing some of the variance in illumination and a less rigorous shape matching approach should provide more matches than we have obtained here.

References

[1] Hahnel, D., Burgard, W., Fox, D., and Thrun, S. An efficient fastSLAM algorithm for generating maps of large-scale cyclic environments from raw laser range measurements. *in Proceedings Intelligent Robots and Systems*, 2003.

[2] Thrun, S., Hahnel, D., Ferguson, D., Montemerlo, M., Triebel, R., Burgard, W., Baker, C., Omohundro, Z., Thayer, S., and Whittaker, W. A system for volumetric robotic mapping of abandoned mines. *in Proceedings International Conference on Robotics and Automation*, 2003.

[3] Gutmann, J.-S. and Konolige, K. Incremental mapping of large cyclic environments. *in Proceedings International Symposium on Computational Intelligence in Robotics and Automation*, 1999.

[4] Yeap, W.K. and Jefferies, M.E., Computing a representation of the local environment. *Artificial Intelligence*, 107: 265-301, 1999

[5] Kuipers, B., The spatial semantic hierarchy. *Artificial Intelligence*, 119: 191-233, 2000

[6] Tomatis, N., Nourbakhsh, I., and Siegwart, R. Hybrid simultaneous localization and map building: Closing the loop with multi-hypotheses tracking. *in Proceedings International Conference on Robotics and Automation*, 2002.

[7] Gallistel, C.R. and Cramer, A.E., Computations on metric maps in mammals: getting oriented and choosing a multi-destination route. *The Journal of Experimental Biology*, 199: 211-217, 1996

[8] Jefferies, M.E., Weng, W., Baker, J.T., and Mayo, M. Using context to solve the correspondence problem in simultaneous localisation and mapping. *in Proceedings 2004 Pacific Rim Conference on Artificial Intelligence*, 2004.

[9] Pollefeys, M. 3D modeling from inages. *in Proceedings Tutorial Notes 2000 European Conference on Computer Vision*, 2000.

[10] Harris, C. and Stephens, M. A combined corner and edge detector. *in Proceedings 4th ALVETY Vision Conference*, 1998.

[11] Klette, R., Schlüns, K., and Koschan, A., *Computer Vision: Three-Dimensional Data from Images*. Singapore: Springer Verlag, 1998

[12] Gupta, G.S., Win, T.A., Messom, C., Demidenko, S., and Mukhopadhyay, S. Defect analysis of grit-blasted or spray-painted surface using vision sensing techniques. *in Proceedings Image and Vision Computing New Zealand*, 2003.

[13] Taylor, C.J. and Kriegman, D., Vision-based motion planning and exploration algorithms for mobile robots. *IEEE Transactions on Robotics and Automation*, 14(3): 417-427, 1998

[14] Thrun, S., Robot mapping: A survey, in *Exploring Artificial Intelligence in the New millenium*, Lakemeyer, G. and Nebel, B., Editors. Morgan Kaufmann, 2002

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

[16] Lamon, P., Tapus, A., Glauser, E., Tomatis, N., and Siegwart, R. Environmental Modeling with Fingerprint sequences for topological global localization. *in Proceedings IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2003.

[17] Kosecka, J. and Li, F. Vision based topological markov localization. *in Proceedings 2004 International Conference on Robotics and Automation*, 2004.

[18] Noble, J. Descriptions of Image Surfaces. PhD Thesis, Robotics Research Group, Department of Engineering Science, Oxford University,1989.