

# Adaptive Structure from Motion with a contrario model estimation

Pierre Moulon, Pascal Monasse, Renaud Marlet

▶ **To cite this version:**

Pierre Moulon, Pascal Monasse, Renaud Marlet. Adaptive Structure from Motion with a contrario model estimation. ACCV 2012, Nov 2012, Daejeon, South Korea. pp.257-270, 10.1007/978-3-642-37447-0_20 . hal-00769266

**HAL Id: hal-00769266**

**https://enpc.hal.science/hal-00769266**

Submitted on 30 Dec 2012

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Adaptive Structure from Motion with *a contrario* model estimation

Pierre Moulon[1,2], Pascal Monasse[1], Renaud Marlet[1]

[1]Université Paris-Est, LIGM (UMR CNRS), Center for Visual Computing, ENPC,
6-8 av. Blaise Pascal, 77455 Marne-la-Vallée, France.
[2]Mikros Image. 120 rue Danton, 92300 Levallois-Perret, France.
[1]firstname.lastname@enpc.fr, [2]pmo@mikrosimage.eu

**Abstract.** Structure from Motion (SfM) algorithms take as input multi-view stereo images (along with internal calibration information) and yield a 3D point cloud and camera orientations/poses in a common 3D coordinate system. In the case of an incremental SfM pipeline, the process requires repeated model estimations based on detected feature points: homography, fundamental and essential matrices, as well as camera poses. These estimations have a crucial impact on the quality of 3D reconstruction. We propose to improve these estimations using the *a contrario* methodology. While SfM pipelines usually have globally-fixed thresholds for model estimation, the *a contrario* principle adapts thresholds to the input data and for each model estimation. Our experiments show that adaptive thresholds reach a significantly better precision. Additionally, the user is free from having to guess thresholds or to optimistically rely on default values. There are also cases where a globally-fixed threshold policy, whatever the threshold value is, cannot provide the best accuracy, contrary to an adaptive threshold policy.

## 1 Introduction

There are numerous approaches to estimate the structure from motion (scene structure and camera motion) from multiple images. Thanks to recent progress in image matching and optimization, it is now possible to compute large scale 3D reconstruction from millions of internet images on reasonable sized cluster [1] or even on a single high-end computer [2]. All these methods aim at working with large datasets of images, but few consider the accuracy of the reconstruction.

Most current Structure from Motion (SfM) pipelines are sequential: they start from a minimal reconstruction and incrementally add new views using pose estimation and 3D point triangulation algorithms. There is no guarantee that the reconstruction converges to the global optimum solution. The implementation often relies on many bundle adjustment steps to optimize the solution and uses hard thresholds for robust model estimation. Recently the $L_\infty$ framework [3] has been shown to solve multi-view geometry problems, minimizing directly the maximal reprojection error rather than the sum of squared error. Although the global minimum is found using convex or linear programming, it becomes computationally expensive when dealing with outliers and large problems [4].

This paper makes use of the *a contrario* theory to study the adaptation of model estimation thresholds to input data. We show how to automatically compute these thresholds and illustrate the advantages: besides user-friendly parameterless procedures, we can also reach optimization levels that would be unattainable with globally-fixed thresholds. Our adaptive thresholds are implemented in a SfM pipeline that targets high precision. Examples output of our SfM pipeline applied to 128 (resp. 119) images are shown in Fig. 1. Note the wide variation of the automatic threshold for pose estimation. Our SfM produces a sparse 3D point cloud based on image feature points, not a dense 3D reconstruction; a subsequent multiple-view stereovision pipeline has to be used for that, such as PMVS [5] or the pipeline described in [6]. The dense reconstruction quality critically depends on the calibration computed from SfM.

The paper is organized as follows. Section 2 recalls the principles of Structure from Motion. Section 3 briefly reviews robust model estimation. Section 4 describes the *a contrario* methodology and its general application to model estimation. Section 5 describes the particular stages of the classical incremental SfM that we replace with a specific *a contrario* model estimation. Section 6 details evaluation results on real and synthetic datasets, and Section 7 concludes.

## 2 Structure from Motion — the classical pipeline

Structure from Motion (Fig. 2) computes an external camera pose per image (the motion) and a 3D point cloud (the structure) representing the pictured scene. Inputs are images and internal camera calibration information. Feature points are detected in each image (e.g., SIFT [9] or SURF [10]) and matched between image pairs. There are two main approaches to correlate detected features and solve the SfM problem: the *incremental* pipeline and the *global* method.

The incremental pipeline is a growing reconstruction process. It starts from an initial two-view reconstruction (the seed) that is iteratively extended by adding new views and 3D points, using pose estimation and triangulation. Due to the incremental nature of the process, successive steps of non-linear refinement, like bundle adjustment and Levenberg-Marquardt steps, are performed to minimize the accumulated error (drift) [11, 12].

The general feature correspondence and SfM processes are described in algorithms 1 and 2. The first algorithm outputs pairwise correspondences that are consistent with the estimated fundamental matrix. Homography estimation is used to choose an initial image pair with numerous correspondences while keeping a wide enough baseline. The second algorithm takes these correspondences as input and yields a 3D point cloud as well as the camera poses. Steps marked with a star ($*$) are those we redefine within the *a contrario* framework. This allows critical thresholds to be automatically adapted to the input images, which yields more accurate SfM as we shall see.

State of the art systems and methods for SfM include Bundler [13], Samantha [14], image triplets based approaches [15, 7] and Visual Odometry systems [16, 17]. All these systems and methods rely on RANSAC-based model estimation

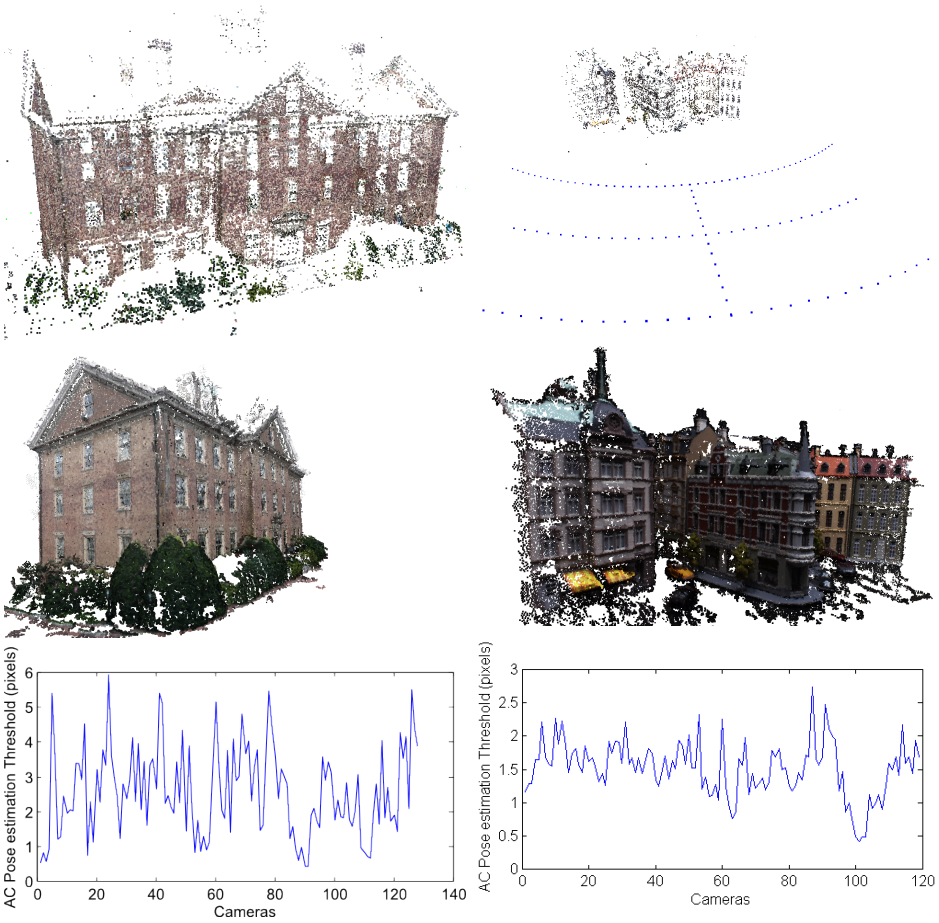



**Fig. 1.** From top to bottom: sparse 3D reconstruction from our SfM pipeline, PMVS densification [5] and variation of the automatic threshold of pose estimation. Left: the 128 images dataset from [7] source code. Right: the 119 images of 004 scene from [8] dataset. Estimated camera positions represented as blue dots.

to be robust to noise/false data. However, it introduces static thresholds, which have to be set empirically.

The global methods compute essential matrix for all possible input pairs and perform the reconstruction in a two-step process. First globally consistent rotations are computed from the relative pairwise rotations (see Martinec and Pajdla [18] and Govindu [19, 20]), then structure and translation equations are solved via the $L_\infty$ constraint [3], or $L_1$ penalization [4] to deal with outliers. As in the incremental pipeline, the basis of the method is a robust estimation of a model that is controlled by a static empirical threshold.

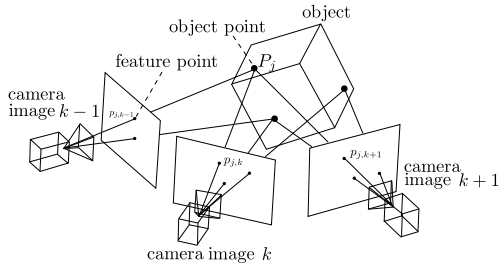



**Fig. 2.** Structure from Motion.

**Algorithm 1** Computation of geometry-consistent pairwise correspondences

**Require:** image set
**Ensure:** pairwise point correspondences that are geometrically consistent
  Compute putative matches:
    detect features in each image and build their descriptor
    match descriptors (brute force or approximate nearest neighbor)
  Filter geometric-consistent matches:
* estimate fundamental matrix $F$
* estimate homography matrix $H$

## 3 Parameterizing robust model estimation

Robust model estimation from noisy data that are corrupted by outliers is often performed with the RANSAC (RANdom SAmple Consensus) algorithm [21], or one of its variants. This is the case for the above-mentioned SfM systems.

RANSAC is a randomized procedure due to complexity considerations. It repeatedly selects random sample sets $S$ from the data, whose minimal size is sufficient to estimate the parameters of a model. At each trial, inliers are defined as the data that fits the model within an acceptable error threshold $T$. After a given number of iterations, the model parameters that maximize the number of corresponding inliers are returned.

The RANSAC algorithm depends on a critical parameter: the choice of threshold $T$. If $T$ is too small, then little data is selected as inliers, which leads to model imprecision and even, sometimes, to the impossibility to estimate a model because the number of inliers is too small. If $T$ is too large, then outliers (false positives) contaminate inliers, which also leads to inaccurate or wrong models.

But the user is generally clueless about the choice of a threshold value. This is very much the case for SfM. Even though SfM thresholds are generally expressed in pixels, which could make sense to the user, they actually refer indirectly to complex operations concerning feature points, and it is practically impossible to adjust or guess any sensible threshold by just looking at the pictures of a dataset. Threshold selection is exemplified in Fig. 3 for image registration.

**Algorithm 2** Incremental Structure from Motion

**Require:** internal camera calibration (matrix $K$, possibly from EXIF data)
**Require:** pairwise geometry consistent point correspondences
**Ensure:** 3D point cloud
**Ensure:** camera poses
  compute correspondence tracks $t$
  compute connectivity graph $G$ (1 node per view, 1 edge when enough matches)
  pick an edge $e$ in $G$ with sufficient baseline (compare $F$ and $H$)
∗ robustly estimate essential matrix from images of $e$
  triangulate $t \cap e$, which provides an initial reconstruction
  contract edge $e$
  **while** $G$ contains an edge **do**
    pick edge $e$ in $G$ that maximizes track($e$) $\cap$ {3D points}
∗   robustly estimate pose (external orientation/resection)
    triangulate new tracks
    contract edge $e$
    perform bundle adjustment
  **end while**



**Fig. 3.** Robust homography estimation. From left to right: RANSAC threshold (transfer error through homography) $T = 0.5$ pixels yields 6 points correspondences, threshold $T = 2$ pixels (i.e., default Bundler threshold) yields 19 points, and threshold $T = 6.8$ pixels yields 50 points as well as a better estimated homography. This last value was actually automatically computed with an *a contrario* technique, that statistically determines a confidence threshold (cf. Section 4).

RANSAC thus faces the user with a dilemma: setting a low threshold and possibly underestimating inliers, which may reduce model accuracy and jeopardize model existence, or setting a high threshold and possibly corrupt data with outliers, which may also decrease precision. In practice, the user relies on default threshold values, that work reasonably well although they might be sub-optimal.

Another issue relates to the globality of parameterization. In practical settings, many instances of a model estimation problem have to be solved independently, for different elements of a dataset. For instance, in SfM, pose has to be estimated many times for a number of different image pairs. The fact is that each problem instance calls for a specific threshold value, adapted to the corresponding data noise. However, most systems only accept a global threshold value for treating a whole dataset. Such a global threshold is naturally too low for some data, and too high for others. There are thus cases where even a perfect oracle can only provide a sub-optimal global parameterization.

## 4 *A contrario* model estimation

Our approach to address the issues listed in Section 3 is to use a methodology for finding a model that best fits the data with a confidence threshold that adapts automatically to noise. For this, we use an *a contrario* model estimation.

In this framework, the computed thresholds are such that they have a good chance of correctly telling apart inliers from outliers. As a result, the accuracy of model estimation tends to be as good as possible (given the sampling strategy), and there are less risks of inadvertently selecting too few inliers for a model to be estimated. Moreover, as thresholds adapt to data, they can vary depending on each image, which allows estimations that would otherwise be impossible with a globally-fixed threshold. Last, the user is free from having to set opaque values or to optimistically rely on default values. Automatic and specific *a contrario* threshold values are illustrated in 3.

### 4.1 The *a contrario* methodology

The *a contrario (AC)* methodology relies on the Helmholtz principle: "an observed strong deviation from the background model is relevant information". In other words, a configuration that is unlikely to be explained by chance is conspicuous. This theory has been first introduced by Desolneux *et al.* in [22] and applied to detection in images.

Applied to model estimation, the *a contrario* approach answers the question "Does this model arise by chance?" and thus decides the meaningfulness of a model. The corresponding statistical criterion is data-specific and avoids empirically setting thresholds for inlier/outlier discrimination. It thus provides a parameter-free evolution of RANSAC, called AC-RANSAC [25]. Additionally, once a meaningful model is found, the convergence of AC-RANSAC can be accelerated by reducing the number of random samples and drawing further samples among the inliers of this model. *A contrario* model estimation has first been introduced to estimate the fundamental matrix under the name of ORSA (Optimized-RANSAC) [24], later renamed as AC-RANSAC and extended to multiple model estimation under the name MAC-RANSAC [23].

AC-RANSAC looks for a consensus set that includes a controlled Number of False Alarms (NFA), as described below. A false alarm in this context is a model that is actually due to chance. This requires the definition of a background model $\mathcal{H}_0$ and of a rigidity measure. $\mathcal{H}_0$, called the null hypothesis, is a model of random correspondence: a pair of independent points that are uniformly distributed in their respective image. The rigidity measure is the residual error (of inliers) with respect to an estimated model.

The generic NFA for a rigid model $M$, which is a generalization of Moisan and Stival's NFA [24], also mentioned in [23], is:

$$NFA(M,k) = N_{\text{out}}(n - N_{\text{sample}}) \binom{n}{k} \binom{k}{N_{\text{sample}}} (e_k(M)^d \alpha_0)^{k - N_{\text{sample}}} \qquad (1)$$

where

- $k$ is the number of hypothesized inlier correspondences,
- $n$ is the total number of correspondences,
- $N_{\text{sample}}$ is the cardinal of a RANSAC sample,
- $N_{\text{out}}$ is the number of models that can be estimated from a RANSAC sample of $N_{\text{sample}}$ correspondences ($N_{\text{sample}}$ is often such that $N_{\text{out}} = 1$),
- $e_k(M)$ is the $k$-th lowest error to the model $M$ among all $n$ correspondences,
- $\alpha_0$ is the probability of a random correspondence having error 1 pixel,
- $d$ is the error dimension: 1 for point-to-line distance, 2 for point-to-point.

$\alpha_0$ is independent on the tested model $M$, being the probability of a random correspondence under *background model* distribution having error 1 pixel: e.g., ratio of area of band of radius 1 and of the area of image for point-to-line distance. The term $e_k(M)^d\alpha_0$ is the probability of a random correspondence having error at most $e_k(M)$. The last factor in the formula is thus the probability of $k - N_{\text{sample}}$ correspondences having error at most $e_k(M)$. The other factors represent a number of tests. In other words, this is an expectation of false alarms for model $M$ having $k$ inliers under the null hypothesis. Model $M$ is considered as valid if

$$NFA(M) = \min_{k=N_{\text{sample}}+1\ldots n} NFA(M,k) \leq \epsilon\,. \tag{2}$$

The only parameter is $\epsilon$. It is usually set to 1, and the inlier/outlier error threshold for model $M$ is $e_k$, with $k$ minimizing (2).

AC model estimation requires finding $\arg\min_M NFA(M)$ among all models $M$ computed from all possible $N_{\text{sample}}$ correspondences. For a given $M$, the complexity of computing $NFA(M)$ is $O(n \log n)$ since it requires sorting the errors $e_k(M)$ of all $n$ correspondences. However, the number of possible models is $N_{\text{out}} \binom{n}{N_{\text{sample}}}$, which becomes exceedingly large as soon as $N_{\text{sample}} > 2$, hence the random model sampling tests of RANSAC.

Minimizing the NFA instead of maximizing the inlier count (if an inlier/outlier threshold $T$ is given) or minimizing the median of errors (in the least median of squares variant) is the task of AC-RANSAC. The definite advantage over standard RANSAC is that the precision $e_k(M)$, that replaces $T$, adapts to the data. In our experiments, we let AC-RANSAC [25] set the threshold without any additional constraint. More precisely, we only impose that the returned model provides at least $2N_{\text{sample}}$ inliers.

### 4.2 Rigidity measures for robust Structure from Motion models

The robust model estimations that are required to define an incremental 3D reconstruction pipeline are the fundamental matrix, homography, essential matrix and pose estimations (see Section 2). Each kind of model has its own definition of rigidity. To devise the *a contrario* rigid model estimation algorithm for these cases, we need to determine the values of $\alpha_0$, $d$, $N_{\text{sample}}$ and $N_{\text{out}}$ assuming a uniform distribution of correspondences. Two main groups of measures are needed: "point to point" and "point to line" distances.

| Model | Fundamental | | Homography | Essential | | Pose estimation | |
|---|---|---|---|---|---|---|---|
| $N_{\text{sample}}$ | 7 | 8 | 4 | 5 (see [26]) | 8 | 4 + K (see [27]) | 6 |
| $N_{\text{out}}$ | 3 | 1 | 1 | 10 | 1 | 1 | 1 |

**Table 1.** Number of samples and number of models for the model estimators.

**Point to point distance.** For homography and camera pose estimation:
- $\alpha_0 = \frac{\pi}{A}$: it is the ratio of the radius 1 disk area to image area $A$.
- $d = 2$: the disk area grows quadratically with its radius.

**Point to line distance.** For essential and fundamental matrix estimation:
- $\alpha_0 = \frac{2D}{A}$: considering a band of "radius" 1 around an image line, whose length cannot exceed the image diameter $D$, $\alpha_0$ is the upper bound of the ratio of areas of such a band to area of the image. Notice this is only an upper bound used for faster computation, which may be more selective than strictly necessary. The actual $\alpha_0$ should depend on the considered line.
- $d = 1$: the band area grows linearly with the distance to the line.

$N_{\text{out}}$ is the maximum number of models that can be computed for a set of $N_{\text{sample}}$ correspondences. It depends on the estimation procedure. The values of $N_{\text{out}}$ are listed in Table 1. Note that $N_{\text{out}}$ may also depend on the actual sample: e.g., computing a fundamental matrix with the 7-point algorithm requires finding roots of a third degree polynomial, which can have 1 or 3 solutions; similarly, the 5-point algorithm for the essential matrix solver involves finding real roots of a 10-degree polynomial. In such a case, we consider the maximum possible number of algorithm outcomes to get an upper bound of the NFA.

AC estimation of a fundamental matrix and of a homography have been described before [24, 23]. In the case of homography estimation, we additionally pick inliers among those that where previously selected for the fundamental matrix estimation. Our AC estimation of the essential matrix and of the pose is original. Note that our pose estimation involves a single image domain instead of two in the other formulations.

## 5 An *a contrario,* incremental Structure from Motion

Robust model estimation in incremental SfM is traditionally implemented using RANSAC and controlled via globally-fixed thresholds, which has the above-mentioned drawbacks (cf. Section 3). Bundler, for instance, uses as default parameters a 9-pixel reprojection threshold for the fundamental matrix estimation, 6 pixels for homography and 4 pixels for pose. These are heuristic choices that yield decent results in many datasets but cannot adapt to all situations.

Using the *a contrario* approach, we have adaptive thresholds for all components of a SfM pipeline that require a robust model estimation (cf. Section 4). Our *a contrario* 3D reconstruction pipeline is separated in two $AC$ blocks: first, the computation of feature correspondences, and second, the SfM process itself.

***A contrario* correspondence** For the computation of fundamental matrices and homographies, we replace estimations by RANSAC with AC-RANSAC. For homography, we additionally pick inliers among those that where previously selected for the fundamental matrix estimation, which reduces the search space. This yields the statistically most consistent set of matches between feature sets as well as a computed threshold for the model found. As we shall see, it selects more stable matches for the camera pose estimation.

***A contrario* camera pose estimation** For pose estimation, which may need the matrix of instrinsics, we also replace RANSAC with AC-RANSAC. The computed threshold of the resection is particularly valuable because it provides a confidence estimate on the current view that is used as a threshold for outlier rejection of new possible triangulated tracks. Each newly triangulated point yielding a larger reprojection error is discarded.

Our reconstruction pipeline is not bound by the usage of a static threshold $T_m$ per kind of model $m$. It provides adaptive thresholds $T_{m,i}$ for each computed model, i.e., for each kind of model and for each model of a given kind to estimate, given corresponding data (typically, a pair of images).

## 6 Experiments

We have implemented an *a contrario,* incremental SfM system, as described in Section 5. Our reconstruction pipeline is entirely written in high level C++, with flexible template modules. In particular, we use a generic AC-RANSAC implementation [25] and new model solvers only need to be warped into a given structure. We plan to open source our system to make available an easy to read/use/modify platform for SfM. Unit tests have been designed for each computer vision building block, that also illustrate how to use the various modules.

In the following, we mainly compare our system with Bundler [13], a popular and efficient system that is open source and fairly easy to use. For comparison, our code has 8,000 lines of code while Bundler has 20,000. AC-RANSAC results for specific kinds of models are also illustrated by comparison to RANSAC only.

To evaluate our approach, we have experimented with datasets where ground truth is available. We have used the datasets of Strecha *et al.* [28], with laser ground truth, dataset 004 of [8] with calibration ground truth and 2 additional synthetic generated dataset.

### 6.1 Threshold variation for fundamental matrix estimation

To assess the interest of adaptive thresholds for feature correspondence estimation, we have estimated fundamental matrices on [28] and measured the average baseline error of the SfM reconstruction, over all views of the dataset, for various values of the corresponding threshold $T_F$. Results are shown in Table 2.

Note that the rank-1 threshold value varies depending on the dataset, meaning there is no ideal static threshold that leads to the best results in the Bundler

| Scene | | Bundler $T_F$ threshold | | | | | AC-SfM $T_F$ threshold | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 3 | 6 | 9 | 12 | auto | min | med | max |
| FountainP11 | error | 0.002 | 0.003 | 0.003 | 0.004 | 0.005 | **0.001** | | | |
| | ranking | **1** | 3 | 2 | 4 | 5 | | 0.57 | 1.00 | 10.5 |
| HerzJesusP8 | error | 0.004 | 0.003 | 0.003 | 0.007 | 0.003 | **0.001** | | | |
| | ranking | 4 | **1** | 3 | 5 | 2 | | 0.63 | 1.88 | 5.26 |
| HerzJesusP25 | error | 0.004 | 0.010 | 0.005 | **0.004** | 0.004 | 0.005 | | | |
| | ranking | 3 | 5 | 4 | **1** | 2 | | 0.23 | 1.53 | 82.8 |
| CastleP19 | error | 8.22 | 0.029 | 0.032 | 0.039 | X | **0.015** | | | |
| | ranking | 4 | **1** | 2 | 3 | X | | 0.69 | 0.91 | 15.7 |
| CastleP30 | error | 0.055 | 0.057 | 0.043 | 0.042 | 0.045 | **0.011** | | | |
| | ranking | 4 | 5 | 2 | **1** | 3 | | 0.55 | 0.92 | 284 |

**Table 2.** Fundamental matrix threshold consequence over reconstruction: average error (in meters) w.r.t. ground truth (over all views of the dataset). For Bundler: average baseline error and corresponding rank, depending on threshold values. Best in **bold**. X denotes a failed calibration, one of the views being rejected by Bundler. For AC-SfM: average baseline error and distribution of the computed threshold values.

chain. This is confirmed by the distribution of the computed AC threshold (over all views of the dataset): there is no stable median value and extreme values (min and max) greatly vary. The average AC baseline error is significantly lower than with the best static threshold in most cases. The error is however almost equal for HerzJesusP25. The reason is that some false matches and bad estimates can still occur in the AC-RANSAC case.

### 6.2 Camera pose estimation

To evaluate camera pose estimation, two views are first used for building a 3D point cloud, then the other images are compared to that point cloud to estimate their pose. The results are displayed in Fig. 4.

The default RANSAC threshold $T = 4$ of Bundler fails for images 0, 2, 10-13 and 23-24, because not enough correspondences with that precision are found. For the other images, the error with respect to the ground truth is worse in baseline and in angle than for the much larger threshold $T = 12$. The AC-RANSAC adaptive thresholds provides errors that are similar to $T = 12$ RANSAC. A closer study shows that the $T = 4$ selection incorporates outliers and thus yields a less accurate result. Naturally, these outliers are also present for $T = 12$, but an averaging effect happens to produce a good accuracy. Still, the *a contrario* pose estimation is more discriminative, strongly adjusting to the context, and yields an accuracy comparable to $T = 12$ with slightly fewer correspondences. No system could register camera 13 though, because of a lack of overlap with the initial pair.

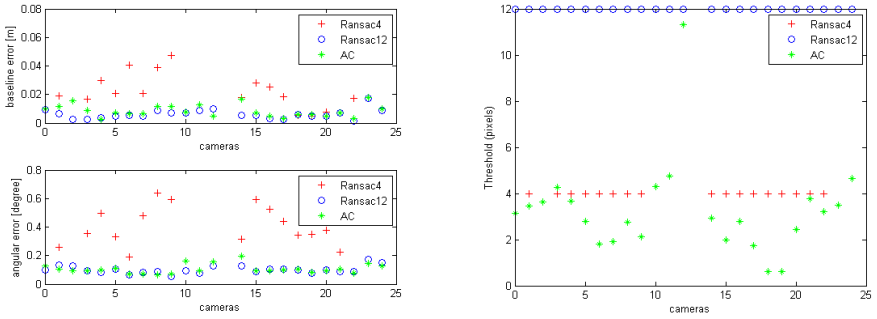



**Fig. 4.** Evaluation of camera pose estimation with RANSAC and AC-RANSAC on HerzJesusP25. Two views are first registered with a fundamental matrix, yielding a 3D point cloud. Then the pose of all other images is computed relatively to that point cloud. Thresholds 4 (Bundler's default) and 12 are tested. On the left is displayed the pose error relative to ground truth (baseline, angle), on the right the used thresholds.

### 6.3 Structure from Motion accuracy comparison

Finally, we have evaluated the whole AC-SfM pipeline, comparing it to Bundler. The same inputs are considered, i.e., SIFT keypoints and a maximum ratio of 0.6 for the best to second best descriptor matches. For this evaluation, we used both real [28, 8] and synthetic datasets. The quality of the reconstructions is evaluated on camera external position and rotation with a 7-degree of freedom rigid transform registration (scale, translation, rotation) [29]. The rigid transform is used to preserve angles and distance ratios. Results are shown in Fig. 5.

The angular error is significantly lower in our pipeline for all datasets while the baseline error is comparable to that of Bundler, although most of the time better and more equally distributed among the views. The exception is Set0004-1200x1600, for which 60% of the baseline errors of the camera are more precise, but all the angular error are better. Figure 6 presents a comparative evaluation on synthetic datasets. *A contrario* SfM gives again significantly more accurate results, thanks to its adaptive thresholds.

It can be noted that Bundler, being considered state of the art, already performs quite well, rotation errors being below a fraction of degree. But there is still room for improvement for applications requiring high-precision, which is what we aim at. In fact, these experiments show that AC consistently yields a better precision (up to factor 10). This provides substantial benefits for 3D reconstruction: a $0.2°$ difference in a ray direction at $10\,m$ distance (typical in most experiments of Figure 5) yields an arc length of $(0.2/180\times\pi)*10m = 3.5\,cm$, whereas for such scenes we would like to achieve a $1\,cm$ precision.

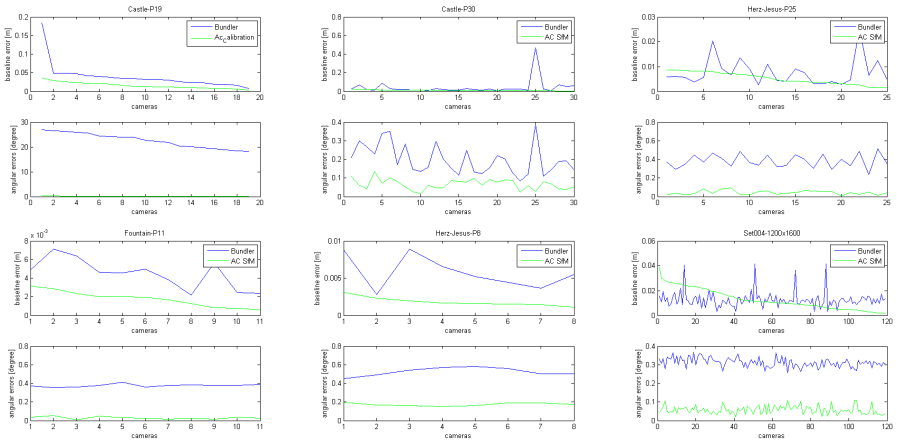



**Fig. 5.** Evaluation of camera calibration with Bundler and *a contrario* SfM on [28], and scene 004 from [8]. To facilitate the graph reading, the error measures (baseline and angle) for the different views are sorted in decreasing order of *a contrario* SfM baseline error. The angular error for Castle-P19, averaging 25° for Bundler, could not be explained and should not be considered in this comparison. On the whole AC SfM is significantly more accurate than Bundler, and has a more equal distribution of errors.

## 7 Conclusion

We have argued the interest of model estimators with fine-grain, adaptive thresholds and described how to automatically perform such model estimations for SfM within the framework of the *a contrario* theory. We have presented a practical 3D reconstruction pipeline that implements these AC estimators and we have shown that our threshold-free system can select inliers with a better discrimination than classical RANSAC, yielding better reconstructions and poses.

Our original contribution includes the *a contrario* threshold definition for the estimation of the essential matrix and for resection. It can be noted that pose estimation involves here a single image, contrary to symmetrical errors used in other parameterizations. Also original is the use, when estimating a homography, of inliers that were selected for the fundamental matrix estimation. Finally, we have systematized AC estimation in a concrete SfM pipeline and showed that it often outperforms state-of-the-art systems. Prior work had indicated feasibility for some components, not efficiency for a complete system and large-scale data.

A few fixed parameters remain, but we believe some of them can be removed too. In particular, we think that the commonly used SIFT descriptor distance ratio for feature matching can be replaced by an *a contrario* descriptor matching [30]. There is also encouraging work on *a contrario* disparity map estimation [31]. This opens the way to a robust, parameter-free, multiple-view stereovision process computing dense point clouds and 3D meshes.

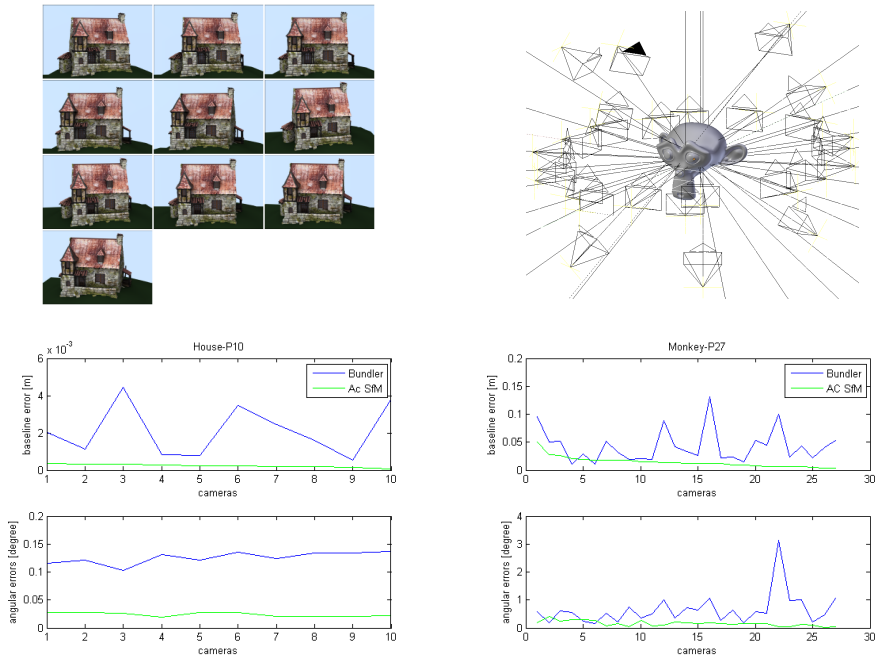



**Fig. 6.** Camera calibration with Bundler and *a contrario* SfM on synthetic datasets.

## Acknowledgments

This work was carried out in IMAGINE, a joint research project between École des Ponts ParisTech (ENPC) and the Scientific and Technical Centre for Building (CSTB) and supported by Mikros Image and Agence Nationale de la Recherche ANR-09-CORD-003 (Callisto project). The original publication is available at springerlink.com.

## References

1. Agarwal, S., Snavely, N., Simon, I., Seitz, S., Szeliski, R.: Building Rome in a day. In: 12th IEEE International Conference on Computer Vision (ICCV). (2009) 72–79
2. Frahm, J., Fite-Georgel, P., Gallup, D., Johnson, T., Raguram, R., Wu, C., Jen, Y., Dunn, E., Clipp, B., Lazebnik, S., et al.: Building Rome on a cloudless day. In: European Conference on Computer Vision (ECCV), Springer (2010) 368–381
3. Kahl, F.: Multiple view geometry and the $L_\infty$-norm. In: ICCV. (2005) 1002–1009
4. Dalalyan, A., Keriven, R.: $L_1$-penalized robust estimation for a class of inverse problems arising in multiview geometry. In: NIPS. (2009) 441–449
5. Furukawa, Y., Ponce, J.: Accurate, dense, and robust multiview stereopsis. IEEE Transactions on Pattern Analysis and Machine Intelligence **32** (2010) 1362–1376
6. Hiep, V., Keriven, R., Labatut, P., Pons, J.: Towards high-resolution large-scale multi-view stereo. In: CVPR. (2009) 1430–1437

7. Zach, C., Klopschitz, M., Pollefeys, M.: Disambiguating visual relations using loop constraints. In: CVPR. (2010) 1426–1433
8. Aanæs, H., Dahl, A., Steenstrup Pedersen, K.: Interesting interest points. International Journal of Computer Vision **97** (2012) 18–35
9. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision (IJCV) **60** (2004) 91–110
10. Bay, H., Tuytelaars, T., Van Gool, L., Van Gool, L.: SURF: Speeded Up Robust Features. European Conference in Computer Vision (ECCV) **3951** (2006) 404–417
11. Lourakis, M.I.A., Argyros, A.A.: SBA: A software package for generic sparse bundle adjustment. ACM Transactions on Mathematical Software (TOMS) **36** (2009)
12. Wu, C., Agarwal, S., Curless, B., Seitz, S.M.: Multicore bundle adjustment. In: CVPR. (2011) 3057–3064
13. Snavely, N., Seitz, S.M., Szeliski, R.: Photo tourism: exploring photo collections in 3D. ACM Transactions on Graphics (TOG) **25** (2006) 835–846
14. Gherardi, R., Farenzena, M., Fusiello, A.: Improving the efficiency of hierarchical structure-and-motion. In: CVPR. (2010) 1594–1600
15. Havlena, M., Torii, A., Knopp, J., Pajdla, T.: Randomized structure from motion based on atomic 3D models from camera triplets. In: CVPR. (2009) 2874–2881
16. Scaramuzza, D., Fraundorfer, F.: Visual odometry: Part I - the first 30 years and fundamentals. IEEE Robot. Automat. Mag. **18** (2011)
17. Fraundorfer, F., Scaramuzza, D.: Visual odometry: Part II - matching, robustness, and applications. IEEE Robot. Automat. Mag. **19** (2012)
18. Martinec, D., Pajdla, T.: Robust rotation and translation estimation in multiview reconstruction. In: CVPR. (2007)
19. Govindu, V.M.: Combining two-view constraints for motion estimation. In: CVPR. Volume 2. (2001) II.218–225
20. Govindu, V.M.: Robustness in motion averaging. In: ACCV. (2006) 457–466
21. Fischler, M.A., Bolles, R.C.: Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM (CACM) **24** (1981) 381–395
22. Desolneux, A., Moisan, L., Morel, J.M.: From Gestalt theory to image analysis: a probabilistic approach. 1st edn. Springer (2007)
23. Rabin, J., Delon, J., Gousseau, Y., Moisan, L.: MAC-RANSAC: a robust algorithm for the recognition of multiple objects. In: Proc. of 3DPTV 2010, Paris (2010)
24. Moisan, L., Stival, B.: A probabilistic criterion to detect rigid point matches between two images and estimate the fundamental matrix. Int. J. of Computer Vision (IJCV) **57** (2004) 201–218
25. Moisan, L., Moulon, P., Monasse, P.: Automatic homographic registration of a pair of images, with a contrario elimination of outliers. Image Processing On Line (2012) http://dx.doi.org/10.5201/ipol.2012.mmm-oh.
26. Nistér, D.: An efficient solution to the five-point relative pose problem. In: CVPR. Volume 2. (2003) II.195–202
27. Lepetit, V., Moreno-Noguer, F., Fua, P.: EP$n$P: an accurate $O(n)$ solution to the P$n$P problem. International Journal of Computer Vision (IJCV) **81** (2009) 155–166
28. Strecha, C., von Hansen, W., Van Gool, L.J., Fua, P., Thoennessen, U.: On benchmarking camera calibration and multi-view stereo for high resolution imagery. In: CVPR. (2008) 1–8
29. Haralick, R.M., Shapiro, L.G.: Computer and Robot Vision. 1st edn. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA (1992)
30. Rabin, J., Delon, J., Gousseau, Y.: A statistical approach to the matching of local features. SIAM J. Imaging Sciences **2** (2009) 931–958

31. Sabater, N., Almansa, A., Morel, J.M.: Meaningful matches in stereovision. IEEE Transactions on Pattern Analysis and Machine Intelligence **99** (2011)