Local Descriptors for Visual SLAM

Mónica Ballesta* Arturo Gil* Óscar Martínez Mozos† Óscar Reinoso*

Abstract

We present a comparison of several local image descriptors in the context of visual Simultaneous Localization and Mapping (SLAM). In visual SLAM a set of points in the environment are extracted from images and used as landmarks. The points are represented by local descriptors used to resolve the association between landmarks. In this paper, we study the class separability of several descriptors under changes in viewpoint and scale. Several experiments were carried out using sequences of images in 2D and 3D scenes.

1 Introduction

Building a map of the environment is a fundamental skill for a mobile robot, since maps are required for a series of high level tasks. Typical approaches use range sensors to build maps in two or three dimensions (e.g. [5, 6] [3, 14]).

Recently, the interest on using cameras as the main sensors to build the map has increased significantly. Such approach is denoted as visual SLAM. Typically, approaches using vision apply a feature-based SLAM (e.g, [2, 4, 8]), in which significant points in the environment are used as landmarks. Two steps can be distinguished in the utilization of visual landmarks: The detection of interest points and the description of the selected points. The first step involves the selection of suitable points in the images that can be used as landmarks. The points should be detected at different distances and viewing angles, since they will be observed by the robot from different poses. In a second step the landmarks are described by a feature vector which is computed using local image information. The descriptor is used to solve the data association problem: when the robot observes a landmark in the environment, it must decide whether the observation corresponds to a previously seen landmark or to a new one. The data association is a fundamental part of the SLAM process, since wrong associations will produce incorrect maps. In practice, however, the interest points detected in the images are not very stable, and the matching between different views becomes difficult. In consequence,

*Miguel Hernández University. E-mail:{m.ballesta|arturo.gil|o.reinoso}@umh.es. Supported by the Spanish Government under projects DPI2004-07433-C02-01 and PCT-G54016977-2005.

†University of Freiburg. E-mail:omartine@informatik.uni-freiburg.de. Supported by the EU under project CoSy FP6-004250-IP.

the problem of selecting a suitable interest point detector and descriptor for visual SLAM is still open. In a previous work [11], we evaluated some interest point detectors to be used as landmarks in visual SLAM. The Harris corner detector was found to be the most suitable for visual SLAM applications. In this paper we present a comparison of different interest point descriptors using Harris corner detector as point detector.

In [10], Mikolajczyk and Schmid evaluated a set of local descriptors using a criterion based on the number of correct and false matches between pairs of images. Instead, in this work we concentrate on the variation of the descriptor when viewed from different angles and distances. We apply a pattern recognition approach using validity clustering measurements [13] to estimate how well the descriptors representing the same landmark along a sequence are grouped in the different descriptor spaces. This measurements will indicate which descriptor has better separability properties, facilitating the data association. Several experiments have been carried out using sequences of real indoor environment images. We believe that these results would help the selection of visual landmarks for SLAM applications.

2 Visual Descriptors

Next, we list the set of different descriptors that have been evaluated in this study. For all of them we compute the descriptors at the local neighborhood of the points detected by Harris.

SIFT: The Scale-Invariant Feature Transform (SIFT) detects distinctive key points in images and computes a descriptor for them. The algorithm, developed by Lowe, was initially used for object recognition tasks [9]. SIFT features are located at maxima and minima of a difference of Gaussian functions applied in scale space. Next, the descriptors are computed based on orientation histograms at a 4x4 subregion around the interest point, resulting in a 128 dimensional vector.

SURF: Speeded Up Robust Features (SURF) is a scale and rotation invariant descriptor presented by Bay *et al.* [1]. The detection process is based on the Hessian matrix. SURF descriptors are based on sums of 2D Haar wavelet responses, calculated in a 4x4 subregion around each interest point. The standard SURF descriptor has a dimension of 64 and the Extended version (e-SURF) of 128. The u-SURF version is not invariant to rotation and has a dimension of 64.

Gray level patch: This method describes each landmark using the gray level values at a subregion around the interest point. This method has been used in [2] as descriptor of Harris points in a visual SLAM framework.

Orientation Histograms: The orientation histograms are computed from the gradient image, which represents the gray value variations in the x and y direction. In [7] orientation histograms are applied for navigation tasks.

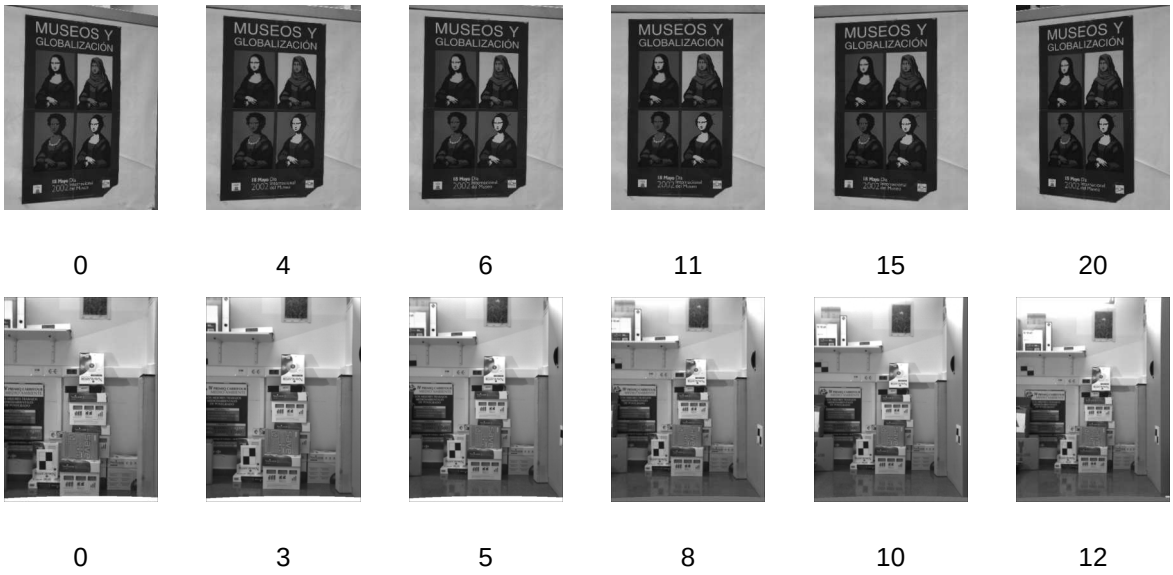

Figure 1: The upper sequence shows a planar object (a poster) under different viewpoints. The bottom sequence depicts a 3D scene under different scale changes.

Zernike Moments: The moment formulation of the Zernike polynomials [15] appears to be one of the most popular in terms of noise resilience, information redundancy and reconstruction capability. They are are constructed using a set of complex polynomials which form a complete orthogonal basis set defined on the unit disc.

3 Descriptor Evaluation

To evaluate the stability of the different interest point descriptors under changes in viewpoint and scale we track each interest point along different images in a sequence. Examples of sequences are shown in Fig. 1. The interest points are extracted using Harris corner detector as shown in [11]. To track the points along the different images we have implemented two different algorithms for 2D and 3D images respectively. In the first case, we used the homography matrix as in [12]. In the case of 3D images, we have implemented a method that is based on the fundamental matrix. This method is divided in two steps. First, seven correspondences between each pair of images are selected, which allows to compute a fundamental matrix F. Using the fundamental matrix F we find a set of preliminary correspondences that are used as input for the computation of a second fundamental matrix F'. In this second step, the fundamental matrix is computed using a RANSAC approach, which results in a more accurate matrix F', that permits to find the correspondences with more precision.

For each tracked interest point p in a sequence of images $S = \{i_1, \ldots, i_N\}$, we obtain a set D_p of descriptor vectors $D_p = \{d_{p_1}, \ldots, d_{p_N}\}$. Each descriptor d_{p_n} represents the interest

Table 1: J_3' values computed in the viewpoint changing sequences

Sequence	SIFT	SURF	e-SURF	u-SURF	Patch	Histogram	Zernike
2D sequences							
1	22.90	36.87	34.18	**126.63**	15.53	2.48	6.39
2	15.89	39.45	34.00	**119.58**	9.03	1.83	2.93
3	10.18	30.49	25.81	**118.64**	6.06	1.85	2.90
4	27.24	68.32	57.81	**184.06**	15.78	2.13	6.54
5	23.75	27.60	28.32	**55.94**	13.59	2.02	5.87
6	13.38	29.45	23.47	**67.36**	6.83	1.77	3.68
3D sequences							
7	5.71	10.70	10.70	**35.93**	2.59	1.46	2.13
8	17.62	16.45	18.96	**73.23**	5.99	1.51	4.33
9	7.11	7.83	7.65	**25.17**	3.33	1.72	2.35
10	16.44	14.47	16.60	**50.58**	7.37	1.54	5.54
11	6.22	9.60	9.41	**30.33**	2.76	1.78	2.25
12	10.26	9.63	11.13	**41.09**	4.00	1.43	3.43

point p in the image i_n. The set D_p forms a cluster in the vector space representing the interest point p in the images along the sequence.

In this work, we use the J_3 separability criterion [13] to measure the separability of the clusters representing the interest points. This measure is based on two scatter matrices: S_w and S_b. S_w is called *within-class scatter matrix*, and measures the compactness of the clusters. The *between-class scatter matrix* S_b measures the separability between vectors belonging to different clusters. In our case, S_w measures the invariance of the descriptor to viewpoint and scale changes, whereas S_b measures the distinctiveness of the points described. The J_3 criterion is defined as:

$$J_3 = trace(S_w^{-1} S_m), \tag{1}$$

where S_m is the *mixture scatter matrix* and is computed as $S_m = S_w + S_b$. A good descriptor has a low value of S_w, since the variability of the vectors describing the same class is small. Furthermore, it is desirable that vectors describing different points are as distinctive as possible, resulting in a high value of S_b. In consequence, a suitable descriptor would have a high value of J_3. This descriptor would have good results in terms of the data association problem, despite of changes in the imaging conditions, such as viewpoint and scale changes. To compare descriptors with different length we use a normalized version $J_3' = \frac{J_3}{N}$, where N is the descriptor length.

4 Experiments

Tables 1 and 2 show the results of applying the J_3' criterion to different sequences of 2D and 3D scenes. The u-SURF descriptor achieves the highest value of separability in 96% of the sequences However u-SURF is not rotational invariant. When comparing only rotational

Table 2: J'_3 values computed in the scale changing sequences

Sequence	SIFT	SURF	e-SURF	u-SURF	Patch	Histogram	Zernike
2D sequences							
1	7.10	3.29	2.87	**8.82**	2.32	1.78	2.15
2	7.97	6.27	5.89	**13.67**	2.59	1.51	2.45
3	9.42	4.47	4.50	**13.03**	3.45	1.92	2.81
4	14.09	7.00	9.05	**26.89**	4.22	1.94	2.70
5	103.36	17.58	38.58	**131.54**	27.73	0.87	14.20
6	4.24	3.51	3.22	**8.56**	2.81	1.12	2.32
7	7.34	4.03	4.90	**12.71**	4.87	1.77	2.73
8	**26.49**	5.99	10.62	22.65	12.34	2.89	9.05
3D sequences							
9	7.06	10.12	10.24	**28.01**	4.47	1.70	3.10
10	14.48	10.39	14.97	**47.48**	5.98	1.67	4.54
11	8.76	9.18	10.02	**24.72**	3.47	2.48	3.95
12	22.22	15.53	23.09	**67.38**	8.50	2.15	5.61
13	6.28	8.84	10.00	**25.56**	3.56	1.94	3.06
14	17.45	11.10	16.86	**42.37**	7.37	2.10	5.88

invariant descriptors, SURF and e-SURF present similar results. In this case, the computational cost of computing the extended version of SURF is not worthy, since the results are not improved substantially. E-SURF always outperforms SIFT in changes in viewpoint. However, in scale changes it is only better in 43% of the cases (2D sequences).

Taking into account the results of Tables 1 and 2 together with the results of our previous work [11], we believe that the u-SURF descriptor in combination with the Harris corner detector is suitable for the common situation in which a robot explores the environment with a camera that only rotates around the vertical axis.

5 Conclusions

We have performed an evaluation of visual local descriptors to be applied for SLAM tasks. For this purpose, we analyzed each descriptor according to its separability. The results of the experiments showed the behavior of seven different descriptors under changes in viewpoint and scale. We believe that this information will be useful when selecting an interest point descriptor as visual landmark for SLAM.

References

[1] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. SURF: Speeded up robust features. In *European Conference on Computer Vision*, 2006.

[2] Andrew J. Davison and David W. Murray. Simultaneous localisation and map-building using active vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2002.

[3] R. Eustice, H. Singh, and J.J. Leonard. Exactly sparse delayed-state filters. In *IEEE Int. Conf. on Robotics & Automation*, 2005.

[4] A. Gil, O. Reinoso, W. Burgard, C. Stachniss, and O. Martínez Mozos. Improving data association in rao-blackwellized visual SLAM. In *IEEE/RSJ Int. Conf. on Intelligent Robots & Systems*, 2006.

[5] G. Grisetti, C. Stachniss, and W. Burgard. Improved techniques for grid mapping with rao-blackwellized particle filters. *IEEE Transactions on Robotics*, 23(1), 2007.

[6] D. Hähnel, W. Burgard, D. Fox, and S. Thrun. An efficient FastSLAM algorithm for generating maps of large-scale cyclic environments from raw laser range measurements. In *IEEE/RSJ Int. Conf. on Intelligent Robots & Systems*, Las Vegas, NV, USA, 2003.

[7] J. Kosecka, L. Zhou, P. Barber, and Z. Duric. Qualitative image based localization in indoor environments. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2003.

[8] J. Little, S. Se, and D.G. Lowe. Global localization using distinctive visual features. In *IEEE/RSJ Int. Conf. on Intelligent Robots & Systems*, 2002.

[9] D.G. Lowe. Object recognition from local scale-invariant features. In *Int. Conf. on Computer Vision*, 1999.

[10] K. Mikolajczyk and C Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10), 2005.

[11] O. Martínez Mozos, A. Gil, M.Ballesta, and O. Reinoso. Interest point detectors for visual slam. In *Proc. of the Conference of the Spanish Association for Artificial Intelligence (CAEPIA)*, 2007.

[12] C. Schmid, R. Mohr, and C. Bauckhage. Evaluaton of interest point detectors. *International Journal of computer Vision*, 37(2), 2000.

[13] S. Theodoridis and K. Koutroumbas. *Pattern Recognition*. Academic Press, third edition, 2006.

[14] R. Triebel and W. Burgard. Improving simultaneous mapping and localization in 3d using global constraints. In *National Conference on Artificial Intelligence (AAAI)*, 2005.

[15] F. Zernike. Diffraction theory of the cut procedure and its improved form, the phase contrast method. *Physica*, 1:689–704, 1934.