Combining SURF and SIFT for Challenging Indoor Localization using a Feature Cloud

Marvin Ferber, Mark Sastuba, Steve Grehl, Bernhard Jung
Institute of Computer Science
Technical University Bergakademie Freiberg, Germany
Email: {marvin.ferber, mark.sastuba, steve.grehl, jung}@informatik.tu-freiberg.de

***Abstract*—Indoor localization for smartphone users enables applications such as indoor navigation or augmented information services. Indoor localization can be achieved by using camera images to resolve the position based on a precomputed training set of images. This technique is widely known as image-based localization. In particular, we create a feature cloud from a Structure from Motion (SfM) approach as training set. At runtime, a feature-based matching identifies similarities between a test image and the trained set in order to solve the perspective n-point (PNP) problem and compute the camera position. Since indoor environments are challenging regarding wall structure, light conditions and glass elements, we combine SIFT and SURF image features to exploit the advantages of both techniques and, thus, provide a highly robust localization technology. We can even show that our novel approach can be used for a realtime image-based localization of a smartphone using remote processing.**

I. Introduction

There is an increasing need for accurate indoor localization using lightweight equipment such as a smartphones or even drones. Such lightweight equipment often lacks range sensors for 3D perception. The sensor information from, e. g., camera, WLAN strength or inertial sensors is limited when compared to, e. g., depth sensors or GNSS systems that are available outdoors. As a result, monocular indoor localization is an active filed of research[1], [2]. However, most techniques are either not robust enough or they require high computational power, which thwarts the execution on a smartphone.

In this paper, we present first results on a novel approach that addresses performance by relocating computationally heavy software parts to the cloud. Furthermore, our approach achieves a high robustness facing indoor conditions by combining multiple image feature detectors such as SIFT and SURF. We compute a 3D model (feature point cloud) of the indoor scene prior to the actual localization. No additional instrumentation of the environment is considered. Another benefit of our approach is its simplicity. Only a minor set of parameters is necessary for the setup.

In this paper, we highlight specific obstacles in the process of building a 3D model from an indoor environment such as blending sunlight through windows or dark areas due to missing ceiling lights. To better understand those implications, our contribution also includes a comparison of SURF and SIFT in indoor environments.

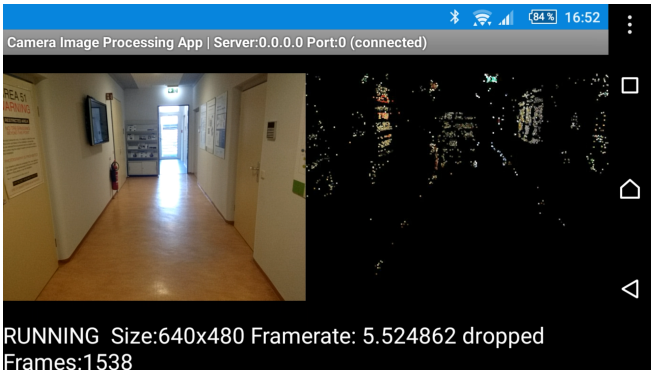

Fig. 1: Android app implementing our localization approach, *left*: camera images, *right:* remotely computed camera pose in a feature point cloud.

An implementation of our localization approach in an Android app is shown in Fig. 1. The app shows a split window with the current camera image on the left and a view from the computed camera pose in the feature cloud on the right. We are able to deliver a localization rate of up to 5 Hz using our approach. The client/server communication is realized using the MAP middleware that already provides streaming image processing and asynchronous remote task execution [3].

This paper is organized as follows. First, we give a brief overview of related work in the field. Some challenging examples are presented in Section III. Section IV introduces our approach. A first evaluation is presented in Section V. Section VI concludes the paper.

II. Related Work

We focus on related works including monocular approaches to 3D model creation and indoor localization with the possibility of realtime computation.

Monocular approaches to Simultaneous Localization and Mapping (SLAM) such as ORB-SLAM [1] have emerged recently. The model is built from a monocular camera image stream. The SLAM approach includes the

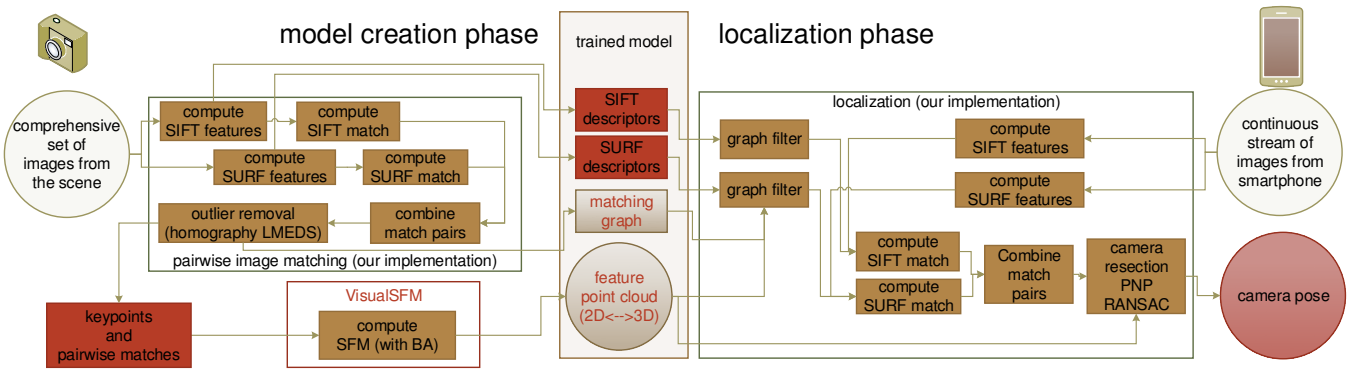

Fig. 2: Overview of our combined SIFT and SURF approach showing the non-recurrent model creation phase, the components of the trained model, and the localization phase using, e. g., camera images from a smartphone.

localization even while the model is built, but can also be used without altering the model as a pure localization. We will evaluate our work against these in a future work.

Our work is based on photogrammetry methods for 3D model creation using the pipeline of Structure from Motion (SfM). SfM is the used to create a 3D point map from the scene using various images. Since SfM extends the 3D map incrementally image by image, a global optimization of the projections (BA) is necessary afterwards. The result is a feature point cloud containing all valid match points that have been used for the 3D reconstruction together with all optimized poses of all cameras that were used to create the model. An extensible software that implements most of this functionality for a pinhole camera model is VisualSFM [4].

A fast indoor localization using the SfM pipeline and VisualSFM similar to our approach is proposed in [2]. It uses the SIFTGPU implementation. The localization is improved by an initialization from an RSSI-based WLAN localization and speed up by a zone-based search space partitioning for the feature cloud. The evaluation does not contain information on the test image quality and camera orientation.

The combination of SIFT and SURF feature detectors can lead to improved detection results as described in [5]. This work targets face recognition. To the best of our knowledge, we are the first to propose a combination of SIFT and SURF for 3D model creation and localization.

III. Feature matching in an indoor setting

Both the creation of a feature cloud and the localization of images (camera poses) are very sensitive to the feature extractor and descriptor used. We highlight some findings when matching images in an indoor environment. The goal is to find many similar image attributes in two images to extract 3D information. Our overall approach is divided into a model creation phase and an arbitrary number of localization phases afterwards, see Fig. 2 for an illustration. In the model creation phase pairwise image matchings have to be identified. Therefore, the keypoints and descriptors are computed first. Then the descriptors are matched finding similarities. A match is a set of two points in two images that describe the same 3D point in the real world. These matches are detected in four steps:

1) compute the best two matches for each descriptor in image A compared to B using L1 norm and remove non-unique matches for better robustness,
2) repeat step 1 using interchanged images,
3) remove matches that do not appear in both results result of step 1 and step 2 (only cross matchings are valid),
4) find the transformation between images A and B (homography) and remove outliers that are considered false positives. We use the LMEDS approach to find the homography.

Feature detectors are varying in robustness against scale, orientation, light conditions, and other attributes. A 3D model can only be created having images that show each scene from different view angles, e. g., 180° in an office floor. We show pairwise image matchings using SIFT and SURF of three challenging areas to show the qualitative differences in Fig. 3.

Fig. 3 (top) shows the robustness against a huge angle between two pictures of approx. 130°. SIFT can handle this situation better. So, less pictures are necessary to reconstruct an indoor scene, because wider angles between pictures are tolerable. A second challenging area has been identified around a glass door with a window in the background. Fig. 3 (middle) shows this configuration. Although, the number of corresponding SURF features seems higher, some false positives can be recognized in the case of SURF (e. g., in the reflection on the floor). The SIFT correspondences are also better here. Finally, a scene with a dark corner shows the advantages of SURF over SIFT in some situations. See

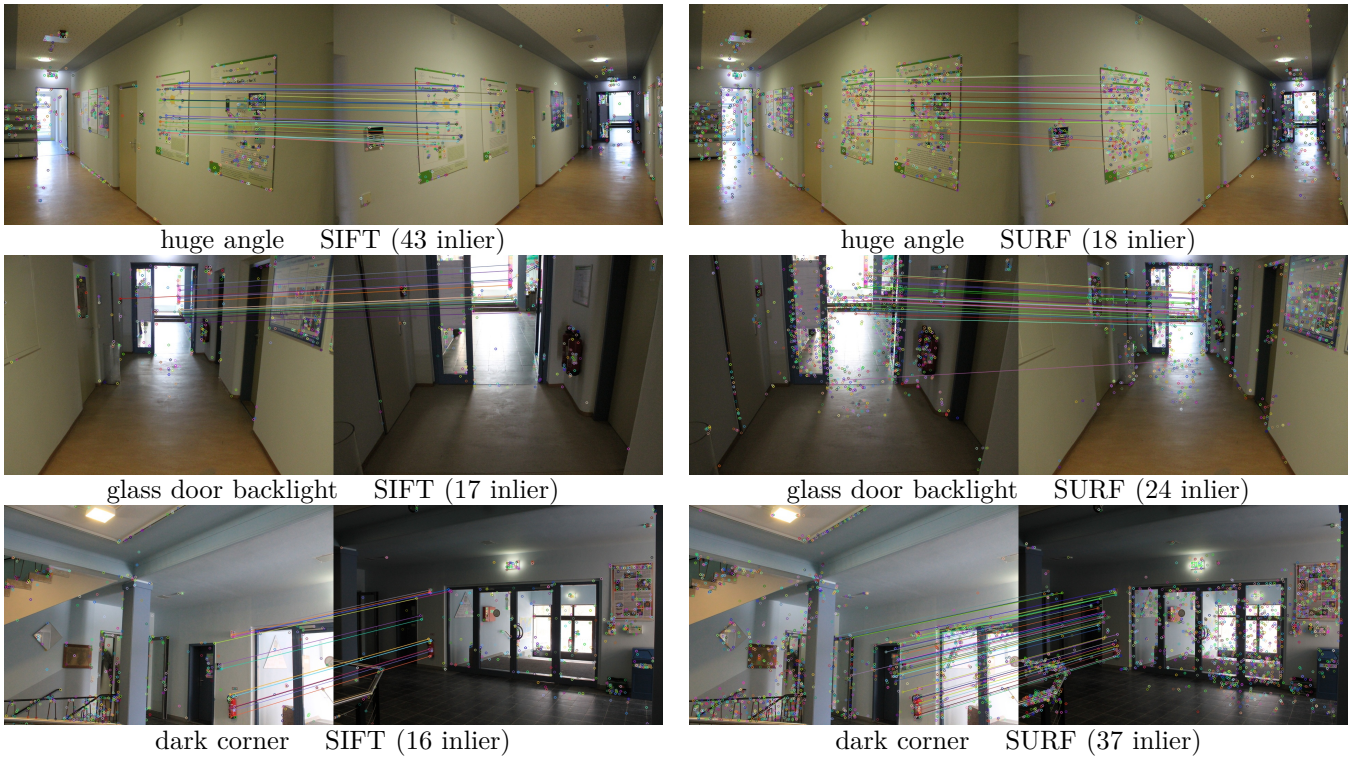

Fig. 3: Comparison of SIFT and SURF matchings in challenging indoor scenarios.

Fig. 3 (bottom) for an illustration. SURF can handle dark scenes often better than SIFT. This is very relevant for indoor environments.

We also tested other combinations of extractor/descriptor types: FAST / FREAK, ORB / ORB, FAST / ORB, and ASIFT / SIFT. These combination either provided much less matchings (e.g. ORB), or took very long to compute with only minor improvements for our test environment (e.g. ASIFT). As a result, both SIFT and SURF proofed to be suitable to precompute matches for an SfM-based 3D reconstruction. To exploit the advantages of both feature detectors, we propose to combine them in order to maximize the quality of the matching in challenging indoor environments.

IV. The combined SIFT/SURF approach

Fig. 2 gives an overview of the approach. The model is created from many overlapping pictures from the scene. The pictures should be taken from poses that are likely to appear during the localization phase, e. g., at the carrying height of the smartphone.

SIFT and SURF keypoints and descriptors are computed from all pictures. Then pairwise SURF matches and SIFT matches are computed separately. Finally, the good matches are combined and outliers are removed as described above. VisualSFM then computes a feature cloud of the scene. The resulting trained model contains the SIFT/SURF descriptor files and the feature cloud. Also, a graph of all adjacent images is created from the matches. Fig. 3 shows examples that hardly match. So, we defined that at least 10 inliers are necessary to extend the graph by an edge between the images.

In the localization phase, SIFT/SURF features are computed from a test image. Considering a stream of images, model descriptors can be reduced in the matching process by only using descriptors that appeared in the highest rated train images from the preceding localization and all of its neighboring images (graph filter). Afterwards, SIFT and SURF matching are performed separately. Both SIFT and SURF results are then combined and the camera resection is solved using a RANSAC approach. The perspective-n-point solver (PNP) computes the position of a camera using a set of 3D points, corresponding image points, and camera intrinsic parameters (500 iterations, reprojection error 2 px, confidence 0.999). The result reflects the best position where a reprojection of the corresponding 3D points to the camera plane is accurate for all image points.

V. Experiments

In this first evaluation, we investigate the robustness of our combined approach against a purely SIFT or SURF approach by localizing a sequence of images taken from a smartphone camera. Our software is implemented

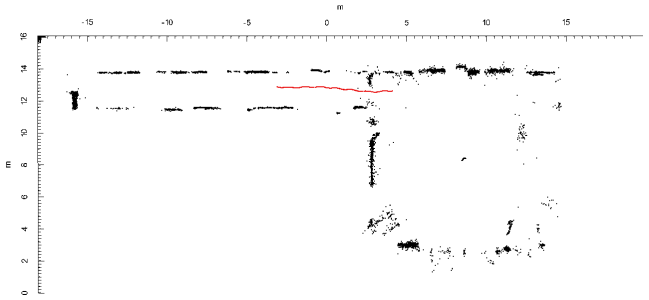

Fig. 4: Top view of the test area (slice of the feature cloud): a foyer with an adjacent office floor. The camera path is shown in red.

in Java using OpenCV (e. g., for SIFT/SURF, PNP, and homography) and JGraphT. We use a challenging area in an office building at the TU Bergakademie Freiberg as a test track. The path crosses a dark scene in an office floor. Fig. 4 shows the top view of the reconstructed area. The camera paths is given in red.

The model was constructed using 447 camera images (1920x1280) taken by a Canon EOS 760D and 18 mm standard lens. Images were not undistorted, but EXIF data such as focal length is used by VisualSFM. Three models could be computed: SURF 59142 3D pts. (291923 descriptors 429 cams used), SIFT 26053 3D pts. (127992 descriptors 394 cams used), COMBINED 65012 3D pts. (363005 descriptors 447 cams used). All models were aligned (scaled, translated, and rotated) using 4 ground control points (GCP) of camera positions.

The test images consist of 301 images (640x480) from an office floor. They were recorded using an Android App on a Sony Xperia Z1 Compact smartphone. The App uses the camera preview mode to realize the recording. The overall length of the camera path in the office floor is approx. 7,65 m and it took around 13 s to record. The speed of movement was approx. 58 cm/s.

We processed the test images one after another on a Core i7-3720QM with 8 GB of RAM and 512 GB SSD drive (Linux Mint 17.3 64-bit, OpenJDK Java 1.7). The intrinsics of the smartphone camera were known.

Table I summarizes the result of the localization tests. It shows the number of recognized images, the average time used for processing per image, and the average number of inlier points in case of a successful localization. We considered an image as localized successfully if the distance between the pose and the ground truth is lower than 1 m and the angle of the rotation vectors in camera coordinate system is lower than 10°. Two configurations of the graph filter have been applied using min. 50 and 100 matches to include neighboring images. Surprisingly, the number of recognized images is higher using SIFT rather than SURF. The COMBINED version can almost recognize all images using any filter.

Table I: Results of localization in an office floor

	office floor (301 images)								
graph filter	SURF			SIFT			COMBINED		
	rec.	*time (ms)*	*in.*	*rec.*	*time (ms)*	*in.*	*rec.*	*time (ms)*	*in.*
none	162	875	60	280	1055	61	299	1874	111
50	206	314	53	275	430	65	296	967	115
100	163	258	63	171	309	84	299	439	120

rec.: successfully recognized images
time: avg. processing time per image
in.: avg. inliers in PNP solver

The number of inliers is significantly higher using the COMBINED approach. Still, the processing speed when comparing the highest recognition ratios (SURF 50, SIFT 50, and COMBINED 100) is competitive for our novel COMBINED approach.

The ground truth camera positions were obtained from an SfM model containing both 447 train images and 301 test images. We created it from a SIFTGPU features and used the VisualSFM tool suite. SIFTGPU computes keypoints/descriptors different from SIFT/SURF. Since this technique is not a real external reference system, the ground truth may include an error of some cm/deg with respect to the real world and can only be used to check the plausibility of our results.

VI. Conclusions

We presented first results on a combined SIFT/-SURF localization approach for indoor environments that outperforms single detector approaches using SIFT or SURF. It only adds a small computation time overhead in the localization phase. We plan to accelerate the localization and to compare it to other works in the field in different case studies. Furthermore, our technique can also be combined with IMU data or an external reference system to improve the localization.

References

[1] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardós, "ORB-SLAM: A Versatile and Accurate Monocular SLAM System," *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, Oct. 2015.

[2] A. Ruiz-Ruiz, P. Lopez-de Teruel, and O. Canovas, "A multisensor LBS using SIFT-based 3d models," in *2012 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, Nov. 2012, pp. 1–10.

[3] M. Ferber and T. Rauber, "MAP: a Cloud-based Middleware for the Provision of fine-grained Compute Services for Mobile Streaming Image Processing Applications," *International Journal of Cloud Computing*, vol. 4, no. 4, pp. 299–316, 2015.

[4] C. Wu, "VisualSFM: A visual structure from motion system," 2011. [Online]. Available: http://ccwu.me/vsfm/doc.html

[5] L. Lenc and P. Král, "A combined SIFT/SURF Descriptor for automatic Face Recognition," in *Proc. of the 6th International Conference on Machine Vision (ICMV 2013)*, 2013.