Distinctive Texture Features from Perspective-Invariant Keypoints

David Gossow, David Weikersdorfer, Michael Beetz
Intelligent Autonomous Systems Group, Technische Universität München.
[gossow—weikersd—beetz]@cs.tum.edu

Abstract

In this paper, we present an algorithm to detect and describe features of surface textures, similar to SIFT and SURF. In contrast to approaches solely based on the intensity image, it uses depth information to achieve invariance with respect to arbitrary changes of the camera pose.

The algorithm works by constructing a scale space representation of the image which conserves the real-world size and shape of texture features. In this representation, keypoints are detected using a Difference-of-Gaussian response. Normal-aligned texture descriptors are then computed from the intensity gradient, normalizing the rotation around the normal using a gradient histogram.

We evaluate our approach on a dataset of planar textured scenes and show that it outperforms SIFT and SURF under large viewpoint changes.

1. Introduction

The detection of local image features that are invariant to changes in scale and rotation, such as SIFT [3] and SURF [1], have become a standard approach in many computer vision applications. However, these algorithms are not invariant with respect to large changes of camera viewpoint, which introduce more complex distortions of the image. Affine invariant detectors as the one described in [5] estimate a local invariant affine transformation from the intensity image, however they require an iterative normalization step that introduces additional instability and lowers distinctiveness.

The availability of cameras that provide accurate depth information for most pixels makes it possible to circumvent this problem, using the surface normals to locally undistort the image. Recently, several such methods have been proposed. In [2], regions are detected using a purely image-based approach, and then rendered from a normal-aligned viewpoint for the descriptor computation. In [6], the scene is first decomposed into planar regions. Then, SIFT is applied to each perspectively normalized plane. The approach described in [7] incorporates the full 3D information directly into the detection and description steps, however it requires a uniformly sampled triangular mesh of the scene.

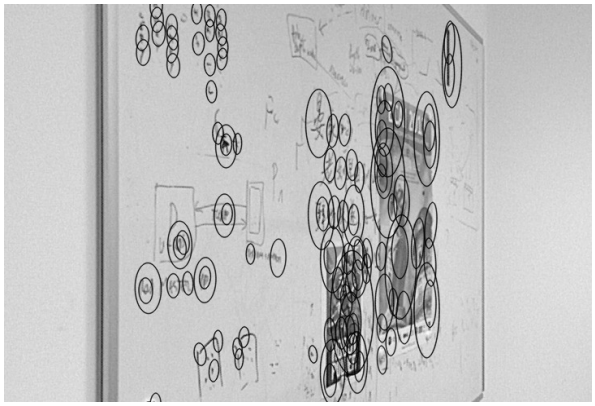

Figure 1. Keypoints detected using our approach

In the following sections, we describe an algorithm that achieves invariance to arbitrary changes in camera pose, working directly on a intensity and depth image without further assumptions about the scene geometry. In section 4, we evaluate the approach against SIFT and SURF and show that it exhibits the desired properties.

2 Feature Detection

For the detection of texture features, we adapt the detection scheme of SIFT [3]. However, in order to remove the need to search and interpolate maxima in the scale dimension, we convolve the image with a Gaussian of a fixed real-world size at each scale level. Using this paradigm, it is sufficient to double the size of the Gaussian between two scale levels.

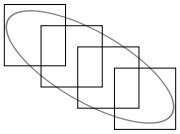

Figure 2. Gaussian approximation using box-shaped samples

For a given image of size $w \times h$, the range of world scales is determined from the depth image in order to yield pixel scales in the range of $[3, min(w,h)/20]$. However, depending on the application domain, this choice can be replaced by a fixed range.

The local projection of the Gaussian into image space is approximated using an affine transformation and the convolution computed using the Feline algorithm [4]. We compute the affine parameters at each pixel from the depth gradient and take a number of equally-weighted box-shaped samples along the major axis using an integral image [1], as shown in figure 2.

In order to compute the local affine parameters, we assume a pinhole camera model with focal length f and and principal point $\vec{c}$ such that a point $\vec{p} = (x, y, z)$ is projected to pixel coordinates (u, v) by

$$(u, v) = \frac{f}{z} \cdot (x, y) + \vec{c} \tag{1}$$

The orthogonal basis of the local affine transformation at world scale s is then given by its major axis $\vec{a}$, and minor axis $\vec{b}$, where $\vec{b}$ is parallel to the depth gradient ∇z:

$$|\vec{a}| = \frac{f}{z} \cdot s, \;\; |\vec{b}| = \frac{|\vec{a}|}{\sqrt{\left\|\nabla z\right\|^2 + 1}} \tag{2}$$

In order to achieve similar invariance properties with respect to the depth gradient ∇z, the depth map is smoothed using $|\vec{b}| = |\vec{a}|$.

Each pixel that is a local extremum in the elliptical neighbourhood of size $2s$ in the difference $\mathrm{D}(s)$ of two consecutive layers generates a keypoint. In addition to a threshold on the value of D, we suppress unstable responses along intensity edges using the princial curvature ratio[3]. Figure 1 shows an example of keypoints that were detected using this algorithm.

3 Descriptor Computation

We use the scheme of SURF [1] to compute keypoint descriptors. In order to do that, we first need to evenly sample intensity gradient values in the neighbourhood of each keypoint $k = (\vec{p}, s)$.

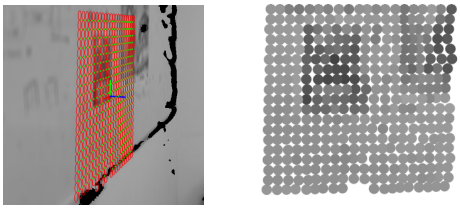

Figure 3. Sample locations and the corresponding points in tangential coordinates used for the descriptor computation

Assuming local planarity, we project evenly spaced locations from the tangent plane into image space and compute the intensity gradients between neighbouring samples. For weighting samples and distributing them into the descriptor bins, we then transform their 3D coordinates into the local tangent space $\mathbb{T}$.

We compute $\mathbb{T}$ by sampling depth values from a 3×3 grid of side length $4s$ using the previously computed local affine approximation. From the corresponding 3D points, the tangential plane and normal axis are computed using principal component analysis.

To normalize the rotation of the tangent plane, we sample a circular neighbourhood with radius $5s$ and step size $0.5s$ in the Gaussian filtered intensity image $\mathrm{I}(2s)$ and compute the discrete intensity gradients. All gradients are then weighted with a 3D Gaussian centered at the keypoint location and $\sigma = \frac{5}{3}s$. When computing the distance to the keypoint center, the position along the normal is scaled by a factor of 5, decreasing the weight of samples that lie outside of the tangent plane. The dominant orientation is computed using an orientation histogram as described in [1].

The sampling and gradient computation for the descriptor is done in the same way (fig. 3), this time considering a window of side length $20s$ in $\mathrm{I}(s)$ using s as step size. The gradients are weighted with a Gaussian with $\sigma = \frac{20}{3}s$ and distributed into 4×4 patches as in [1]. In order to avoid border effects, each sample is distributed into the four neighbouring bins using bilinear interpolation. Each bin is then separately normalized by the sum of bilinear weights of gradients that contributed to it, which increases robustness towards missing depth information. If one of the bins does not contain any information, the keypoint is omitted.

When matching the feature vectors, the search space can be significantly reduced in comparison to purely image-based descriptors by taking into account the real-world scale.

4 Evaluation

We use the same evaluation framework as in [5] to compare our approach (dubbed Depth-Adaptive Feature Transform, or DAFT) to SIFT [3] and SURF [1]. We created multiple datasets using the Microsoft Kinect sensor at an image resolution of 1280×960, with the depth image upscaled to the same resolution. Each one shows the same planar scene under a change of the camera's perspective, such as the angle between camera axis and plane normal (*viewpoint*), the rotation around the plane normal (*rotation*) or the distance of the camera to the scene (*scaling*). In order to test with partially planar scenes, we implemented a step in the repeatability evaluation to omit points outside of the planar image area using a manually created mask. During matching, the nearest neighbour distance ratio is used as the rejection criterion.

We evaluate the detector by means of repeatability and the descriptor by means of precision/recall (for a definition, see [5]). In addition, we compute the maximal F_1 score from each precision/recall graph in order to show the decay of descriptor performance with respect to the strength of an image transformation, where

$$F_1 := 2 \cdot \text{precision} \cdot \text{recall} \div (\text{precision} + \text{recall})$$

Since the evaluated algorithms detect a similar type of feature, we adjusted the thresholds of the detectors to yield between 300 and 500 features for the first image in each data set, depending on the scene. We adjusted the size of the region assigned to a keypoint with respect to its scale to be equal for all algorithms using artificial images.

The average run times on a 2.7 GHz CPU are 0.1s for SURF and 0.6s for SIFT. The run time of DAFT varies between 1.2s and 2.2s, since the number of probes necessary for the Gaussian approximation depends on the surface tilt.

Figure 5 shows the evaluation results. As to be expected, results for the different approaches are similar in case of a rotation around the camera axis and scaling, while the noisiness of the depth information negatively affects repeatability of DAFT when the camera moves far away from the scene. In the case of large viewpoint changes, the additional invariance properties of DAFT strongly improve its repeatability in comparison to SIFT and SURF.

Note that, since SIFT and SURF generate circular image regions and the evaluation computes corresponding regions by how much the projected regions overlap, SIFT and SURF can not generate correspondences at a viewpoint angle of 60° and a rotation of more than 45° around the surface normal. We found that one reason for the repeatability going down even for DAFT is that some features become too small to be detected, and that for some features that become larger in the image, the corresponding descriptor window exceeds the image boundaries or contains too much missing depth information.

While the test set is relatively small, it demonstrates that the algorithm's principle works as expected, increasing detection and matching performance under large changes of camera perspective.

5 Conclusion

We have presented a novel algorithm that extracts viewpoint-invariant texture features from an image by incorporating depth information into both it's detection and description step. In our evaluation, we have demonstrated that it clearly outperforms SIFT and SURF under large viewpoint changes.

While we extend the detection scheme of SIFT and the description step of SURF to 3D surface textures, the described framework can be applied to achieve the same invariance properties for other feature extraction schemes.

The source code and data sets used in this paper are available at *http://ias.in.tum.de/people/gossow/rgbd*.

References

[1] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool. Speeded-up robust features (surf). *Computer Vision and Image Understanding*, 2008.

[2] K. Koser and R. Koch. Perspectively Invariant Normal Features. *International Conference on Computer Vision (ICCV)*, 2007.

[3] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2004.

[4] J. Mccormack, R. Perry, K. I. Farkas, and N. P. Jouppi. Feline : Fast Elliptical Lines for Anisotropic Texture Mapping. In *SIGGRAPH*, 1999.

[5] K. Mikolajczyk and C. Schmid. An affine invariant interest point detector. In *European Conference on Computer Vision (ECCV)*, 2002.

[6] C. Wu, B. Clipp, X. Li, J.-M. Frahm, and M. Pollefeys. 3D model matching with Viewpoint-Invariant Patches (VIP). *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, June 2008.

[7] A. Zaharescu, E. Boyer, K. Varanasi, and R. Horaud. Surface feature detection and description with applications to mesh matching. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 373–380. Ieee, June 2009.

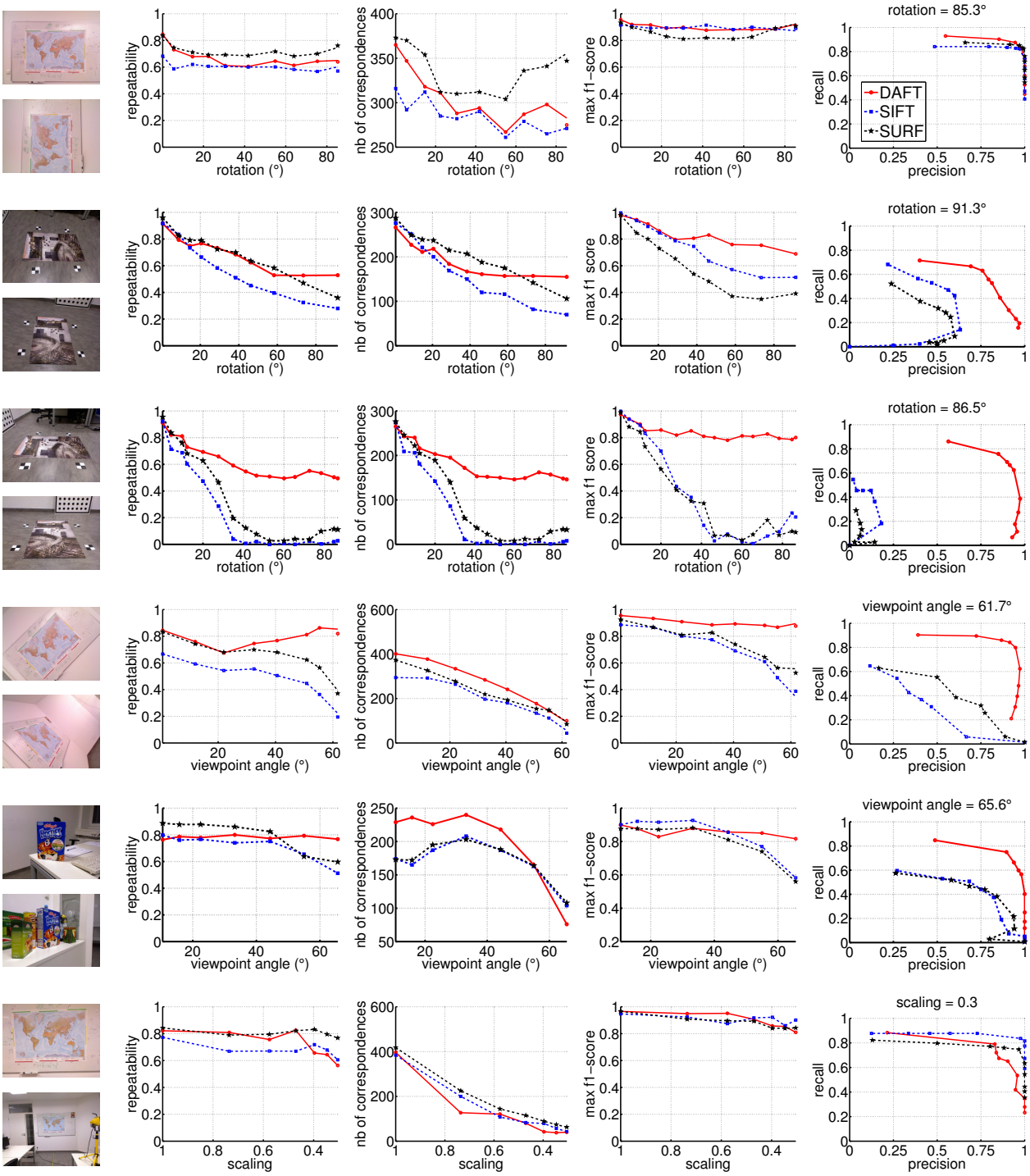

Figure 4. Evaluation results for a set of textured scenes. Each row contains from left to right: First and last image of the sequence, repeatability and number of correspondences (feature detection), maximal f1 scores for all images and precision/recall for the last image in the sequence (descriptor matching). The first 3 datasets contain a rotation around the surface normal at a viewpoint angle of 0°, 40°and 60°, followed by two datasets with increasing viewpoint angle and one with the camera moving away from the planar surface.