

# SIFT VS. SOFT - A COMPARISON OF FEATURE AND CORRELATION BASED ROTATION ESTIMATION FOR PANORAMIC IMAGES

Timo Schairer, Sebastian Herholz, Benjamin Huhle, Wolfgang Straßer

University of Tübingen, WSI/GRIS

## ABSTRACT

Orientation estimation based on image data is a key technique in many applications. Robust estimates are possible in case of omnidirectional images due to the large field of view of the camera. Traditionally, techniques based on local image features have been applied to this kind of problem. Another very efficient technique is to formulate the problem in terms of correlation on the sphere and to solve it in Fourier space. While both methods claim to provide accurate and robust estimates, a quantitative comparison has not been reported yet. In this paper we evaluate the two approaches in terms of accuracy, image resolution and robustness to noise by comparing the estimated rotations of virtual as well as real images to ground-truth data.

**Index Terms** — image registration, omnidirectional vision, orientation estimation, scale invariant features, spherical fourier transform

## 1. INTRODUCTION

The problem of estimating the ego-motion of a camera based on image data has been studied extensively over the last years. Typically, these algorithms have been developed for conventional perspective images and were later adapted to different image modalities, such as panoramic images. In various application scenarios, the use of omnidirectional images is increasingly popular. They are used, for example, in mobile robotics, scene acquisition as well as for surveillance systems since the larger information content inherently present in images with a wide field of view obviously bears important advantages and panoramic image data may even be considered the optimal modality for the recovery of ego-motion [1].

Omnidirectional vision plays an important role in the field of 3DTV: Spherical image data is widely used in applications for background model acquisition (*e.g.* for setups like [2]) and in different semi-3D applications like image based rendering or realistic image based lighting as well as in mixed-reality applications.

Typically, local salient features are detected in the two images. Examining the correspondences allows for an estimation of quite large camera motions. For an application on panoramic images see, *e.g.*, [3]. On a global scale the optical flow can be computed to extract the motion parameters if the motions are sufficiently small (differential motions), see *e.g.* [4].

An orthogonal approach builds on the fact that panoramic images can easily be mapped onto the unit sphere, which in turn allows for the use of spherical signal analysis. A method of rotation estimation directly from images defined on the sphere was presented by Kostelec and Rockmore [5]. Their approach is related to the method of estimating the relative translational movement

between two planar images by the use of phase correlation in the Fourier domain (see *e.g.* [6]). Since these techniques do not rely on correspondences of local image features they are expected to be less affected by small changes in dynamic environments and in contrast to optical-flow methods, they should perform well even if the images change significantly (*i.e.*, in case of large movements).

In this work we focus on the problem of estimating the rotation  $R$  that separates two panoramic images  $I_{\text{ref}}$  and  $I_{\text{rot}}$ , corresponding to the relative rotation of the camera. Without loss of generality, we specifically try to recover the rotation that transforms  $I_{\text{ref}}$  to  $I_{\text{rot}}$

$$I_{\text{rot}} = \Lambda(R) I_{\text{ref}}, \quad (1)$$

with rotation operator  $\Lambda$ . The most likely rotation  $\hat{R} \approx R$  can be found using either of the approaches mentioned above, but unfortunately there are no comparative measurements available that give insight on the performance that is to be expected. Therefore, we explicitly compare a method that examines correspondences between image features to a global technique based on correlation. We provide qualitative and quantitative results using real as well as virtual image data.

In the following, we review these two alternative approaches to rotation estimation. In Section 2.1 we briefly describe a standard algorithm based on correspondences between image features and discuss the correlation based rotation estimation in Section 2.2. We compare the two methods in terms of speed, accuracy and robustness to image noise and resolution. Using rendered data of a photorealistic 3D model as well as real-world data recorded with a spherical camera system mounted onto a highly accurate pan-tilt unit we validate both approaches. Results of our experiments and a discussion of the characteristics and applicability of both methods is presented in Section 3.

## 2. ROTATION ESTIMATION

A widely used representation of omnidirectional images is the so-called “latitude/longitude” or *quirectangular* projection. This projection leads to distortions of sizes and shapes, increasing from the equator to the poles. Compared to other projections, *e.g.* the dome projection, it is still a suitable representation for finding corresponding features in spherical images. We use this projection as input to the algorithm described in Section 2.1. In Figure 1 the effect of rotating a panoramic image can be seen on real and virtual data.

### 2.1. Feature based rotation estimation

In the feature based approach correspondences between image features, extracted from  $I_{\text{ref}}$  and  $I_{\text{rot}}$ , are used to calculate the rotation separating the two images. We extract scale, rotation and illumination invariant features in  $I_{\text{ref}}$  and  $I_{\text{rot}}$  using the *SIFT* feature

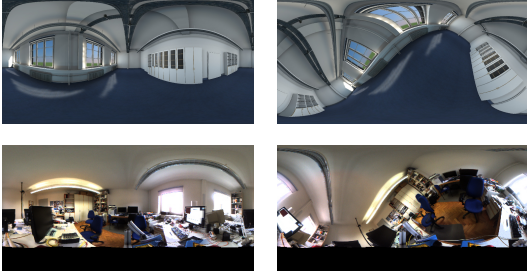


Figure 1. Examples of the effect of rotating an omnidirectional camera in a real and virtual scene. Default orientation (left) and rotated around horizontal and vertical axis (right). Note the black areas in the lower images due to the limited field of view of the camera.

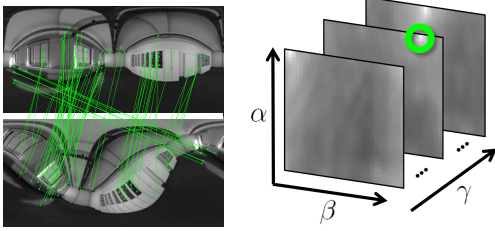


Figure 2. Illustration of SIFT feature points and correspondences (left). Grid of Euler angles  $G$  with highlighted maximum value (right).

detector [7]. As can be seen in Figure 1 the rotation of an omnidirectional camera not only causes variations in position, scale and orientation of the features, but also locally leads to affine transformations, that are not handled explicitly by the *SIFT* feature detector. Note, that small affine transformations could be handled by applying a modified variant of *SIFT*, namely *ASIFT* ([8]), for feature detection. This method, however, performed poorly in terms of speed and robustness compared to the purely *SIFT* based estimation and was therefore discarded from our experiments section.

In a first step, the feature detector is used to extract a set of feature points  $F_{\text{ref}} = \{v_1, \dots, v_n\}$  in the reference image  $I_{\text{ref}}$  and  $F_{\text{rot}} = \{w_1, \dots, w_m\}$  in the rotated image  $I_{\text{rot}}$ , respectively. Then, the feature points in both sets are matched to corresponding points in the other set according to the minimal Euclidian distance of the respective descriptors, resulting in two sets of corresponding pairs  $C_{\text{ref}}$  and  $C_{\text{rot}}$ . To reduce the possibility of mismatches of feature points, the final set of correspondences is computed as the intersection  $C = C_{\text{ref}} \cap C_{\text{rot}}$ .

Then, the feature points are converted from pixel positions via spherical coordinates to unit vectors in Cartesian space. Hence, all vectors lie on the unit sphere and corresponding feature pairs are separated only by a rotation

$$w_j = Rv_i, \quad (2)$$

with a  $3 \times 3$  rotation matrix  $R$ . In theory, only two corresponding feature pairs are needed for the calculation of  $R$ . In practice, however, the images are corrupted by noise and sampling artifacts. Furthermore, they suffer from various other defects due to the applied stitching and/or warping necessary to acquire panoramic image data. This can lead to mismatched pairs of features that severely impact the rotation estimation. It is therefore inevitable to prune outliers and to calculate the best estimate  $\hat{R}$  for  $R$  with the remaining feature pairs. This filtering is performed by applying the *random sample consensus* paradigm (RANSAC, [9]) to rotation estimation:

In each iteration of the algorithm three random feature pairs from the set  $C$  are selected and an estimate  $\tilde{R}$  of  $R$  is calculated. Note, that  $\tilde{R}$  has to be orthogonalized to assure that it meets the criteria of a rotation matrix. Using *Singular Value Decomposition* (SVD)  $\tilde{R}$  can be decomposed into

$$\tilde{R} = U S V^T. \quad (3)$$

The closest orthogonal matrix is then derived replacing  $S$  by the identity matrix  $I$ ,

$$\tilde{R}' = U I V^T. \quad (4)$$

This rotation,  $\tilde{R}'$ , is applied to each feature pair  $c_k \in C$  and  $c_k$  is added to the set of candidates  $\tilde{C}$  if the angular error between  $w_j$  and  $\tilde{R}'v_i$  is smaller than a given threshold. If the size of the set  $\tilde{C}$  is larger than the best candidate set  $\hat{C}$ , then  $\hat{C}$  is replaced by  $\tilde{C}$ . This is done iteratively until a maximum number of iterations is reached or until  $\hat{C}$  contains a certain percentage of all feature pairs.

The set of maximum cardinality  $\hat{C}$  is now used to calculate the final estimation of the rotation matrix  $\hat{R}$  by least squares fitting followed by the orthogonalization procedure described above. An illustration of the detected feature correspondences can be seen in Figure 2.

## 2.2. Correlation based rotation estimation

An omnidirectional image can be considered as a function  $f(\theta, \phi) = f(\omega)$  on the 2-sphere, where  $\theta \in [0, \pi]$  denotes the colatitude and  $\phi \in [0, 2\pi)$  denotes the azimuth. Driscoll and Healy [10] showed that the spherical harmonic functions  $Y_l^m$  form a complete orthonormal basis over the unit sphere and that any square-integrable function  $f \in L^2(S^2)$  can be expanded as a linear combination of spherical harmonic functions (Spherical Fourier Transform, *SFT*)

$$f(\omega) = \sum_{l \in \mathbb{N}} \sum_{m \in \mathbb{Z}, |m| \leq l} \hat{f}_l^m Y_l^m(\omega), \quad (5)$$

where  $\hat{f}_l^m \in \mathbb{C}$  are the complex expansion coefficients. The spherical harmonic function  $Y_l^m$  of degree  $l$  and order  $m$  is given by

$$Y_l^m(\theta, \phi) = \sqrt{\frac{(2l+1)(l-m)!}{4\pi(l+m)!}} P_l^m(\cos \theta) \exp(im\phi), \quad (6)$$

with  $P_l^m$  denoting the associated Legendre polynomials. Our input data, the spherical functions  $f$ , are defined on a uniformly sampled equiangular grid. A perfect reconstruction from a  $2B \times 2B$  grid is possible when bandlimiting  $f$  to  $B$ .

Similar to the phase correlation method on planar images, Kostelec and Rockmore [5] present a fast method to estimate the alignment of images defined on the sphere using cross-correlation as similarity measure. They showed that the correlation between two omnidirectional images  $I_{\text{ref}}$  and  $I_{\text{rot}}$  as a function

$$C(R) = \int_{S^2} I_{\text{ref}}(\omega) \Lambda(R) I_{\text{rot}}(\omega) d\omega \quad (7)$$

of rotations can efficiently be evaluated in the Fourier domain. Here,  $\Lambda$  denotes the rotation operator corresponding to the rotation  $R = R(\alpha, \beta, \gamma)$  where  $\alpha, \beta, \gamma$  are the Euler angles (in *ZYX* representation) defining the rotation. Further, the spherical harmonic functions  $Y_l^m$  form an orthonormal basis for the representations of  $SO(3)$  and the  $SO(3)$  Fourier transform (*SOFT*) coefficients

of the correlation of two spherical functions can be obtained directly by calculating the bandwise outer product (denoted by  $\diamond$ ) of their individual *SFT* coefficients. Taking the inverse *SOFT*

$$C(R) = SOFT^{-1} \left( \hat{I}_{\text{ref}} \diamond (\hat{I}_{\text{rot}})^* \right), \quad (8)$$

where  $(\hat{I}_{\text{rot}})^*$  denotes the complex conjugate of  $\hat{I}_{\text{rot}}$ , yields the correlation  $C(R)$  evaluated on the  $2B \times 2B \times 2B$  grid of Euler angles  $G$  and its maximum value ideally indicates the rotation separating the two images. The accuracy of the rotation estimate  $\hat{R} = \arg \max_{(\alpha, \beta, \gamma) \in G} C(R(\alpha, \beta, \gamma))$  is directly related to the resolution of the likelihood grid which in turn is specified by the number of bands used in the *SFT*. Given images of bandwidth  $B$ , the resolution of the likelihood grid implicates an inaccuracy of up to  $\pm(\frac{180}{2B})^\circ$  in  $\alpha$  and  $\gamma$ , and  $\pm(\frac{90}{2B})^\circ$  in  $\beta$ . The cubic computational cost when evaluating the grid, in practice, restricts this method to bandwidths up to  $B = 256$ . An illustration of the grid is depicted in Figure 2.

When acquiring omnidirectional images, typically, the sensors do not cover the whole sphere and the images have limited support. Huhle et. al [11] show that the spatially normalized cross-correlation (NCC) of two spherical images can be expanded in terms of simple correlations and therefore can be computed with multiple applications of the inverse *SOFT* transform. In the remainder of the paper we use this function as a similarity measure when estimating the orientation.

### 3. EXPERIMENTS AND RESULTS

The evaluation of the two methods was done using Matlab along with MEX-files as interfaces to C-routines. The feature based approach makes use of the *SIFT for Matlab* implementation by Vedaldi [12]. For the correlation based approach we apply the *SFT* and *SOFT* transforms provided by the *S2Kit* [13] and *SOFT library* [5]. Timings were recorded on a standard quad core machine using grayscale images. Note, that no further processing has been performed on the rotation estimates. To account for the temporal coherence and to compensate outliers and the quantization effect of the *SOFT* based orientation estimates, respectively, a particle filter could be used, e.g., as described in [14].

We compared the feature based method to the correlation based technique using sets of virtual, as well as real-world images with known camera rotations. Using a CAD rendering system based on global illumination, four synthetic sets of photorealistic images were generated. Each set corresponds to rotations of  $-90^\circ$  to  $+90^\circ$ , with increments of  $5^\circ$  per frame, around the rotation vector  $r$ , that we set to  $(1, 0, 0)^T$ ,  $(0, 1, 0)^T$ ,  $(0, 0, 1)^T$  and  $(1, 0, 1)^T$ , respectively. Note, that for all experiments, the image center corresponds to the  $Y$  axis and the optical axis of the camera is aligned to the  $Z$  axis. To further examine the robustness of the methods to noise in the input images, white Gaussian noise of different levels was added to the source images (floating point values in  $[0, \dots, 1]$ ) after resizing. The real-world images were acquired using a *LadyBug2* camera system that we mounted on a highly accurate pan-tilt unit to control the camera rotation. The camera system consists of six wide angle cameras covering about  $360^\circ \times 130^\circ$  of the entire sphere and the stitching of the single XGA-frames is computed in hardware in real-time. Note, that in this experiment, the camera origin does not coincide with the rotation axis. Therefore, an additional translational movement (up to

ca. 20cm and ca. 50cm in the first and second experiment, respectively) is induced which however does not affect the comparison. Contrarily, a correct rotation estimate approves a certain robustness to translational movements. Due to mechanical limitations rotations around the  $X$  axis in the real scene #2 could only be recorded from  $-65^\circ$  to  $65^\circ$ .

We evaluated both methods in terms of noise resistance, input image resolution and runtime requirements. These criteria can be considered the boundary conditions for a concrete application and our results presented in Table 1 provide an indication for the estimation performance that can be expected. Note, that we did not include the results for rotations around the  $X$  axis since the numbers were very similar to the ones obtained by rotating around the  $Y$  axis. Examining the results of virtual scene #1 and #2 shows that on noise-free data, both methods perform equally well, with *SIFT* outperforming *SOFT* on higher resolutions and vice versa.

Increasing the amount of noise, the accuracy of the *SIFT* based approach degrades very slowly for higher resolutions but quite fast for smaller input images, whereas the *SOFT* based method proved to be very robust to noise even in combination with very low resolution input data.

When using the equirectangular projection, rotations around the optical axis of the camera, as in the virtual scene #3 and real scene #1, do not distort the image but only cause a circular shift in image content promoting the feature based technique in such situations. Here, the quantization effects are very apparent if the *SOFT* based approach is used.

Real scene #2 is challenging for both approaches, since due to the limited support of the camera images (Fig. 1) a great amount of new image content is introduced.

At first glance the *SIFT* based approach seems superior in terms of speed, but looking at specific achievable accuracies both methods provide comparable runtimes, especially on noisy data (e.g. comparing the measurements annotated by  $\dagger$  and  $\ddagger$ , respectively). Note, that both methods could be optimized by a parallel formulation of the algorithms running on multiple cores or on the GPU.

To compare the two methods in respect to their performance as a function of the rotation angle, we included two plots in Figure 3 that correspond to virtual scene #1 without adding noise. The approach based on features exhibits a very good performance when estimating the rotation of smaller angles and degrades as the angles become larger; the level of degradation scaling inversely with the size of the input data. The correlation based approach shows quantization artifacts that are inherently given by the grid of Euler angles, the level of accuracy, however, is not affected by the magnitude of the rotation.

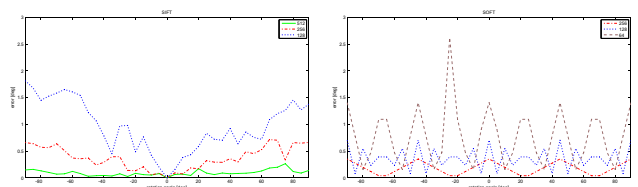


Figure 3. Detail plots of the single estimates of virtual scene #1 without additional noise added. The plots show the error in degrees of the *SIFT* based approach (left) and of the method using *SOFT* (right) against the rotation angle.

Scene	Noise	Resolution							
	$\sigma$	512 <sup>2</sup>		256 <sup>2</sup>		128 <sup>2</sup>		64 <sup>2</sup>	
		<i>SIFT</i>	<i>SOFT</i>	<i>SIFT</i>	<i>SOFT</i>	<i>SIFT</i>	<i>SOFT</i>	<i>SIFT</i>	<i>SOFT</i>
<b>Virtual Scene 1</b> , $r = (0, 1, 0)^T$	0	0.011	-	0.183 <sup>†</sup>	0.044	1.147	0.178 <sup>‡</sup>	x	0.875
	0.001	0.023	-	0.272	0.044	4.972	0.179	x	1.722
	0.005	0.082	-	0.437	0.044	198.990	0.207	x	2.891
	0.01	0.264	-	2.530	0.044	x	0.328	x	4.227
<b>Virtual Scene 2</b> , $r = (1, 0, 1)^T$	0	0.006	-	0.087	0.381	0.670	1.171	x	5.177
	0.001	0.016	-	0.193	0.381	1.590	1.171	x	5.408
	0.005	0.061	-	0.306	0.384	641.411	1.143	x	5.367
	0.01	0.101	-	0.986 <sup>†</sup>	0.365	342.117	1.205 <sup>‡</sup>	x	6.891
<b>Virtual Scene 3</b> , $r = (0, 0, 1)^T$	0	0.003	-	0.004	0.282	0.013	1.128	0.498	4.511
	0.001	0.007	-	0.039	0.282	0.366	1.128	2.010	4.511
	0.005	0.018	-	0.132	0.282	1.316	1.128	29.876	4.511
	0.01	0.057	-	0.356	0.282	4.514	1.152	378.085	4.511
<b>Real Scene 1</b> , $r = (0, 0, 1)^T$	n/a	3.122	-	2.955 <sup>†</sup>	1.636	4.139	2.104 <sup>‡</sup>	3.908	4.812
<b>Real Scene 2</b> , $r = (1, 0, 0)^T$	n/a	11.961	-	14.112	13.856	1783.504	11.990	1544.654	7.216
<b>Runtime</b> [sec]		2.93	-	0.79	6.74	0.22	0.71	0.06	0.01

Table 1. Mean squared error (MSE [ $deg^2$ ]) of rotation estimates w.r.t. ground-truth (virtual scenes) and encoder readings of the pan-tilt unit (real scenes). Experiments where more than 10% of the estimated rotations could not be estimated due the low number of correspondences were considered as failed and are denoted by “x”. Note, that in case of *SOFT*, input image resolutions of 512<sup>2</sup> or more pixels (corresponding to bandwidths  $> 128$ ) are technically possible on 64-bit machines but the memory consumption is rather prohibitive. The runtime is measured using a non-optimized single-threaded Matlab implementation on a standard quad core machine.

#### 4. CONCLUSION

We presented a comparison of two alternative approaches to rotation estimation on panoramic images. Using virtual as well as real scenes the performance of a feature based approach versus a technique based on correlation was evaluated according to accuracy, image resolution and robustness to noise. The experiments showed, that for a given anticipated accuracy both methods are comparable in terms of runtime. The *SIFT* based rotation estimation leads to superior estimates on larger input images that contain a low amount of noise while the accuracy turned out to be inversely proportional to the magnitude of the rotation. On the contrary, the *SOFT* based approach is very robust to noise and performs well, even on very small images. Additionally, it is not affected by the amount of rotation. Further experiments could be conducted to evaluate the sensitivity to varying lighting conditions and changes in the scene.

#### 5. REFERENCES

- [1] C. Fermüller and Y. Aloimonos, “Ambiguity in structure from motion: Sphere versus plane,” *Int. J. Comput. Vision*, vol. 28, no. 2, pp. 137–154, 1998.
- [2] S. Fleck, F. Busch, P. Biber, and W. Straßer, “Graph cut based panoramic 3d modeling and ground truth comparison with a mobile platform - the wägele,” *Image Vision Comput.*, vol. 27, no. 1-2, pp. 141–152, 2009.
- [3] M. Fiala, “Structure from motion using sift features and the ph transform with panoramic imagery,” in *Proc. Canadian conference on Computer and Robot Vision (CRV)*, 2005.
- [4] J. Lim and N. Barnes, “Directions of egomotion from antipodal points,” in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2008, pp. 1–8.
- [5] P.J. Kostelec and D.N. Rockmore, “Ffts on the rotation group,” Tech. Rep., Fe Institutes Working Paper Series, 2003.
- [6] E. De Castro and C. Morandi, “Registration of translated and rotated images using finite fourier transforms,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 9, no. 5, pp. 700–703, 1987.
- [7] David G. Lowe, “Distinctive Image Features from Scale-Invariant Keypoints,” *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, 2004.
- [8] Jean-Michel Morel and Guoshen Yu, “Asift: A new framework for fully affine invariant image comparison,” *SIAM J. Img. Sci.*, vol. 2, no. 2, pp. 438–469, 2009.
- [9] Martin A. Fischler and Robert C. Bolles, “Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography,” *Readings in Computer Vision: Issues, Problems, Principles and Paradigms*, vol. 24, no. 6, pp. 726–740, 1987.
- [10] J.R. Driscoll and D. M. Healy, Jr., “Computing fourier transforms and convolutions on the 2-sphere,” *Adv. Appl. Math.*, vol. 15, no. 2, pp. 202–250, 1994.
- [11] Benjamin Huhle, Timo Schairer, and Wolfgang Straßer, “Normalized cross-correlation using SOFT,” in *Proc. Int. Workshop on Local and Non-Local Approximation in Image Processing (LNLA)*, 2009.
- [12] A. Vedaldi, “An open implementation of the SIFT detector and descriptor,” Tech. Rep. 070012, UCLA CSD, 2007.
- [13] P.J. Kostelec and D.N. Rockmore, “S2kit: A lite version of spharmonickit,” Tech. Rep., Dartmouth College, 2004.
- [14] Timo Schairer, Benjamin Huhle, and Wolfgang Straßer, “Application of particle filters to vision-based orientation estimation using harmonic analysis,” in *Proc. IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2010.