Geometric and Probabilistic Image Dissimilarity Measures for Common Field of View Detection

Marcel Brückner   Ferid Bajramovic   Joachim Denzler
Chair for Computer Vision, Friedrich Schiller University of Jena
Ernst-Abbe-Platz 2, 07743 Jena, Germany
{marcel.brueckner, ferid.bajramovic, joachim.denzler}@uni-jena.de

Abstract

Detecting image pairs with a common field of view is an important prerequisite for many computer vision tasks. Typically, common local features are used as a criterion for identifying such image pairs. This approach, however, requires a reliable method for matching features, which is generally a very difficult problem – especially in situations with a wide baseline or ambiguities in the scene.

We propose two new approaches for the common field of view problem. The first one is still based on feature matching. Instead of requiring a very low false positive rate for the feature matching, however, geometric constraints are used to assess matches which may contain many false positives. The second approach completely avoids hard matching of features by evaluating the entropy of correspondence probabilities.

We perform quantitative experiments on three different hand-labeled scenes with varying difficulty. In moderately difficult situations with a medium baseline and few ambiguities in the scene, our proposed methods give similarly good results to the classical matching based method. On the most challenging scene having a wide baseline and many ambiguities, the performance of the classical method deteriorates, while ours are much less affected and still produce good results. Hence, our methods show the best overall performance in a combined evaluation.

1. Introduction

For many computer vision tasks, like 3D reconstruction [5, 8, 11], multi camera calibration [6, 1], or multi camera object tracking [10], it is important to know which pairs of a given set of images have a common field of view. This is especially important in case of multi camera systems including pan-tilt units and mobile cameras. If knowledge about common fields of view is not available from other sources, like the temporal order of the images in a video or user input, it has to be gained from the images themselves. In this paper, we present two new approaches to this problem and compare them to an existing one.

Note that our methods are not intended for the related problem of image retrieval, which poses a "one of *n*" problem with large *n*, whereas common field of view detection poses an "*m* vs. *k*" problem with moderate *m* and *k*.

Typically, local feature correspondences are used to identify images with a common field of view [11, 2]. Snavely et al. [11] use SIFT features [7] with strict rejection in case of ambiguities during the matching process. Their approach works well if image pairs with a common field of view share highly distinctive features. However, if features are less distinctive, e.g. due to a wide baseline or ambiguities within the images, the strict rejection may miss all correspondences.

In order to handle such difficult situations, we propose using a less restrictive matching method and applying geometric constraints to assess the resulting point correspondences. In particular, we suggest using the uncertainty measures presented by Bajramovic and Denzler [1], which were originally used to evaluate relative pose estimates. This approach is roughly similar to the rejection of image pairs employed by Martinec and Pajdla [8] in the context of multi camera 3D reconstruction, but uses a probabilistic model instead of an inlier threshold.

The least restrictive matching method possible consists of avoiding matching altogether. In case of ambiguities, we can retain the whole information without risking incorrect matches. Domke and Aloimonos [3] suggest estimating relative poses from point correspondence probabilities based on Gabor filters. We adopt the idea of expressing possible correspondences in probability distributions. Instead of Gabor filters, however, we use SIFT features to gain rotation an scale invariance [7, 9]. Furthermore, we present two alternative correspondence probability models.

Our hypothesis is that the correspondence probability distributions are usually peaked in case of image pairs with a common field of view, but tend towards uniform distri-

978-1-4244-3991-1/09/$25.00 ©2009 IEEE

butions for unrelated images. Hence, we suggest using the joint entropy of the distributions as a measure for images with a common field of view.

The remainder of this paper is structured as follows: In section 2, we summarize Snavely's method. We present the geometric measures in section 3 and the probabilistic ones in section 4. Our experimental analysis is provided in section 5. Conclusions are given in section 6.

2. Snavely's Criterion

The criterion of Snavely et al. [11] is based on counting point correspondences. It is motivated by the fact that correct correspondences can only exist between images with a common field of view. In order for their approach to work, they require a reliable correspondence extraction method with a very low false positive rate.

The usual way of extracting point correspondences from an image pair consists of three steps: detecting interest points $C = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ and $C' = \{\boldsymbol{x}'_1, \ldots, \boldsymbol{x}'_{n'}\}$ in both images, computing a descriptor $\mathbf{des}(\boldsymbol{x}_i)$ for each of these points, and matching the points based on a distance measure $\mathrm{dist}(\cdot, \cdot)$ on the descriptors. Snavely et al. [11] use the difference of Gaussian (DoG) detector, the SIFT descriptor, the Euclidean distance, and the two nearest neighbors matching with rejection as proposed by Lowe [7].

As false positive matches typically result from ambiguities, they modify the matching process in order to achieve a very low false positive rate as follows: They use a stricter threshold than Lowe for the two nearest neighbors rejection. Additionally, they remove from these correspondences all point pairs containing multiply matched points.

As there might still be a few – necessarily incorrect – correspondences between images *without* a common field of view, a certain minimum number of correspondences is required to classify an image pair as having a common field of view. In the experiments, we vary this threshold in a Receiver Operator Characteristic (ROC) plot.

3. Geometric Dissimilarity Measures

In case of a wide baseline setup or ambiguities within the scene, the rejection of ambiguities used for Snavely's criterion often returns only very few or no correspondences at all—despite a common field of view. In order to handle such difficult situations, a less strict rejection of ambiguities is required in the matching step. As this leads to more correspondences also in case of images *without* a common field of view, simply counting correspondences is not very promising. However, the correspondences between images *without* a common field of view can be expected to be very unstructured. Hence, we suggest using a geometric measure based on the epipolar constraint to assess the set of correspondences between two images. For this approach, we have to assume that there is a translation between the two cameras (opposed to pure rotation about the optical center).

Bajramovic and Denzler [1] propose three uncertainty measures for relative pose estimates based on a global evaluation of a probability density function introduced by Engels and Nistér [4]. We suggest using these measures to evaluate the quality of point correspondences. As argued above, the quality of correspondences between images with a common field of view can be expected to be higher than between unrelated images.

3.1. Less Restrictive Matching

A less strict variant of the matching procedure described in section 2 can be achieved by simply choosing a less restrictive two nearest neighbors rejection threshold. Alternatively, the two nearest neighbors rejection criterion can be left out completely. Furthermore, the elimination of multiply matched points can be replaced by the following greedy strategy: Let $\mathcal{D}^*$ denote the preliminary set of correspondences resulting from the (two) nearest neighbor(s) matching. The final correspondence set $\mathcal{D}$ is constructed from $\mathcal{D}^*$ by adding one pair after the other with increasing descriptor distances provided that both points in each pair have not yet been added to $\mathcal{D}$. As Bajramovic and Denzler [1], we limit the number of correspondences in $\mathcal{D}$ by choosing the best k point pairs according to their descriptor distances.

3.2. Geometric Measures

Bajramovic and Denzler [1] model the probability density function $p(\boldsymbol{R}, \boldsymbol{t}^* \mid \mathcal{D})$ of a relative pose $\boldsymbol{R}, \boldsymbol{t}^*$ (rotation and translation direction) given point correspondences $\mathcal{D}$ using the Blake-Zisserman distribution [5] and the epipolar constraint. A discretely represented approximate marginalization $p(\boldsymbol{t}^* \mid \mathcal{D})$ of $p(\boldsymbol{R}, \boldsymbol{t}^* \mid \mathcal{D})$ is computed as well as an estimate $\hat{\boldsymbol{R}}, \hat{\boldsymbol{t}}^*$ for the relative pose. For this task, a robust sampling algorithm [4, 1] similar to MLESAC [13] is applied based on the five point algorithm [12].

Bajramovic and Denzler [1] derive three geometric uncertainty measures from the results:

1. the information $-\log p(\hat{\boldsymbol{R}}, \hat{\boldsymbol{t}}^* \mid \mathcal{D})$,
2. the entropy $-\int p(\boldsymbol{t}^* \mid \mathcal{D}) \log p(\boldsymbol{t}^* \mid \mathcal{D}) d\boldsymbol{t}^*$,
3. and the so called "smoothed information" $-\log \int \mathcal{N}(\boldsymbol{t}^*; \hat{\boldsymbol{t}}^*, \boldsymbol{\Sigma}) p(\boldsymbol{t}^* \mid \mathcal{D}) d\boldsymbol{t}^*$, where $\mathcal{N}(\boldsymbol{t}^*; \hat{\boldsymbol{t}}^*, \boldsymbol{\Sigma})$ denotes a Gaussian distribution with mean $\hat{\boldsymbol{t}}^*$ and covariance matrix $\boldsymbol{\Sigma} = \sqrt{5}\boldsymbol{I}$.

4. Probabilistic Dissimilarity Measures

If two points are matched incorrectly, e.g. due to ambiguities in the image, the resulting point correspondence $(\boldsymbol{x}_i, \boldsymbol{x}'_j)$ is useless. Probabilistic point correspondences, as suggested by Domke and Aloimonos [3] for relative pose estimation, on the other hand, provide a way of avoiding hard matching decisions, while retaining all information. Instead of using a set of point pairs, the correspondence information between two images is described by conditional probability distributions $p(\boldsymbol{x}'_j \mid \boldsymbol{x}_i)$. We adopt this idea to the common field of view problem.

Due to their invariance properties [7, 9], we use SIFT descriptors [7] instead of Gabor filters to construct the conditional probability distributions $p(\boldsymbol{x}'_j \mid \boldsymbol{x}_i)$. In each image, we extract interest points using the DoG detector and compute the SIFT descriptor for each of these points. We model $p(\boldsymbol{x}'_j \mid \boldsymbol{x}_i)$ using the exponential distribution and the Euclidean distances $d_{ij} = \mathrm{dist}(\mathbf{des}(\boldsymbol{x}_i), \mathbf{des}(\boldsymbol{x}'_j))$ between pairs of SIFT descriptors:

$$p\left(\boldsymbol{x}'_j \mid \boldsymbol{x}_i\right) \propto \exp\left(-\frac{d_{ij}}{\lambda}\right) \ , \qquad (1)$$

where λ denotes the inverse scale parameter of the exponential distribution.

Inspired by the good performance of the Snavely method (see sections 2 and 5), we construct an alternative distribution by incorporating the distance to the nearest neighbor $d_{\mathrm{N}}(\boldsymbol{x}_i) = \min_j(d_{ij})$ of the point $\boldsymbol{x}_i$:

$$p\left(\boldsymbol{x}'_j \mid \boldsymbol{x}_i\right) \propto \exp\left(-\frac{d_{ij} - d_{\mathrm{N}}(\boldsymbol{x}_i)}{\lambda\, d_{\mathrm{N}}(\boldsymbol{x}_i)}\right) \ . \qquad (2)$$

Note that this variant is similar to the two nearest neighbors rejection method. If the distance to the second nearest neighbor is large compared to the distance to the first nearest neighbor, there will be a single high peak in the distribution.

Finally, each of the resulting conditional probability distributions $p(\boldsymbol{x}'_j \mid \boldsymbol{x}_i)$ has to be normalized such that the following condition holds: $\sum_{j=1}^{n'} p(\boldsymbol{x}'_j \mid \boldsymbol{x}_i) = 1$.

If two images share a common field of view, it is likely that the conditional probability distributions $p(\boldsymbol{x}'_j \mid \boldsymbol{x}_i)$ will have some clear peaks. On the other hand, the conditional probability distributions $p(\boldsymbol{x}'_j \mid \boldsymbol{x}_i)$ of two images showing two different scenes can be expected to tend towards uniform distributions. Hence, we propose using the normalized joint entropy as a measure for common fields of view:

$$H(C, C') = -\frac{1}{\alpha} \sum_{i=1}^{n} \sum_{j=1}^{n'} p(\boldsymbol{x}_i) p\left(\boldsymbol{x}'_j \mid \boldsymbol{x}_i\right) \log\left(p(\boldsymbol{x}_i) p\left(\boldsymbol{x}'_j \mid \boldsymbol{x}_i\right)\right) \ , \qquad (3)$$

where $\alpha = \log(nn')$ and $p(\boldsymbol{x}_i)$ is a uniform distribution if no prior information about the interest points is available. The joint entropy is maximized if all conditional probability distributions $p(\boldsymbol{x}'_j \mid \boldsymbol{x}_i)$ are uniform. In case of the second model in equation (2), it is minimized if every interest point in the first image has a unique corresponding point with an identical descriptor in the second image. In case of equation (1), the minimum is only reached if additionally the non-corresponding descriptors have infinite distance.

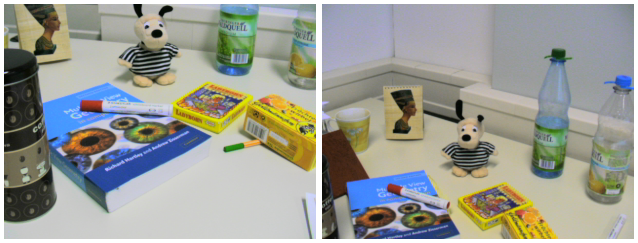

Figure 1. Image number 16 of the panning camera (left) and image number 5 of the tilting camera (right) from sequence `desk1`.

Since every point in an image can only have a single corresponding point in the second image, we enforce $|C| = |C'| = m$ by selecting exactly m points from each of the two point sets C and C'. The procedure for this selection depends on the chosen distribution. In case of the distribution in equation (1), the point pairs are sorted ascendingly by their descriptor distances d_{ij}. According to this order, the m best points of each point set are chosen.

We use a similar selection of m points from each of the two point sets in case of the second distribution in equation (2). However, instead of sorting the point pairs by their descriptor distances, they are sorted descendingly by the conditional probability $p(\boldsymbol{x}'_j \mid \boldsymbol{x}_i)$ using all points of each set. After the selection, the conditional probabilities need to be recomputed using the smaller point sets.

An obvious upper bound for the value m is $\min(|C|, |C'|)$, but smaller values can give better results (see section 5.1).

5. Experiments

In order to evaluate and compare our dissimilarity measures, we use two cameras mounted on top of pan-tilt units. While one camera carries out a pan movement and records 25 images, the other one records 9 images in a tilt sequence. In the experiment `desk1`, the baseline between the cameras is rather small and the scene does not have many ambiguities. An example image pair of this sequence sharing a common field of view is shown in Figure 1. For the experiment `desk2`, we choose a wide baseline and place different ambiguities (identical objects) in the scene. An impression of this sequence is given by the image pair in Figure 2. The task for each of these two sequences consists of finding image pairs between the two cameras with a common field of view. Only about one quarter of the image pairs share a common field of view.

We perform a third experiment called `robo` with a camera mounted onto a robot arm, which we use to record 14 images showing different parts of the surrounding room. Two example images of this sequence are shown in Fig-

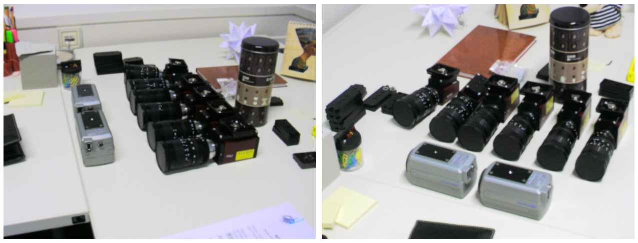

Figure 2. Image number 12 of the panning camera (left) and image number 4 of the tilting camera (right) from sequence `desk2`.

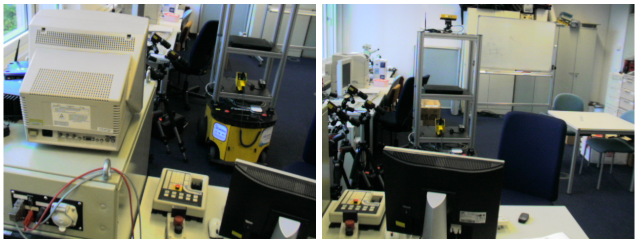

Figure 3. Images number 10 and 12 of the `robo` sequence.

ure 3. As this sequence is recorded using a single camera as opposed to two pan-tilt cameras, the task consists of finding images pairs with a common field of view considering all possible pairs (except for pairs of identical images).

Ground truth is obtained by manually marking polygons of common fields of view in all image pairs. An image pair is considered having a common field of view if the area of the marked polygon in each image is at least ten percent of the total image area.

We compute Receiver Operator Characteristic (ROC) curves for each individual experiment as well as jointly for all three scenes (denoted `all`) using common thresholds. For the sake of clarity, however, we condense most of the ROC curves into a single value by computing the area under the curve. In the first part of the evaluation, we analyse the influence of the parameters m and λ of the probabilistic measures by plotting the ROC areas for varying parameter values. In the second part, we present a comparison of the individual methods using ROC curves as well as a bar chart showing ROC areas. In the plots, we use the following terms to refer to the individual methods:

- `Snavely`: the criterion described in section 2 using a two nearest neighbors rejection threshold of 0.6,
- `geometric information/entropy/smoothed NN/2NN`: respectively, the geometric information, entropy or smoothed information measure (section 3) using nearest neighbor (section 3.1) or two nearest neighbors matching with a rejection threshold of 0.8 (section 2) to extract $k = 50$ correspondences,
- `probabilistic (NN)`: the probabilistic measure (section 4, equation (1)) and its nearest neighbor variant (equation (2)).

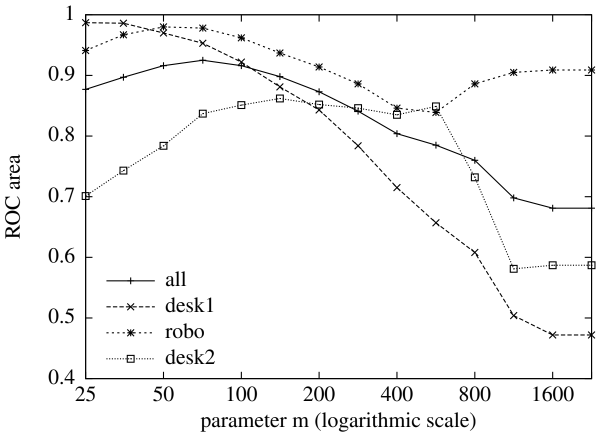

Figure 4. Comparison of the ROC area results for various parameters m of the `probabilistic` method.

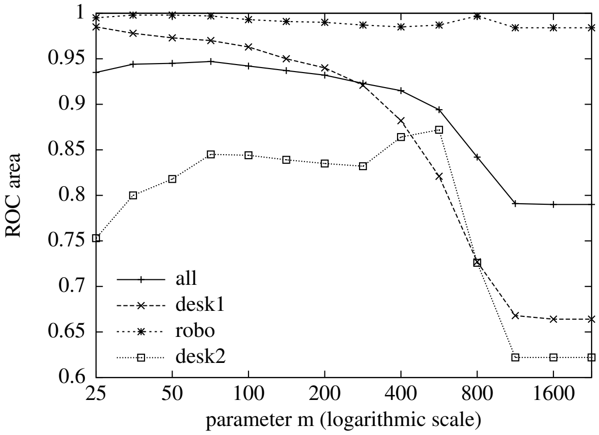

Figure 5. Comparison of the ROC area results for various parameters m of the `probabilistic NN` method.

5.1. Parameters of the Probabilistic Methods

Figures 4 and 5 show the ROC areas for different values of the parameter m (see section 4), starting with $m = 25$ and increasing by a factor of $\sqrt{2}$. In both cases, the best overall result is achieved by a value of $m = 71$, which we use in all other experiments. Interestingly, the best result on the `desk2` sequence using `probabilistic NN` is achieved by setting $m = 500$.

Another parameter we investigate is the inverse scale parameter λ in the exponential distribution in equation (1) and (2) used by the `probabilistic` and the `probabilistic NN` methods. Figures 6 and 7 show the resulting ROC areas using various values for λ. In case of `probabilistic`, the best results are achieved using

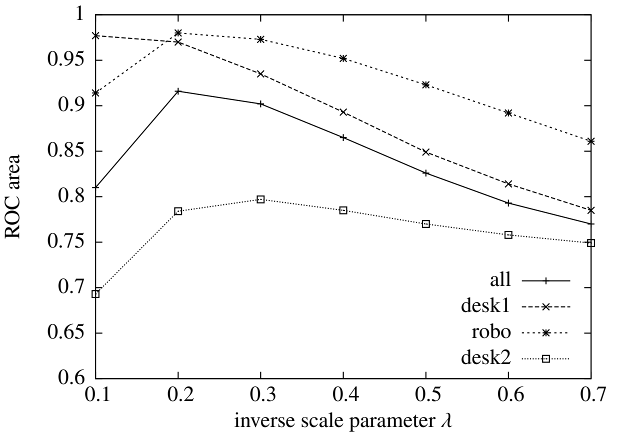

Figure 6. Comparison of the ROC area results for various parameters λ of the `probabilistic` method.

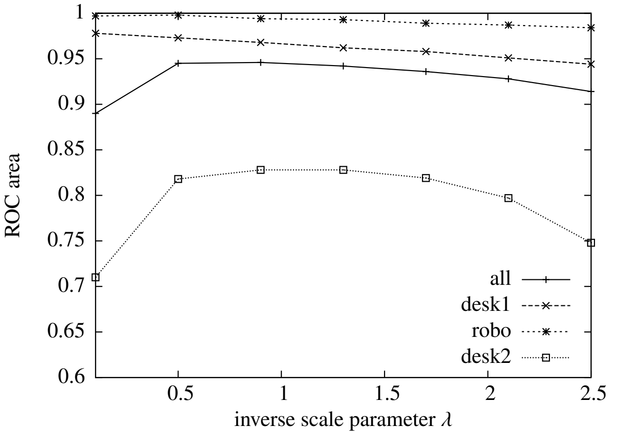

Figure 7. Comparison of the ROC area results for various parameters λ of the `probabilistic NN` method.

$\lambda = 0.2$. The method `probabilistic NN` performs best using $\lambda = 0.5$. We use these values in all other experiments.

Note that in case of `probabilistic NN`, the choice of the parameters m and λ is not very critical, since higher values lead to only slightly worse results.

5.2. Comparison of Methods

Figures 8 and 9 show the results of all three experiments combined in one ROC plot for each method. At very low false positive rates, `Snavely` shows the best performance, closely followed by `probabilistic NN` and `probabilistic`. However, starting at slightly higher false positive rates, `probabilistic NN` and `probabilistic` achieve a higher true positive rate. `Probabilistic NN` clearly shows the best overall performance. Only at high false positive rates, `geometric entropy NN` reaches higher true positive rates than `Snavely`. Within the group of geometric measures, `geometric entropy NN` gives the best results. The variant using the two nearest neighbors rejection during correspondence matching (`geometric entropy 2NN`) performs slightly worse. This indicates that strict rejection of ambiguities during the matching process is disadvantageous for the geometric measures. The remaining two measures, `geometric information NN` and `geometric smoothed NN`, perform worse than the entropy measure. `Geometric smoothed NN` produces acceptable results only at low false positive rates.

Figure 10 shows the ROC area results of all methods evaluated on each of the three sequences as well as jointly on all sequences. In case of the two easier sequences, `desk1` and `robo`, the methods `Snavely` and `probabilistic NN` perform almost identically. On the difficult sequence `desk2`, however, our methods `probabilistic`

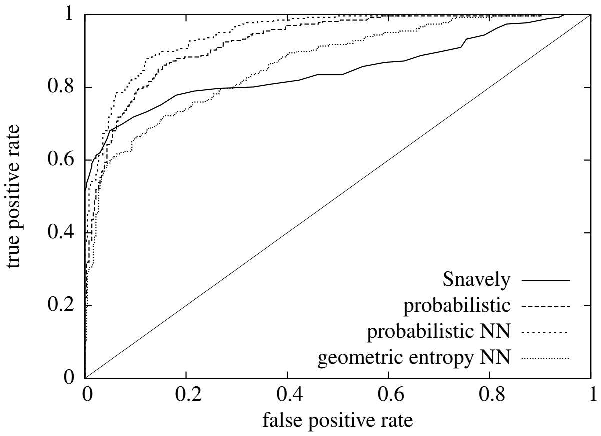

Figure 8. ROC plot comparing the methods `Snavely`, `geometric entropy NN`, `probabilistic`, and `probabilistic NN` jointly on all three sequences.

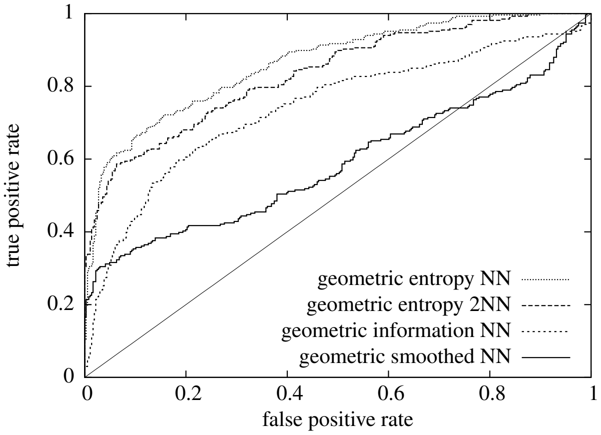

Figure 9. ROC plot comparing the methods `geometric entropy NN`, `geometric entropy 2NN`, `geometric information NN`, and `geometric smoothed NN` jointly on all three sequences.

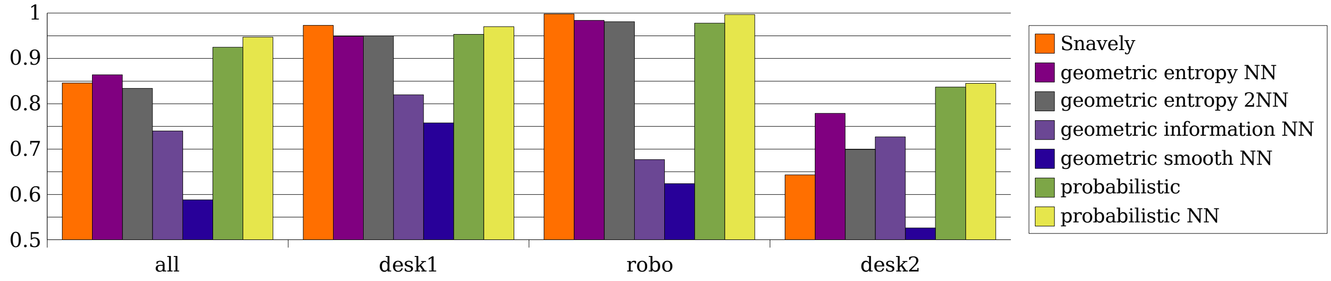

Figure 10. Comparison of the ROC areas of all methods evaluated jointly on all three sequences and also individually on each single one.

`NN`, `probabilistic` and `geometric entropy NN` perform much better than `Snavely`. The best overall results are achieved by `probabilistic NN`.

The good results of the probabilistic dissimilarity measures confirm our hypothesis that avoiding hard matching decisions leads to improved common field of view detection. The observation that `probabilistic NN` outperforms `probabilistic` is consistent with Lowe's [7] observation, that absolute SIFT descriptor distances are inferior to a comparison of the first and second nearest neighbor. Somewhat surprisingly, the geometric methods were not able to outperform the results of the `Snavely` method.

6. Conclusions

We presented two new approaches to the common field of view problem. The first one uses the uncertainty of relative pose estimates as an indication for images with a common field of view. We compared three such uncertainty measures [1]. The second approach is based on the entropy of correspondence probabilities, for which we presented two different models.

In quantitative experiments on three hand-labeled image sequences of varying difficulty, we assessed the performance of our new methods and compared them to the criterion used by Snavely et al. [11]. We showed that both variants of our probabilistic method achieve a great improvement over the Snavely method in complicated situations containing ambiguities and a wide baseline.

7. Acknowledgements

This research was partially funded by the Carl-Zeiss-Stiftung.

References

[1] F. Bajramovic and J. Denzler. Global Uncertainty-based Selection of Relative Poses for Multi Camera Calibration. In *Proceedings of the British Machine Vision Conference (BMVC)*, volume 2, pages 745–754, September 2008.

[2] R. Chang, S.-H. Ieng, R. Benosman, L. Lachèze, and T. Debaecker. Auto-Organized Visual Perception Using Distributed Camera Network. In *OMNIVIS'2008, the Eighth Workshop on Omnidirectional Vision, Camera Networks and Non-classical Cameras*, 2008.

[3] J. Domke and Y. Aloimonos. A Probabilistic Notion of Camera Geometry: Calibrated vs. Uncalibrated. In *Symposium of ISPRS Commission III Photogrammetric Computer Vision (PCV)*, pages 260–265. ISPRS, 2006.

[4] C. Engels and D. Nistér. Global uncertainty in epipolar geometry via fully and partially data-driven sampling. In *ISPRS Workshop BenCOS: Towards Benchmarking Automated Calibration, Orientation and Surface Reconstruction from Images*, pages 17–22, 2005.

[5] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2nd edition, 2003.

[6] E. Hörster and R. Lienhart. Calibrating and optimizing poses of visual sensors in distributed platforms. *ACM Multimedia System Journal, Special Issue on Multimedia Surveillance System*, 12(3):195–210, 2006.

[7] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[8] D. Martinec and T. Pajdla. Robust Rotation and Translation Estimation in Multiview Reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2007.

[9] K. Mikolajczyk and C. Schmid. A Performance Evaluation of Local Descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 27(10):1615–1630, 2005.

[10] E. Oto, F. Lau, and H. K. Aghajan. Color-Based Multiple Agent Tracking for Wireless Image Sensor Networks. In *Proceedings of the 8th International Conference on Advanced Concepts for Intelligent Vision Systems (ACIVS)*, pages 299–310, 2006.

[11] N. Snavely, S. Seitz, and R. Szeliski. Modeling the World from Internet Photo Collections. *International Journal of Computer Vision (IJCV)*, 80(2):189–210, 2008.

[12] H. Stewénius, C. Engels, and D. Nistér. Recent Developments on Direct Relative Orientation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 60(4):284–294, 2006.

[13] P. Torr and A. Zisserman. MLESAC: A New Robust Estimator with Application to Estimating Image Geometry. *Computer Vision and Image Understanding*, 78(19):138–156, 2000.