Visual Landmark Generation and Redetection with a Single Feature Per Frame

S. Frintrop and A. B. Cremers

Abstract— **In this paper we show that visual landmark generation and redetection is possible with a single feature per frame. The approach is based on the assumption that highly discriminative regions are easily redetectable in subsequent frames as well as in frames visited from different viewpoints. We investigate which feature detectors fit for this purpose and under which conditions the discriminability applies. The approach is tested in a topological localization scenario in which the best feature is tracked over several frames to build landmarks. We show that we can represent a large environment with a few salient landmarks and that a large percentage of these landmarks is robustly redetectable from different viewpoints.**

I. INTRODUCTION

Self localization and navigation belong to the key competences of mobile robots and have been a topic of intensive research during the last decades. Vision-based approaches are of special interest in many applications, since cameras are light-weight, low-cost, passive sensors, that additionally offer rich information about the environment [1], [2], [3]. Visual localization and navigation is often based on landmarks, that means on objects or regions in the environment that serve as reference points for the robot. Ideally, they shall be easily redetectable from different viewpoints, under changing illumination conditions, and in the presence of disturbances such as walking people.

The first step of visual landmark detection is usually the feature detection. However, not all detected features are useful landmark candidates. Especially corner features are often detected at intersections of objects and thus not stable [4]. Furthermore, most feature detectors obtain a feature repeatability of 50 – 80%, depending on the scene and the transformation between frames [5]. That means, a large amount of the detected features is not redetected in a following frame. Only a few of the features are stable enough to survive the tracking over several frames. To find stable landmarks, a common approach is to extract a large amount of features (usually several hundred per frame), track them over several frames and keep only the most stable ones [4].

However, detecting, matching and storing of hundreds of features per frame and the comparison to a large image database is costly. Robots usually have to operate in real-time and additionally have to share resources between different modules and tasks. While there have been successful approaches to deal even with large amount of features [6], [7], it is certainly preferable if it is possible to solve the task with less features. Desirable would be to know in advance which features will turn out to be stable and thus will be good candidates for landmarks.

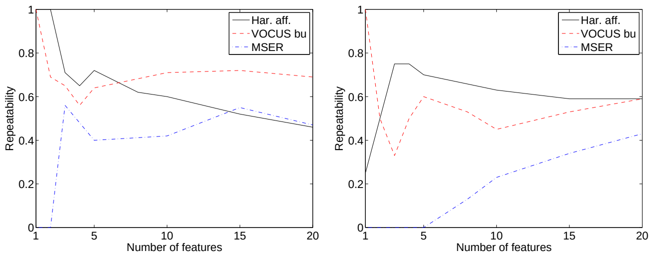

Fig. 1. Repeatability depending on number of features per frame. Features are selected by their quality as defined in sec. II-A.4. Examples determined on data sets 1 and 5 of Fig. 3 for 4 different viewpoints (after 50, 100, 150, and 200 frames) for Harris-affine regions, MSERs, and bottom-up salient regions (VOCUS bu). Two typical cases occur: either the best feature is very poor or extremely stable.

When investigating sparse sets of features (1–20 features per frame, features selected by their quality as defined in sec. II-A.4), we found two typical cases for the distribution of the repeatability values: Before converging to stable repeatability values, the repeatability of the best feature was either very poor or extremely high, often reaching 100% repeatability (cf. Fig. 1). This behaviour depended on the scene and was observed for all the investigated detectors. The proportion of poor versus high performance cases however differed among the detectors.

Outgoing from this observation, we pose the following questions: is it possible to exploit the fact that the best feature often is extremely stable? Is it possible to generate landmarks and redetect them reliably with only one feature per frame? For topological localization, it is in principle enough to have one landmark every few meters. The robot does not have to know its exact position and it is not necessary to see a landmark in each frame, as long as the scene is recognized reliably from time to time. A certain redundancy is necessary anyway since some landmarks may be occluded or removed upon revisiting that place, but as long as a few stable landmarks per environment remain, this is sufficient.

In this paper, we show that topological localization is possible with a single feature per frame. First, we investigate which feature detectors are suitable to be restricted to a sparse set and which quality measure suits to determine the best feature. We investigate Harris-affine regions [5], maximally stable extremal regions (MSERs) [8] and a saliency detector [9]; we finally chose the last one for

The authors are with the Institute of Computer Science III, University of Bonn, 53117 Bonn, Germany {frintrop}/{abc}@iai.uni-bonn.de

further investigations and show that it is possible to build stable landmarks from the most salient feature. In a scene classification experiment, we show that a reliable redetection of landmarks from different viewpoints is possible and that a test sequence can be reliable allocated to the correct scene.

Feature selection has been investigated before in several ways. In applications in which training data is available, machine learning methods often determine the best features for a class of objects from a pool of training images [10], [11]. The reduced set however contains usually still several dozens of features per frame or object and reduces only the features in the database, not the ones obtained during testing. Other approaches compare descriptors applied to the detected regions and keep only the most discriminative ones [12]. The main difference in our approach is that we start much earlier with the preselection, namely already during feature selection. In applications in which no training data is available, e.g. visual SLAM (simultaneous localization and mapping), some people use thresholds to reduce the number of features. E.g., [2] only add features to their map if the number of features visible in the robot view is below a threshold and [3] keep only landmarks that perform well over a sequence of frames. Preliminary investigations on the repeatability of a single stable feature have been made in [13]. Here, we extend this study by using detectors that are known to perform well in other applications (Harris-affine, MSERs), by introducing a more adequate repeatability measure, and by performing more detailed experiments. Completely new in this paper is the integration of the single-feature approach into a topological localization scenario.

II. Feature Detection

In this section, we discuss and evaluate the feature detection. First, we describe the investigated feature detectors and the quality measure to determine the best feature (sec. II-A). Second, we present the performance measure repeatability and extend the definition to image sequences (sec. II-B). Finally, we investigate in several experiments which feature detector provides the most stable feature in tracking and redetection situations (sec. IV-A).

A. Feature Detectors

1) Harris-Affine Regions: Harris-affine regions are computed by detecting interest points with the Harris detector in scale-space and determining an elliptical region for each point based on the second moment matrix of the intensity gradient [5].[1] For each pixel $\vec{x} = (x, y)$, the Harris detector determines its *cornerness* $c(\vec{x})$ (also strength or Harris response) as $c(\vec{x}) = \det(M) - \alpha trace^2(M)$, where M is the second moment matrix describing the local neighborhood of $\vec{x}$. This detector is applied to multiple scales and the characteristic scale is chosen to obtain scale-invariance. Finally, the affine region is determined according to [14]. If the cornerness exceeds a certain threshold, the pixel is defined as a corner.

2) MSERs: Maximally Stable Extremal Regions (MSERs) were introduced in [8] and have shown high repeatability results under various image transformations [5].[2] The MSER algorithm first detects several nested sets of extremal regions $Q_1, ..., Q_k$. Each Q_i is a region such that for all pixels $p \in Q_i, q \in \partial Q_i : I(p) > I(q)$ (MSER+) or $I(p) < I(q)$ (MSER-), where ∂Q_i is the boundary of Q_i, consisting of pixels that are adjacent but do not belong to Q_i, and $I(p)$ is the intensity value of p. A region Q_i is maximally stable iff the stability $q(i) = |Q_{i+\Delta} - Q_{i-\Delta}|/|Q_i|$ has a local minimum at i. Usually, a fixed Δ is used. This however results often in a set of regions with the same stability value (e.g. $q(i) = 0$) making it impossible to determine a single best MSER. Increasing Δ results in fewer regions with higher repeatability but usually lower $q(i)$, while a too large Δ might result in no MSERs in certain images. In our approach, we increase the Δ automatically until the MSER with the lowest $q(i)$ is non-ambiguous.

3) Biologically-inspired salient regions: Biologically-inspired attention systems compute the saliency of regions based on concepts of the human visual system [15]. They have shown to outperform other methods such as intensity contrasts, local oriented edge density, or entropy in terms of predicting human eye movements [16]. Here, we use the attention system VOCUS [9] that is real-time capable (20 ms for a 320×240 pixel image, on a 2.5 GHz PC [17]) and has a top-down part to search for targets.

VOCUS creates a saliency map by computing image contrasts and uniqueness of a feature. The feature computations for the features intensity, orientation, and color are performed on 3 different scales with image pyramids. Two intensity feature maps, for on-off and off-on contrasts, are computed by *center-surround mechanisms*. Similarly, 4 orientation maps ($0°, 45°, 90°, 135°$) and 4 color maps (green, blue, red, yellow) are computed (cf. [9]).

The core of the saliency detector is the *uniqueness weight* that is applied before feature channels are fused: a feature which occurs seldomly in a scene is assigned a higher saliency than a frequently occurring feature. This is a mechanism which enables humans to instantly detect outliers. The uniqueness $\mathcal{W}$ of map X is computed as $\mathcal{W}(X) = X/\sqrt{m}$, where m is the number of local maxima that exceed a threshold. Note that this weighting, together with the parallel investigation of different feature channels, distinguishes this detector from standard detectors such as Harris corners or MSERs because it considers the global instead of the local discriminability of a region.

The weighted feature maps are summed up to 3 conspicuity maps I (intensity), O (orientation) and C (color) and combined to the *saliency map*: $S_{bu} = \mathcal{W}(I) + \mathcal{W}(O) + \mathcal{W}(C)$. The salient regions, the *VOCUS-ROIs*, are the local maxima in S above a threshold, extended to a region with a region growing approach [18].

4) Sorting features: To determine the best features, we need a measure for the quality. This depends on the detector,

[1] We used the detector from http://www.robots.ox.ac.uk/∼vgg/research/affine

[2] We used the MSER code from the VLFeat library: http://www.vlfeat.org

for Harris-affine regions we chose the cornerness, for MSERs the stability, for VOCUS-ROIs the saliency. This results for each detector in an ordered list of features $F_i = (F_1, ..., F_n)$, where F_1 is the best feature and F_j has a higher quality than F_{j+1}.

B. Performance Measure: Repeatability

The performance measure to compare the stability of features is the repeatability that is defined as follows:

$$R(I_j, I_k) = \frac{\textit{\# features in } I_j \textit{ with correspondence in } I_k}{\textit{\# features in } I_j} .$$

for parts of the scene visible in both frames I_j and I_k. To be a valid correspondence, about 50% of the regions have to overlap. This allows a relatively large overlap error but a powerful descriptor is still able to match such regions successfully (cf. [5]). A symmetric measure can be obtained as follows:[3]

$$R_{sym}(I_j, I_k) = (R(I_j, I_k) + R(I_k, I_j))/2 .$$

To extend the repeatability definition to image sequences or sets, we distinguish two different versions: we define the *tracking repeatability* as the average repeatability between consecutive frames:

$$R_T(I_{1:t}) = \frac{\sum_{i=2}^{t} R_{sym}(I_{i-1}, I_i)}{(t-1)} .$$

for an image sequence $I_{1:t} = I_1, ..., I_t$. It is called tracking repeatability because it is mainly of interest when features are tracked over frames. The *viewpoint repeatability* on the other hand is defined as the average repeatability between a frame I_i and the remaining images of the sequence or set:

$$R_V(I_i, I_{1:t}) = \frac{\sum_{j=1, Ij \neq Ii}^{t} R_{sym}(I_i, I_j)}{(t-1)} .$$

It is called viewpoint repeatability because, in contrast to tracking, the viewpoint between considered frames might change strongly, usually the more the longer the sequence.

III. Landmark Generation

While a feature is a 2D region in an image, a landmark is a region in the 3D world that can be observed from different viewpoints. To create landmarks, the detected feature is tracked over several frames. The resulting list of features represents a landmark. The *length* of a landmark is the number of elements in the list, which is equivalent to the number of frames the feature was detected in.

To compute the landmarks, we match new features to features from previous frames whereas we allow gaps of up to 2 frames. We finally consider only landmark with a length $\geq k$ (here: $k = 5$). This enables to determine which landmarks are stable over time.

To match two features, we use the SIFT descriptor [12] that has outperformed most other descriptors in terms of matching performance [19]. Usually, SIFT descriptors are computed at intensity extrema in scale space [12] or at Harris-Laplacians [5]. Here, we calculate one SIFT descriptor for each VOCUS-ROI. The center of the ROI provides the position and the size of the ROI determines the size of the descriptor grid. The grid should be larger than the ROI to allow catching information about the surrounding but should also not include too much background and stay within the image borders[4]. The procedure to generate landmarks is illustrated in Fig. 2.

IV. Experiments

In our experiments, we investigate three questions. First: Which is the best feature detector for our purpose? This experiment investigates the repeatability in tracking situations as well as under strong viewpoint changes. Second: Is it possible to create stable landmarks from a single feature per frame? And third: Can localization be performed based on such a sparse landmark representation?

A. Which is the best feature detector for our purpose?

To test which feature detector suits best for our purpose, we investigated the repeatability of features on 7 image sequences of 200-400 frames of size 320 × 240 (cf. Fig. 3). In all sequences, strong viewpoint changes occur. Data sets 1–4 show natural scenes in an office environment and contain objects which were especially designed to be salient for humans: a green exit sign, a magnet clamp, a red circle containing a warning remark, and, in data set 4, a fire extinguisher and a red piece of paper at the wall. The last 3 data sets show natural, cluttered office environment scenes. We investigated the tracking repeatability as well as the viewpoint repeatability on these data sets. The results are displayed in Table I. As to be expected, the tracking repeatability is almost always higher than the viewpoint repeatability. Worth to note is also that the viewpoint repeatability naturally goes down the more the viewpoint changes.

It turns out that the Harris regions as well as the salient VOCUS-ROIs perform well in most cases, whereas the MSERs show a considerably lower performance. The VOCUS-ROIs outperform the Harris regions on average since the attention system is able to capture the uniqueness of features in more cases. The low performance of the MSERs can be explained as follows: usually, MSERs are stable, if all possible MSERs in a scene are considered (as in [5]). But, since all MSERs have an equal stability value, it is hard to determine a stable subset or even a best feature. So, if reduction of the number of features is desired, the other detectors seem to be the better choice.

We decided to use the salient VOCUS-ROIs for our application, first, because they yielded the highest repeatability

[3] The symmetric measure in [5] divides instead by the smaller of the number of regions in both frames. This might however result in problems if the number of features in the 1st frame is a subset of the features in the 2nd frame. The measure would report a repeatability of 100%, even if the number of features in the 2nd frame is considerably larger. This is especially a problem for small numbers of features.

[4] We chose a grid size of 1.5 times the maximum of width and height of the ROI.

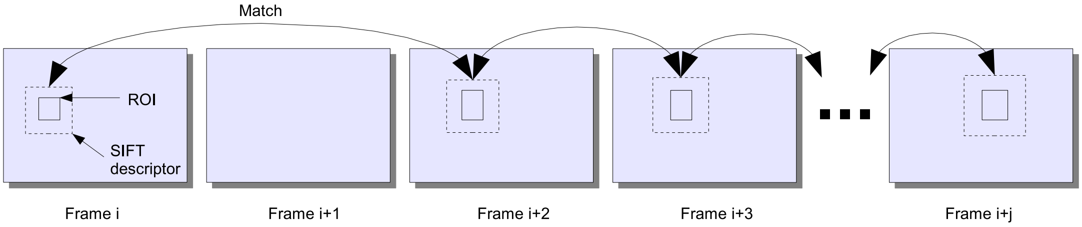

Fig. 2. The process to generate landmarks: For each feature (ROI, solid rectangle), a SIFT descriptor is computed (area in dashed rectangle). The descriptors of the ROIs of consecutive frames are compared. If they match, a landmark is created. Gaps of up to 2 frames are allowed and only landmarks of length $\geq k = 5$ are considered.

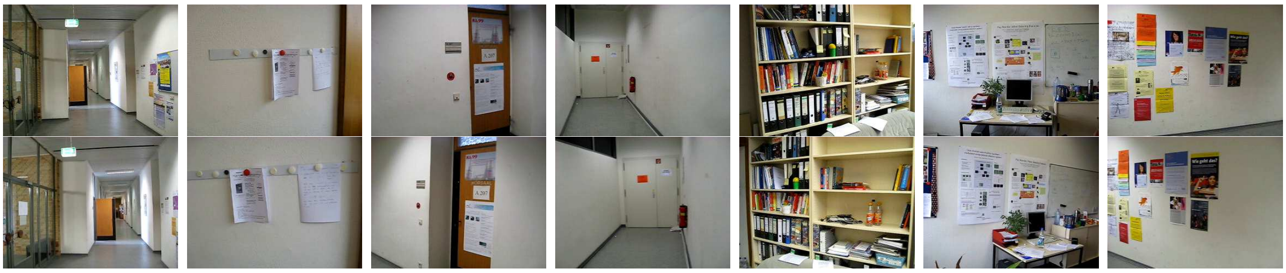

Fig. 3. Data sets. 1st row: 1st frame, 2nd row: last frame of sequence

data set	# frames	Tracking repeatability [%] Harris	MSER	VOCUS-ROI	Viewpoint repeatability [%] Harris	MSER	VOCUS-ROI
1.	259	96	25	97	96	18	97
2.	210	76	80	100	78	89	100
3.	315	94	20	90	77	5	82
4.	254	95	33	83	92	21	29
5.	254	85	5	100	19	11	100
6.	209	76	13	82	61	1	14
7.	341	86	10	84	19	9	72
av.		87	27	91	63	22	71

TABLE I

THE TRACKING REPEATABILITY $R_T(I_{1:t})$ AND THE VIEWPOINT REPEATABILITY $R_V(I_1, I_{2:t})$ OF THE BEST FEATURE F_1 (SELECTED ACCORDING TO SEC. II-A.4) ON THE DATA SETS OF FIG. 3.

and second, because it is possible to adapt the attention system to search for expected regions in a top-down manner. We plan to exploit this in future work.

B. Is it possible to create stable landmarks from a single feature per frame?

In this section, we investigate whether the VOCUS-ROIs can be used to create stable landmarks. A stable landmark should be visible over several frames and should be redetectable under viewpoint and illumination changes. We tested our approach in 5 scenes of a typical office building: 3 corridors on different levels of the same building (scene 2,3,4) and two open areas (scene 1 and 5) (cf. Fig. 4). The corridors, especially scene 3 and 4, are very similar, resulting in matching ambiguities. The experiments were performed during normal working hours, i.e. people walked around, doors were opened or closed etc. In each of the scenes, we recorded two image sequences (denoted a and b in the following) with a mobile camera mounted on a moving vehicle. Each track had a length of about $100\,m$, images had a resolution of 320×240.

First, we test whether a single feature per frame is sufficient at all to build landmarks. Remember that a feature has to be seen and matched over at least 5 frames to become a valid landmark. Thus, if repeatability is too low, the system will not create any landmarks. The results are shown in Table II. We obtained between 9 and 62 landmarks per scene, depending on the length of the sequence. Each landmark consists of $7 - 16$ ROIs, on average 10 ROIs. That means, a feature that was used to create a landmark was on average visible over 10 frames. This shows that it is possible to create landmarks even from a single feature per frame. The approach can also be applied for tasks like visual SLAM (simultaneous localization and mapping) in which no previous training is possible.

Next, we investigate whether these landmarks can be redetected under viewpoint and illumination changes. Especially for a sparse landmark representation this is not obvious and has to be investigated further.

To test the redetection of landmarks, we divided the image

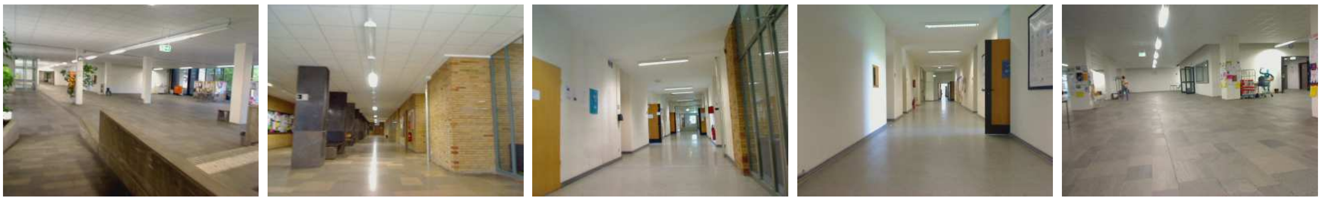

Fig. 4. Example frames of the 5 scenes we investigated for scene recognition

TABLE II
LANDMARK GENERATION

Scene	# Frames	# landmarks	av. # ROIs per LM
1.a	539	13	8
1.b	598	26	10
2.a	1194	31	7
2.b	1144	56	9
3.a	828	32	12
3.b	749	17	10
4.a	1720	62	16
4.b	1064	48	11
5.a	580	26	12
5.b	568	9	7

TABLE III
LANDMARK REDETECTION. LEFT COLUMN: S_i/S_j MEANS THAT LANDMARKS WERE OBTAINED FROM REFERENCE SEQUENCE S_i AND REDETECTED IN TEST SEQUENCE S_j.

Scene	redetected LMs [%]
1.a/1.b	84
2.a/2.b	83
3.a/3.b	75
4.a/4.b	61
5.a/5.b	73
1.b/1.a	69
2.b/2.a	79
3.b/3.a	94
4.b/4.a	38
5.b/5.a	78
average	73

sequences into train and test sequences. In a first run, the sequences denoted by a (1.a, 2.a, ..., 5.a) are used as training data $S_i, i \in \{1, .., 5\}$, the ones denoted by b as test sequences $S_j, j \in \{1, .., 5\}$. In a second run, we applied the sequences vice versa. The redetection ratio was determined by matching the detected VOCUS-ROIs of each frame of test sequence S_j to all landmarks obtained from the training sequences $S_i, i \neq j$. (Remember that only one of these sequences is from the same environment as S_j, the other sequences are from different environments.) Some matching examples are displayed in Fig. 5; the percentage of redetected landmarks is shown in Tab. III. It shows that generally the majority of the landmarks, on average 73%, is redetected in a test sequence. Thus, stable landmarks can be created from a single feature per frame and reliably redetected.

C. Can we perform topological localization with such a sparse landmark representation?

In this section, we show that the sparse landmark representation that we obtained in the previous section can be used to reliably localize a system in an office scenario. We use the same sequences as in the previous section and show that we can reliably assign the correct location to a sequence of images. To show this, we cross-validated the matching performance of all sequences to each other, i.e. we considered one sequence S_i as training data and another sequence S_j as test data. For each sequence combination (S_i, S_j) we compute the confidence that the test sequence S_j was obtained in the same environment as the reference sequence S_i:

$$C(S_i, S_j) = \frac{M(S_i, S_j)}{\sum_{k=1, k \neq j}^{N} M(S_k, S_j)}, \quad \forall i \neq j$$

where N is the number of sequences, here $N = 10$, and $M(S_i, S_j)$ denotes the number of ROI matches between S_i and S_j.

The confidence values are shown in Tab. IV. It can be seen that the confidence values for test sequences from the same environment as the reference sequences are considerably higher (bold numbers). In most cases, they are between 95 and 100%. Only the matching confidences for scenes 3 and 4, two very similar corridors, are a little lower. The similarity of the two scenes results in several false detections. Still, the confidence for the correct sequence is always more than three times as high as the confidence for each other sequence. The final decision of the robot for a test sequence S_j is:

$$Estimated\ scene = argmax_i\ C(S_i, S_j) \qquad (1)$$

Based on this decision rule, the system determines the correct scene for all of the test sequences. Thus, we have shown that it is possible to reliably localize a system based on a single feature per frame. This is also applicable if no training phase is possible as in visual SLAM.

V. DISCUSSION AND CONCLUSION

In this paper, we have shown that visual localization and scene recognition is possible with a very sparse landmark representation. Focusing on the most salient feature in a frame enables to select the most discriminative regions in an environment as landmark candidates. While this approach does not detect landmarks in each part of the environment (if there is nothing salient, no landmarks are found), it works well as long as the environment contains some discriminative parts. Especially human-made environments have plenty of such salient objects: fire extinguishers, exit signs, doors or posters can serve as valuable landmarks. The advantage of such landmarks is that they are easily redetected from

TABLE IV

THE CONFIDENCE VALUES FOR THE LANDMARK MATCHING. ROWS DENOTE THE REFERENCE SEQUENCES S_i, COLUMNS THE TEST SEQUENCES S_j. BOLD NUMBERS HIGHLIGHT THE HIGHEST VALUE IN A COLUMN.

	1.a	1.b	2.a	2.b	3.a	3.b	4.a	4.b	5.a	5.b
1.a		**100**	0	0	0	0	0	0	0	0
1.b	**100**		0	0	0	0	0	0	0	0
2.a	0	0		**99**	2	1	0	1	0	0
2.b	0	0	**95**		1	1	0	0	1	0
3.a	0	0	3	1		**82**	11	19	1	1
3.b	0	0	1	0	**75**		7	11	0	0
4.a	0	0	0	1	8	5		**68**	0	0
4.b	0	0	1	0	14	1	**82**		0	0
5.a	0	0	0	0	1	0	0	0		**98**
5.b	0	0	0	0	1	0	0	0	**98**	

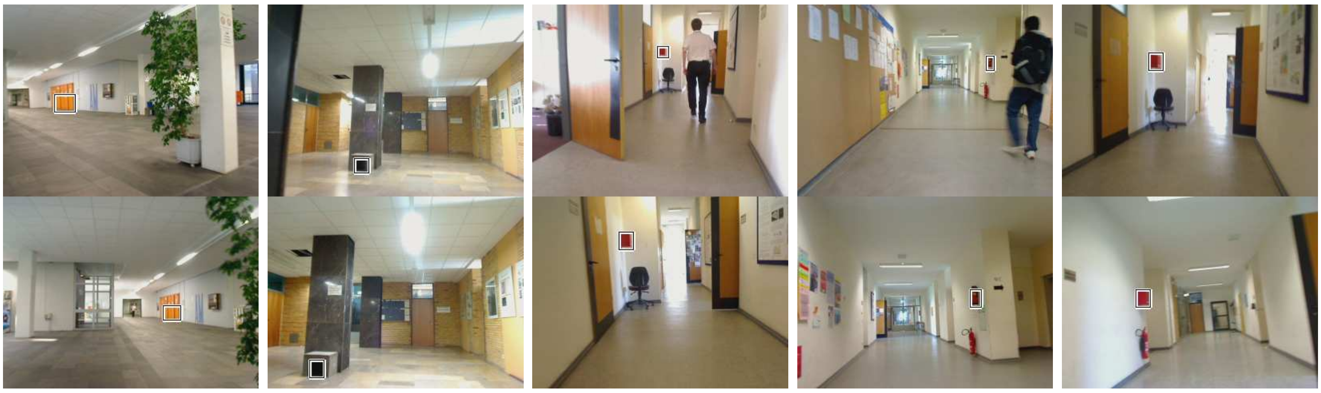

Fig. 5. Some examples of matching ROIs. Top: test sequence, bottom: reference sequence. Four successful matches and one false match (right) are shown.

different viewpoints. We show in several experiments that the one-feature-per-frame approach is well suited for landmark generation and redetection.

While the approach works well in the presented setting, it could be even improved with active camera control and top-down feature search. This would enable the robot to actively search for salient landmarks. In future work, we plan to integrate the one-feature approach to a SLAM setting with active camera control as the one in [3]. We also plan to investigate how the approach copes with long-term changes in the environment.

REFERENCES

[1] S. Se, D. Lowe, and J. Little, "Mobile robot localization and mapping with uncertainty using scale-invariant visual landmarks," *Int'l J. of Robotics Research*, vol. 21, no. 8, pp. 735–758, August 2002.

[2] A. Davison and D. Murray, "Simultaneous localisation and map-building using active vision," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 24, no. 7, 2002.

[3] S. Frintrop and P. Jensfelt, "Attentional landmarks and active gaze control for visual SLAM," *IEEE Trans. on Robotics, Special Issue on Visual SLAM*, vol. 24, no. 5, Oct 2008.

[4] J. Shi and C. Tomasi, "Good features to track," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 1994.

[5] K. Mikolajczyk and C. Schmid, "A comparison of affine region detectors," *International Journal of Computer Vision (IJCV)*, vol. 65, no. 1-2, pp. 43–72, 2005.

[6] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2006.

[7] J. Wang, H. Zha, and R. Cipolla, "Coarse-to-fine vision-based localization by indexing scale-invariant features," *IEEE Trans. on Systems, Man, and Cybernetics – Part B: Cybernetics*, vol. 36, no. 2, 2006.

[8] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide baseline stereo from maximally stable extremal regions," in *Proc. of the British Machine Vision Conference*, 2002.

[9] S. Frintrop, "VOCUS: a visual attention system for object detection and goal-directed search," Ph.D. dissertation, Rheinische Friedrich-Wilhelms-Universität Bonn, Germany, July 2005, published 2006 in Lecture Notes in Artificial Intelligence (LNAI), Vol. 3899, Springer.

[10] A. Pronobis and B. Caputo, "Confidence-based cue integration for visual place recognition," in *Proc. of IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems (IROS'07)*, 2007.

[11] F. Li and J. Kosecka, "Probabilistic location recognition using reduced feature set," in *ICRA*, 2006, pp. 3405–3410.

[12] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int'l J. of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[13] S. Frintrop, "The high repeatability of salient regions," in *Proc. of ECCV workshop "Vision in Action: Efficient Strategies for Cognitive Agents in Complex Environments"*, 2008.

[14] T. Lindeberg and J. Garding, "Shape-adapted smoothing in estimation of 3-D shape cues from affine deformations of local 2-D brightness structure," *Image and Vision Computing*, vol. 15, no. 6, pp. 415–434, 1997.

[15] A. M. Treisman and G. Gelade, "A feature integration theory of attention," *Cognitive Psychology*, vol. 12, pp. 97–136, 1980.

[16] L. Itti and P. Baldi, "Bayesian surprise attracts human attention," in *Advances in Neural Information Processing Systems (NIPS)*, vol. 19. Cambridge, MA: MIT Press, 2006, pp. 547–554.

[17] S. Frintrop, M. Klodt, and E. Rome, "A real-time visual attention system using integral images," in *Proc. of the 5th Int'l Conf. on Computer Vision Systems (ICVS)*, Bielefeld, Germany, March 2007.

[18] R. Adams and L. Bischof, "Seeded region growing," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 16, no. 6, pp. 641 – 647, 1994.

[19] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Transactions of Pattern Analysis and Machine intelligence*, vol. 27, no. 10, 2005.