2013 IEEE Conference on Computer Vision and Pattern Recognition

SWIGS: A Swift Guided Sampling Method

Victor Fragoso Matthew Turk
University of California, Santa Barbara
{vfragoso, mturk}@cs.ucsb.edu

Abstract

We present SWIGS, a Swift and efficient Guided Sampling method for robust model estimation from image feature correspondences. Our method leverages the accuracy of our new confidence measure (MR-Rayleigh), which assigns a correctness-confidence to a putative correspondence in an online fashion. MR-Rayleigh is inspired by Meta-Recognition (MR), an algorithm that aims to predict when a classifier's outcome is correct. We demonstrate that by using a Rayleigh distribution, the prediction accuracy of MR can be improved considerably. Our experiments show that MR-Rayleigh tends to predict better than the often-used Lowe's ratio, Brown's ratio, and the standard MR under a range of imaging conditions. Furthermore, our homography estimation experiment demonstrates that SWIGS performs similarly or better than other guided sampling methods while requiring fewer iterations, leading to fast and accurate model estimates.

1. Introduction

Many computer vision tasks have to deal with incorrect image feature correspondences to estimate various types of models, such as homography, camera matrix, and others. Estimating these models robustly and quickly is very important for applications such as image registration [3], image-based localization [12, 13, 16], and many others. RANSAC [9] has been the method of choice for robustly estimating these models, and several improvements have been developed to increase its efficiency and improve its accuracy, *e.g.*, [2, 5, 6, 7, 10, 19, 20].

For many applications these estimates have to be computed as quickly and efficiently as possible. With this objective in mind, many approaches exploit some prior information (*e.g.*, geometrical and matching information) to compute a set of confidences that are used to select image feature correspondences for generating models. In this work, we focus on exploiting the matching scores to compute these confidences. In Fig. 1 we depict the general pipeline of a "guided sampling" robust model estimation.

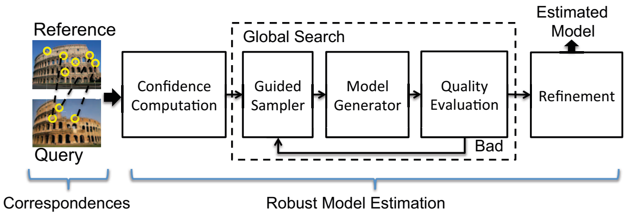


Figure 1. A set of image feature correspondences are passed to the robust model estimation process. For every correspondence, a correctness-confidence is computed. Subsequently, these confidences are used for sampling image correspondences to generate a model hypothesis. Finally, the estimation process stops iterating when a good model is found.

We introduce SWIGS, an efficient and fast method for guiding the data selection process, also referred as global search, that has only a single parameter. Our approach performs as well as the aforementioned methods but requires fewer iterations. SWIGS assumes that every feature correspondence has its own pair of correct and incorrect matching scores distributions, and it computes a confidence on the fly using our proposed MR-Rayleigh confidence measure for every correspondence by analyzing the closest matching scores obtained by the feature matcher. Assigning a confidence on the fly can be important for applications where environmental conditions can drastically change the matching scores distributions for correct and incorrect image feature correspondences, or when real-time performance is desired.

Moreover, we show that MR-Rayleigh can be used to predict correct image feature correspondences more accurately than Lowe's ratio [14], Brown's ratio [3], and Meta-Recognition [18] under a variety of different imaging conditions. Predicting when an image feature correspondence is correct is used and can be extremely beneficial in image based localization [12, 13, 16], where the prediction is used to keep "good" matches, and other applications. MR-Rayleigh can be easily applied to any recognition task as well, such as biometrics, object recognition, and others.

Our contributions are:

1. MR-Rayleigh: A confidence metric that allows more accurate predictions of correct matches and enables an

1063-6919/13 $26.00 © 2013 IEEE
DOI 10.1109/CVPR.2013.357

efficient and quicker guided sampling, under the assumption that every correspondence has its own correct and incorrect matching scores distributions.

2. SWIGS: A fast and efficient guided sampling process for robust model estimation based on MR-Rayleigh confidences that only has a single parameter to tune and does not need an offline stage.

2. Previous work

There exists a rich literature about computing weights or confidences to bias the selection of image feature correspondences to generate models in a robust model estimation process. These approaches in general exploit prior information such as matching scores ([2, 11, 19]) or geometrical cues ([5, 17]) to compute these weights. In Fig. 1 we show an overview of the main loop in a robust model estimation, where the confidences or weights are used to select feature correspondences and generate hypotheses. We call "sampling strategy" to the selection of image feature correspondences using the computed confidences.

In this section, we review the approaches that use matching scores as priors to compute a sampling strategy, as well as methods for predicting correct correspondences. Note that we use the terms match and correspondence interchangeably throughout the rest of the paper.

2.1. Prediction of correct matches

Lowe's ratio [14] has been one of the most efficient and widely used heuristics for predicting the correctness of a putative correspondence. The ratio compares the first nearest neighbor matching score against the second nearest neighbor matching score. This ratio exploits the fact that correct matching scores tend to be distant from the incorrect matching scores, consequently producing lower values (assuming a distance-based matching score). Finally, a threshold on the ratio is used for predicting correctness.

Brown *et al.* [3] extend Lowe's ratio by comparing the first nearest neighbor matching score against the average of the second nearest neighbor matching scores of multiple correspondences. Brown and colleagues report that this extension improved prediction performance.

A more elaborate method for predicting correct matches was introduced by Cech *et al.* [4]. This method uses a sequential co-segmentation process to obtain more information about the correctness of the correspondence. The method stops co-segmenting when it has enough evidence to declare a putative correspondence correct.

Our predictor uses only matching scores, as collecting other cues as in [4] can take extra time. We show that using the closest matching scores to the nearest-neighbor score can reveal useful information about the correctness of a putative correspondence, which boosts the prediction performance considerably. We will analyze and discuss this point throughly in Sec. 3.

2.2. Guided sampling using matching scores

Tordoff and Murray [19] calculate the correctness-confidence by considering the matching scores and the number of correspondences to which a feature was matched. Their method requires fitting appropriate curves to the empirical correct/incorrect densities a priori. Then, the probability that a match is correct, given all the matching scores of its potential matches, is calculated and used for biasing the selection of matches that are more likely to be valid.

The BEEM's global search mechanism [11] estimates the correct and incorrect correspondence distributions for a given pair of images by considering Lowe's ratio as the random variable. BEEM estimates these distributions by using kernel density estimation after classifying each correspondence as correct/incorrect using Lowe's ratio. Subsequently, BEEM estimates the mixing parameter between the computed distributions, and calculates the correctness confidences using the distributions and the mixing parameter. BEEM then assumes that the statistics of the matching scores are fixed for the given pair of images.

The BLOGS global search mechanism [2] computes the confidences by considering the highest and the two closest similarity scores. An important feature of this method, which is similar to our approach, is that it computes the confidences on the fly as it only requires the similarity scores of every feature. Hence, BLOGS considers that the statistics of the matching scores are defined per correspondence.

In contrast with most of the previous approaches, except BLOGS, SWIGS assumes that every correspondence has its own correct and incorrect matching scores distributions. To compute the confidence for every match, we exploit information from the tail of the incorrect matching score distribution. We will discuss in more depth these confidence computations in the following section.

3. Swift guided sampling

In this section we describe the keypoint matching process used in SWIGS. Given a query descriptor $\mathbf{q}_i$ and a pool of reference descriptors $\{\mathbf{r}\}_{j=1}^{n}$, a feature matcher decides the best putative correspondence following the nearest-neighbor rule:

$$j^{\star} = \arg\min_{j} \|\mathbf{q}_i - \mathbf{r}_j\|_2, \tag{1}$$

where $s_{i,j} = \|\mathbf{q}_i - \mathbf{r}_j\|_2$ is the matching score (or score for short). Two descriptors are correctly matched when their associated features correspond to the same location in the scene. In practice, the minimum matching score can belong to a correct or incorrect match due to several nuisances; *e.g.*, a minimum matching score produced by an incorrect image correspondence can be obtained when the scene contains repetitive structures.

We can consider the sequence of matching scores $\{s_{i,1}, \ldots, s_{i,n}\}$ for a single query descriptor $\mathbf{q}_i$ as a sequence composed by scores generated independently from a correct matching-scores distribution F_c and an incorrect matching-scores distribution $F_{\bar{c}}$. The matcher selects the minimum score from the sequence as the best match, which can either correspond to a minimum score generated from F_c or $F_{\bar{c}}$. The correct score (if any) can be the second, third, or other ranked score in the sequence, and hence we must consider overlapping distributions.

3.1. Meta-Recognition

In this section, we briefly review Meta-Recognition (MR) [18] and discuss some of its challenges in the context of feature matching. The objective of MR is to predict the correctness of a classifier; in our context we are interested in knowing whether a putative match is likely to be correct or incorrect. To achieve this objective, MR considers a ranked sequence of scores for a given query and selects the best ranked k scores $s_{1:k}$ (the k lowest scores). Subsequently, MR fits a Weibull distribution (W) to the selection discarding the lowest score s_1; *i.e.*, it uses $s_{2:k}$ for modeling the tail of $F_{\bar{c}}$. Finally, MR tests if s_1 is an outlier of the tail model in order to classify it as correct. The MR-Weibull prediction process can be summarized as follows:

$$\text{Prediction}(s_1) = \begin{cases} \text{Correct} & \text{if } W(s_1; \lambda, \eta) > \delta \\ \text{Incorrect} & \text{otherwise} \end{cases} \qquad (2)$$

where $W(s_1; \lambda, \eta)$ is the Complementary Cumulative Distribution Function (CCDF); λ and η are the scale and shape parameters; and δ is a threshold.

$$W(s; \lambda, \eta) = e^{-\left(\frac{s}{\lambda}\right)^{\eta}} \qquad (3)$$

As discussed earlier the lowest (best) matching score s_1 was generated either by $F_{\bar{c}}$ or F_c (depicted in Fig. 2a). Meta-Recognition's goal is to classify s_1 as correct or incorrect, and a threshold α corresponding to the crossover of F_c and $F_{\bar{c}}$ suffices for the task. However, we do not have enough information to determine exactly F_c or $F_{\bar{c}}$. MR-Weibull takes a different approach to achieve this goal by leveraging the fact that we have more samples from $F_{\bar{c}}$, and models its tail with a Weibull distribution W (depicted in Fig. 2b). Under the assumption that F_c is predominantly to the left of $F_{\bar{c}}$, MR-Weibull uses W (the CCDF of the tail model) for testing whether s_1 is an outlier, in which case it is classified as a correct match (see Fig. 2d). Nevertheless, the tail-fitting process in MR-Weibull can be affected by correct matching scores that are present in $s_{2:k}$ causing a bad model of the tail W and affecting the prediction. In particular, this scenario can happen when dealing with scenes that contain repetitive textures.

3.2. MR-Rayleigh

We have found that the accuracy and robustness of Meta-Recognition (MR) can be improved in the context of feature matching by using a Rayleigh distribution. Rayleigh's CCDF,

$$R(s; \sigma) = e^{-\frac{s^2}{2\sigma^2}}, \qquad (4)$$

has a single parameter to estimate and can reduce sensitivity. We estimate σ from the closest scores $s_{2:k}$ using the maximum-likelihood formula:

$$\hat{\sigma} = \sqrt{\frac{1}{2\,(k-1)} \sum_{j=2}^{k} s_j^2} \qquad (5)$$

where s_j is the j-th ranked matching score. An advantage of this new MR-Rayleigh approach is that $\hat{\sigma}$ can be computed efficiently, which is a desired property for applications that demand (near) real-time.

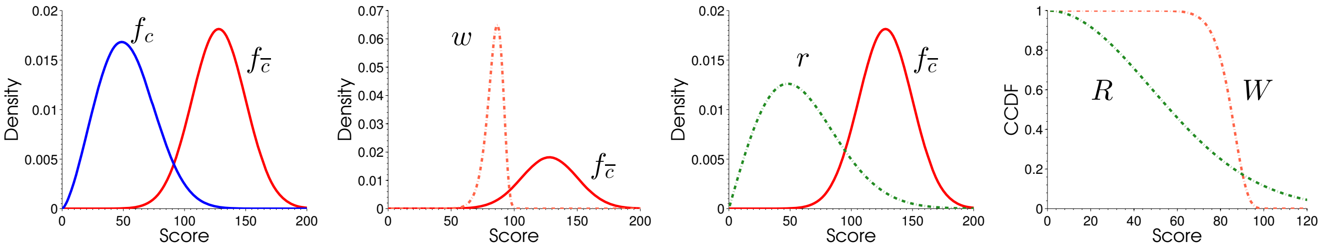
Intuitively, MR-Rayleigh finds a CCDF (R) that sets most of its mass over the support of F_c; the support is assumed to be predominantly to the left of $F_{\bar{c}}$ (depicted in Fig. 2c). Moreover, R decays gradually as soon as the matching score approaches the region of the tail of $F_{\bar{c}}$. Hence, MR-Rayleigh assigns a higher confidence to those matching scores that fall to the left of $F_{\bar{c}}$ and a lower confidence to those that fall near $F_{\bar{c}}$, in contrast with MR-Weibull, which assigns the confidence of one over the support of F_c (illustrated Fig. 2d), and abruptly falls near $F_{\bar{c}}$. Hence, MR-Weibull can assign a high confidence to scores corresponding to incorrect matches that fall near the distribution's crossover, yielding false-alarms.

3.3. Guided Sampling using MR-Rayleigh

The main idea of guided sampling for model fitting from feature correspondences is to use the computed confidences $\{c_l\}_{l=1}^{N}$ of being a correct match, where l indicates the index of a putative correspondence. Several approaches estimate c_l by using the following relationship

$$c_l = p(c|x_l) = \frac{p(x_l|c)p(c)}{p(x_l|c)p(c) + p(x_l|\bar{c})\,(1 - p(c))} \qquad (6)$$

where x_l (the random variable) can be either matching scores [19] or Lowe's ratios [11], $p(x_|c)$ and $p(x_|\bar{c})$ are the likelihoods of being correct and incorrect matches respectively, and $p(c)$ is the probability that a correct correspondence is selected. To estimate these confidences with this relationship, we therefore need to know the likelihoods $p(x_|c)$ and $p(x_|\bar{c})$, and the prior $p(c)$. Several approaches spend time in estimating these data either offline [19] or online [11], where the former requires a representative dataset of matching scores, and the latter uses only the matching



(a) (b) (c) (d)

Figure 2. Illustration of densities involved in a keypoint matching process per query feature[1]. (a) The lowest matching score is produced either by f_c (the matching scores distribution of correct matches) or $f_{\bar{c}}$ (the matching scores distribution of incorrect matches). The goal of this work is to estimate which distribution produced the minimum score. (b) Meta-Recognition models the tail of $f_{\bar{c}}$ with a Weibull distribution w to then calculate a confidence using the CCDF and predict correctness (d). MR-Rayleigh approximates the support of f_c by computing a Rayleigh distribution r from data taken from the tail to calculate a confidence using the CCDF and predict correctness (d).

scores obtained from the query and reference images to calculate the required Lowe's ratios. Other approaches, *e.g.*, BLOGS [2], define their own confidence using the best and closest similarity scores:

$$c_l = \left(1 - e^{-m_l}\right)^2 \left(1 - \frac{m_{lr}}{m_l}\right) \left(1 - \frac{m_{lc}}{m_l}\right) \qquad (7)$$

where c_l is the confidence assigned to the l-th correspondence, m_l is the best similarity score, and m_{lr} and m_{lc} are the two closest similarity scores obtained from a similarity matrix. BLOGS assigns a higher confidence when the greatest similarity score is high and is distant from its closest scores, and the confidence is severely penalized when its closest scores are near the greatest similarity score.

We calculate the confidence of being a correct match using MR-Rayleigh (see Sec. 3.2):

$$c_l = R(s_1^{(l)}; \hat{\sigma}) \qquad (8)$$

where $s_1^{(l)}$ is the best matching score of the l-th correspondence, and $\hat{\sigma}$ is calculated with its $k-1$ closest scores ($s_{2:k}^{(l)}$).

SWIGS avoids any density estimation and/or an offline stage; instead, it calculates a confidence on the fly. It requires a single parameter to tune k (discussed in Sec. 4), and avoids other complicated parameters such as a kernel to be used and its bandwidth, or distributions/curves to fit. SWIGS requires only the calculation of $\hat{\sigma}$, which can be calculated quickly and efficiently via Eq. (5).

4. Experiments

To assess the performance of SWIGS, we present two experiments: correct matches prediction accuracy, and a guided sampling experiment for estimating homography from image feature correspondences.

4.1. Datasets

In all the experiments we used the publicly available Affine Covariant Features Dataset used in [15]. This dataset contains eight sub-datasets (graf, wall, bark, boat, bikes, trees, leuven, and ubc), each with systematic variations of a single imaging condition: viewpoint (graf, wall), scale (bark, boat), image blur (bikes, trees), illumination (leuven), or jpeg compression (ubc). Every sub-dataset contains six images: a reference image and five query images of the same scene varying a single imaging condition. In addition, every sub-dataset provides five homographies that relate the reference image with each of the query images in the sub-dataset.

We used OpenCV's Hessian keypoint detector for finding approximately 2000 interest points per image. We used OpenCV's implementation of SIFT [14] and SURF [1] for describing the keypoints and we included (non-optimized) C++ code to calculate MR-Rayleigh, MR-Weibull, Brown's ratio, and Lowe's ratio into the brute-force feature matcher in OpenCV. With these modifications, the matcher returns either a confidence (MR-Rayleigh or MR-Weibull) or a ratio (Lowe's or Brown's) for every putative correspondence. We matched the reference keypoints (found in the reference image) against the query keypoints (detected per query image) for every sub-dataset. We then identified the correct matches by evaluating the following statement

$$\|\mathbf{x_q} - \mathbf{H}\mathbf{x_r}\|_2 < \epsilon \qquad (9)$$

where $\mathbf{x_q}$ and $\mathbf{x_r}$ are the query and reference keypoints, $\mathbf{H}$ is the homography transformation provided in the dataset that relates the reference and the query image, and $\epsilon = 5$ pixels is a threshold. Those matches that did not comply were labeled as incorrect matches. These identified correct correspondences were verified manually and used as our ground truth in our experiments.

We generated a tuning dataset for determining the values of k and δ for MR-Rayleigh and MR-Weibull, and

[1] We drew S = 1000 samples from $F_{\bar{c}} = \mathcal{N}(128, 30)$ to compute W with $k = 5\%$ of S and R with $k = 0.1\%$ of S.

the threshold τ_{BR} for Brown's ratio. For every sub-dataset we generated eight different random affine transformations. Then, we use these generated transformations to obtain eight images from the reference image of each sub-dataset. We then detected approximately 1000 interest points on every image, we calculated their descriptors (SIFT and SURF), and computed the correspondences between the reference image and each generated image per descriptor. Subsequently, we then identified the correct correspondences in a similar manner as described earlier but using the affine transformations instead of the homographies in Eq. (9).

4.2. Correct match prediction experiment

In this experiment we are interested in measuring the performance of MR-Rayleigh on detecting correct matches; and we use the labeled correct correspondences as our ground truth. We considered a True-Positive when the predictor accurately detects a correct match, and a False-Positive when the predictor inaccurately detected a positive match, *i.e*., a false alarm. We used the False-Positive rate (FPR) and True-Positive rate (TPR) to determine the values of k and δ for MR-Rayleigh and MR-Weibull (see [8] for FPR and TPR calculation). We ran large series of predictions using the tuning dataset mentioned earlier and selected k and δ for MR-Rayleigh and MR-Weibull per descriptor such that the operation point $\text{op} = (\text{FPR}, \text{TPR})$ was as close as possible to the ideal operation point $\text{op}^{\star} = (0, 1)$, *i.e*., the lowest FPR and the maximum TPR. We found that $k_{\text{MR-Rayleigh}} = 0.5\%$ of n and $k_{\text{MR-Weibull}} = 2\%$ of n worked the best for SIFT and SURF matches, where n is the number of reference features. We also tuned Brown's ratio (BR) on the same dataset and in the same manner and found $\tau_{\text{BR}} = 0.73$ and $\tau_{\text{BR}} = 0.709$ were good thresholds for SIFT and SURF matches respectively, while for Lowe's ratio (LWR) we used the recommended threshold of $\tau_{\text{LWR}} = 0.8$ for both SIFT and SURF matches.

We present five different receiver operating characteristics (ROC) curves per descriptor in Fig. 3. The top row corresponds to SIFT matches and the bottom row to SURF matches, and each column presents results for a different imaging condition; with the exception of the first column, which presents the results over all imaging conditions. We used the best values found for k where $n \approx 2000$ reference features (*i.e*., $k_{\text{MR-Weibull}} \approx 40$ and $k_{\text{MR-Rayleigh}} \approx 10$). For Lowe's ratio and Brown's ratio we predict a correct match when such a ratio is lower than a threshold τ. We varied every threshold of each predictor in the range of 0 to 1 with steps of size 10^{-4}.

We can observe in the first column of Fig. 3 that MR-Rayleigh (MRR) outperformed MR-Weibull (MRW), Lowe's ratio (LWR), and Brown's ratio (BR) over all imaging conditions for SIFT and SURF matches. From the subsequent columns we can conclude that MR-Rayleigh tends to perform better in most cases: the rate of true-positives overall tends to be higher than for MR-Weibull, Lowe's ratio and Brown's ratio. A problem we observed with MR-Weibull is the sensitivity of its threshold: the effective range is between 0.9 and 1 to be discriminative; in fact, this is the reason for choosing a small step size for the thresholds. This threshold sensitivity explains the abrupt "jumps" in the ROC curves for SIFT matches in the first, second, fourth, and fifth columns, as a tiny variation in the threshold can affect drastically the prediction accuracy of MR-Weibull; the True-Positive rate drastically drops when the False-Positive rate is low. Consequently, MR-Weibull can struggle in detecting correct matches when a low False-Positive rate is required. In contrast, MR-Rayleigh does not suffer this threshold sensitivity and it can be used when a low False-Positive rate is required. Lowe's ratio in general performs competitively for SIFT and SURF matches, whereas, Brown's ratio tends to perform competitively for SIFT matches but tends to fall short for SURF matches.

Table 1. Optimal operation points for predictors.

Predictor	Thld.	FPR	TPR	F
SIFT				
Lowe's ratio	0.8000	**0.07**	0.76	0.78
Brown's ratio	0.7300	0.10	0.80	0.78
MR-Weibull	0.9999	0.21	**0.90**	0.74
MR-Rayleigh	0.6000	0.11	0.85	**0.80**
SURF				
Lowe's ratio	0.8000	**0.04**	0.64	0.73
Brown's ratio	0.7090	0.05	0.61	0.68
MR-Weibull	0.9999	0.06	0.69	0.72
MR-Rayleigh	0.6000	0.06	**0.71**	**0.75**

We also conducted an experiment on detecting correct matches per descriptor on the entire testing dataset using the thresholds found during our tuning stage. The goal of the experiment is to assess the performance of these predictors using the best parameters found in our tuning stage. We present False-Positive rate (FPR), True-Positive rate (TPR), and the F-score per descriptor (see [8] for F-score calculation) as the results of this experiment in Table 1. We calculated the F-score to assess performance in a unified manner. From the results of this experiment we can conclude that Lowe's ratio returned the lowest False-Positive rate (FPR) regardless of the descriptor. MR-Weibull produced the highest True-Positive rate for SIFT matches but with the highest False-Positive rate, while MR-Rayleigh produced a high True-Positive rate and a low False-Positive rate. For SURF matches MR-Rayleigh produced the highest True-Positive rate and a low False-Positive rate. MR-Rayleigh has the highest F-score for SIFT and SURF matches, which suggests that MR-Rayleigh is a good detector of correct image feature correspondences.

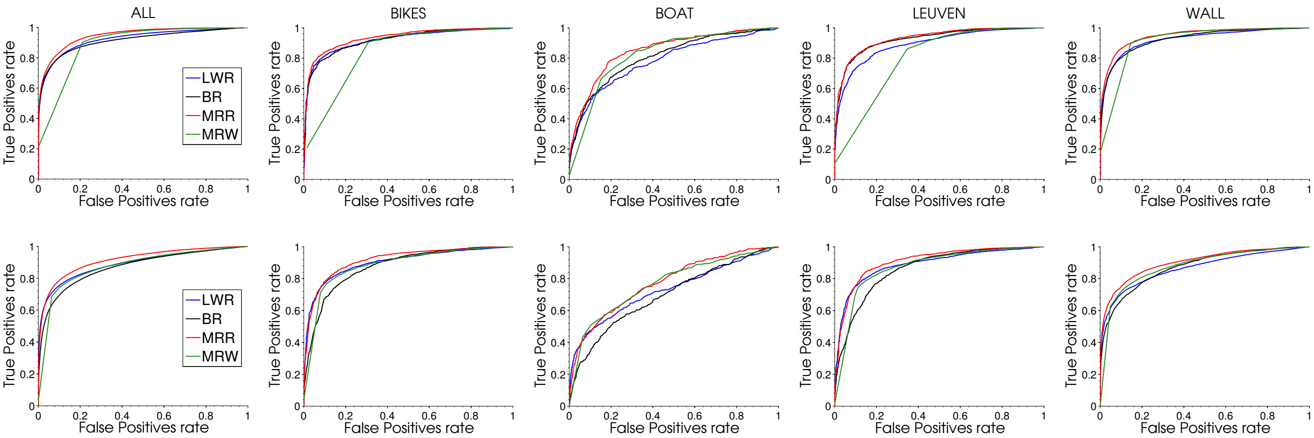


Figure 3. ROC curves for evaluating correct matches predictions by using MR-Weibull (MRW), MR-Rayleigh (MRR), Lowe's ratio (LRW), and Brown's ratio (BR). The top row presents the results for SIFT matches and bottom row for SURF matches. The first column presents the results over all sub-datasets while the second-to-fourth columns show the results on bikes, boat, leuven, and wall sub-datasets.

4.3. Homography estimation experiment

In this experiment we aim to evaluate the performance of SWIGS for homography estimation in a dense matching scenario, and compare it with several other sampling methods: BEEM [11]; a Guided-Sampling [19] with a general distribution considering all the imaging conditions (GEN); a Guided-Sampling [19] that considers only the distribution for a specific imaging condition (SPEC); BLOGS [2] where $m_l = {}^1\!/_{s_1}$, and $m_{lr} = m_{lc} = {}^1\!/_{s_2}$ as our approach considers a different matching procedure; and a classical random sampling (uniform distribution) for a baseline.

Each of these methods was plugged in to our own MLESAC [21] implementation, where the standard deviation of the residuals distribution was set to $\bar{\sigma} = 5$ pixels, and $w = 20$ as the parameter for the mismatched residuals distribution. Matlab was used to obtain the distributions required for the two Guided-Sampling [19] methods and to fit Weibull and Generalized Extreme Value distributions for correct and incorrect matches respectively (see Fig. 4). We implemented only the prior estimation stage of BEEM and BLOGS' global search mechanism, as we aim at comparing the confidence mechanism used for data sampling in a robust estimation. We used OpenCV's findHomography function (without the RANSAC option) and the correct matches identified by each method to estimate the homography.

We executed the experiment 5000 times with a stopping criterion of 100% of correct matches found and a maximum of 1000 iterations, since we are interested in applications that have a limited budget of iterations; an iteration is a completion of the loop in Fig. 1. We report the median of the number of iterations a method took to find the best model within the required number of iterations and the median of the percentage of correct matches that the best model found considered as a correct match. We used the same ground truth as in the previous experiment.

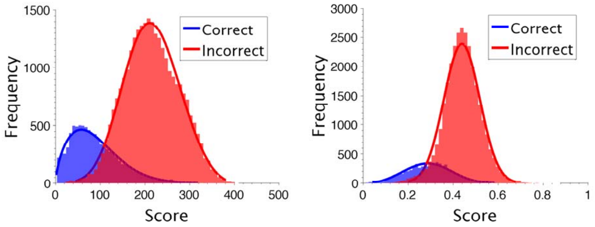


Figure 4. Fitted distributions for SIFT matches (left) and SURF matches (right) used in GEN. Similar distributions were obtained per sub-dataset for SPEC.

The results are shown in Fig. 5 (see supp. material for more detailed plots), where the first two rows show the results obtained for SIFT, and the rest for SURF matches. The percentage of correct matches are presented in the first and third rows, while the iterations are in the second and fourth rows. The x-axis indicates the index of the images contained in the considered sub-datasets (omitting the reference image, which is index 1); an increasing index represents increasing variation with respect to the reference image. Each column presents the results for a different sub-dataset: bikes, boat, graf, trees, and wall, from left-to-right.

We can observe that SWIGS tends to require in general fewer iterations than the other methods (second and fourth rows) to find models that consider a comparable or higher percentage of correct matches within the allowed number of iterations (first and third row). We note that SWIGS, SPEC, and BEEM tend to find models that consider approximately the same number of matches. The GEN method struggles more to find models that consider a high percentage of correct matches in scenes with repetitive textures, *e.g.*, wall, and trees sub-datasets; repetitive textures can cause a considerable overlap between correct and incorrect matching

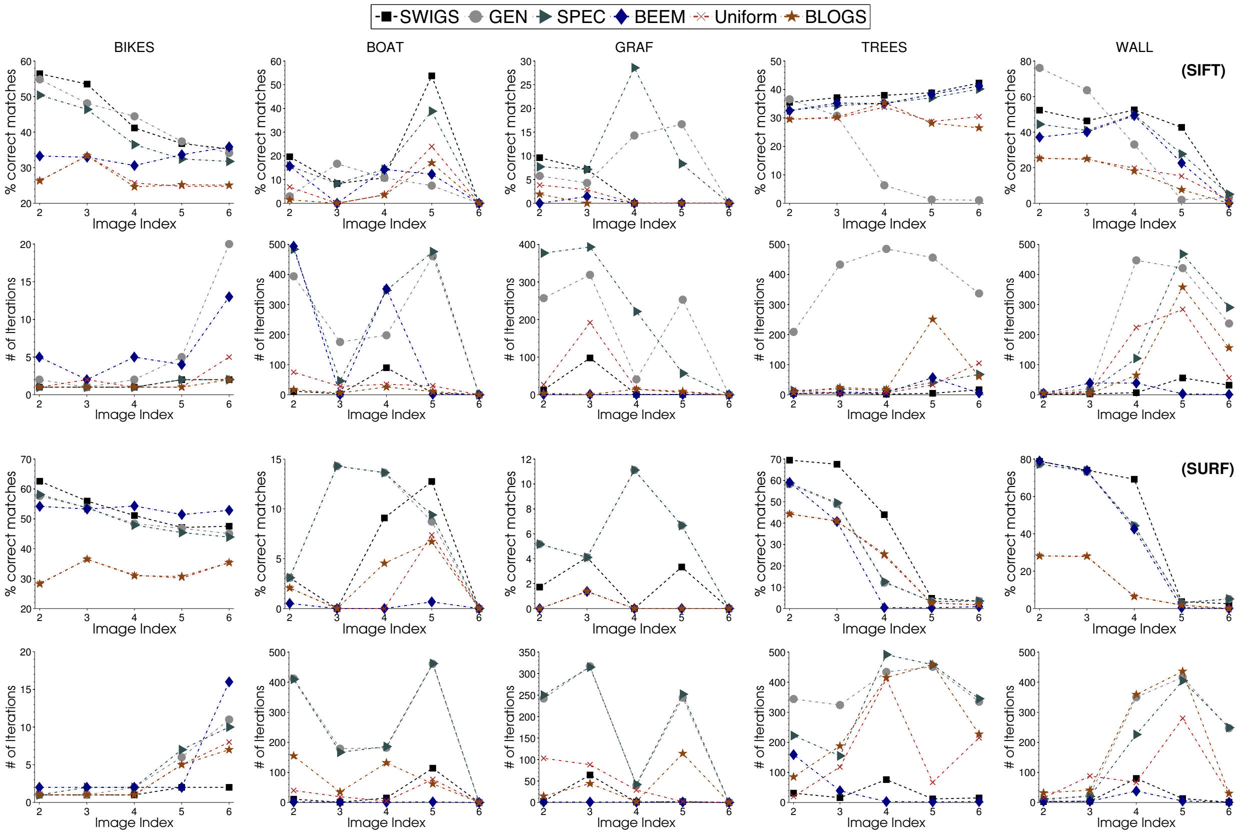


Figure 5. Performance evaluation across several sub-datasets (bikes, boat, graf, trees, wall from left-to-right). Of all the 5000 repetitions of the experiment, the first and third rows present the median of the percentage of correct matches found by the best computed models within the allowed number of iterations, while the second and fourth rows present the median number of iterations at which the best model was found. The first and second rows present the results for SIFT, and the third and fourth for SURF.

scores distributions (see Fig. 4). BLOGS and a random sampling (Uniform) method perform similarly in finding models that consider a high portion of the correct matches.

The experimental results presented in this section demonstrate that SWIGS can perform similarly or better in finding models that consider a good portion of correct matches in a dense matching scenario. The experiments also show that SWIGS tends to require fewer iterations than the other guiding sampling methods without sacrificing the number of correct matches found. Moreover, this confirms that MR-Rayleigh confidences tend to identify good matches, and these confidences yield an efficient and accurate sampling strategy. In Fig. 6 we show two different sets of correct image feature correspondences found with SWIGS and MLESAC.

5. Conclusions and future directions

We have introduced MR-Rayleigh, a confidence measure based on Meta-Recognition (MR) [18] for predicting correct image feature correspondences more accurately. MR-Rayleigh computes the confidence considering the k clos-

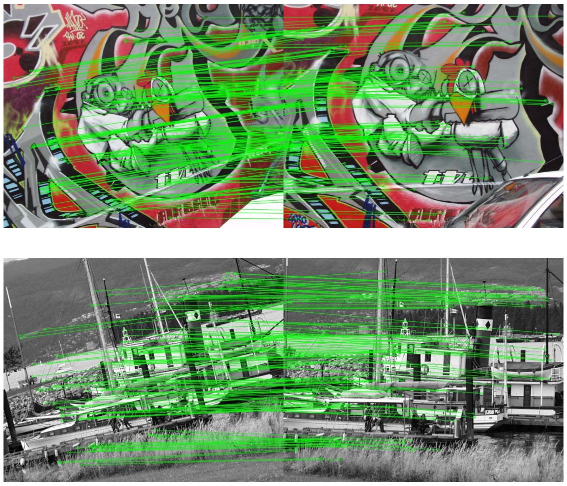


Figure 6. Correct SIFT matches found between the reference image (top-right) and query image (top-left) of the Graf dataset and correct SURF matches between reference image (bottom-right) and query image (bottom-left) of the Boat dataset using SWIGS and MLESAC.

est matching scores produced by the matcher when comparing the query descriptor against the reference descriptors. MR-Rayleigh assigns a higher confidence when the lowest matching score is closer to zero and gradually decays as it gets closer to the tail of the incorrect matching scores distribution. Moreover, MR-Rayleigh estimates a single parameter which is more efficient to compute and can be more robust to the data used for its estimation than the two Weibull's parameters required in MR-Weibull [18].

Our experiments showed that MR-Rayleigh outperformed Lowe's ratio, Brown's ratio, and MR-Weibull in predicting correct matches across several image correspondences obtained in different imaging conditions. This prediction is efficient to compute and can be useful in many applications such as image-based localization where only good matches are kept; in estimating the inlier-ratio, which can be used to estimate the maximum number of iterations in RANSAC, and others.

We also presented SWIGS, an efficient method to sample data in a guided manner for robust model fitting that exploits the confidence delivered by MR-Rayleigh. In comparison with other guided sampling methods (*e.g.*, BEEM [11] and Guided-MLESAC [19]) that assume a correct or incorrect matching score distribution for a pair of images or for an entire dataset, SWIGS considers that every query feature has a correct and incorrect matching scores distributions. SWIGS then computes the confidence of every correspondence on the fly and uses these confidences for sampling matches to estimate a model such as a homography.

Our homography estimation experiment suggests that SWIGS achieves competitive or better results than BEEM's and BLOGS's [2] guided sampling mechanisms, and Tordoff and Murray's guided MLESAC [19]. We believe that SWIGS can help applications that have no prior information of the environment where they will be used, such as image registration, feature-based tracking, SLAM, and other applications that use putative correspondences for estimating different models.

For future work, we plan to evaluate MR-Rayleigh's performance on other applications, such as object recognition, image retrieval, and others. In addition, we plan to extend MR-Rayleigh for similarity metrics.

Acknowledgments. V. Fragoso would like to thank UC MEXUS-CONACYT (Fellowship 212913) for the funding. We also would like to thank the anonymous reviewers and area chairs for the valuable feedback. This research was supported in part by NSF under award number 121926.

References

[1] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-up robust features (surf). *Comput. Vis. Image Underst.*, 110(3):346–359, June 2008. 4

[2] A. S. Brahmachari and S. Sarkar. BLOGS: Balanced local and global search for non-degenerate two view epipolar geometry. In *Proc. IEEE Intl. Conf. on Computer Vision*, 2009. 1, 2, 4, 6, 8

[3] M. Brown, R. Szeliski, and S. Winder. Multi-image matching using multi-scale oriented patches. In *Proc. IEEE Computer Vision and Pattern Recognition*, 2005. 1, 2

[4] J. Cech, J. Matas, and M. Perdoch. Efficient sequential correspondence selection by cosegmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(9):1568–1581, Sept. 2010. 2

[5] T.-J. Chin, J. Yu, and D. Suter. Accelerated hypothesis generation for multistructure data via preference analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(4):625–638, 2012. 1, 2

[6] O. Chum and J. Matas. Matching with PROSAC – progressive sample consensus. In *Proc. IEEE Computer Vision and Pattern Recognition*, 2005. 1

[7] O. Chum, J. Matas, and J. Kittler. Locally optimized RANSAC. In *Proc. Pattern Recognition (DAGM Symposium)*, 2003. 1

[8] T. Fawcett. An introduction to ROC analysis. *Pattern Recogn. Lett.*, 27(8):861–874, Jun. 2006. 5

[9] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, June 1981. 1

[10] L. Goshen and I. Shimshoni. Guided sampling via weak motion models and outlier sample generation for epipolar geometry estimation. In *Proc. IEEE Computer Vision and Pattern Recognition*, 2005. 1

[11] L. Goshen and I. Shimshoni. Balanced exploration and exploitation model search for efficient epipolar geometry estimation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(7):1230–1242, July 2008. 2, 3, 6, 8

[12] A. Irschara, C. Zach, J.-M. Frahm, and H. Bischof. From structure-from-motion point clouds to fast location recognition. In *Proc. IEEE Computer Vision and Pattern Recognition*, 2009. 1

[13] Y. Li, N. Snavely, and D. P. Huttenlocher. Location recognition using prioritized feature matching. In *Proc. European Conf. on Computer Vision*, 2010. 1

[14] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Intl. Journal of Computer Vision*, 60(2):91–110, Nov. 2004. 1, 2, 4

[15] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(10), Oct. 2005. 4

[16] T. Sattler, B. Leibe, and L. Kobbelt. Fast image-based localization using direct 2d-to-3d matching. In *Proc. IEEE Intl. Conf. on Computer Vision*, 2011. 1

[17] T. a. Sattler. SCRAMSAC: Improving RANSAC's efficiency with a spacial consistency filter. In *Proc. IEEE Intl. Conf. on Computer Vision*, 2009. 2

[18] W. J. Scheirer, A. Rocha, R. J. Micheals, and T. E. Boult. Meta-Recognition: The theory and practice of recognition score analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(8):1689–1695, Aug. 2011. 1, 3, 7, 8

[19] B. Tordoff and D. W. Murray. Guided sampling and consensus for motion estimation. In *Proc. European Conf. on Computer Vision*, 2002. 1, 2, 3, 6, 8

[20] P. H. S. Torr. Bayesian model estimation and selection for epipolar geometry and generic manifold fitting. *Intl. Journal of Computer Vision*, 50(1):35–61, Apr. 2002. 1

[21] P. H. S. Torr and A. Zisserman. MLESAC: A new robust estimator with application to estimating image geometry. *Comput. Vis. Image Underst.*, 78(1):138–156, Apr. 2000. 6