Optimal Randomized RANSAC

Ondřej Chum, *Member*, *IEEE*, and Jiří Matas, *Member*, *IEEE*

Abstract—A randomized model verification strategy for RANSAC is presented. The proposed method finds, like RANSAC, a solution that is optimal with user-specified probability. The solution is found in time that is close to the shortest possible and superior to any deterministic verification strategy. A provably fastest model verification strategy is designed for the (theoretical) situation when the contamination of data by outliers is known. In this case, the algorithm is the fastest possible (on the average) of all randomized RANSAC algorithms guaranteeing a confidence in the solution. The derivation of the optimality property is based on Wald's theory of sequential decision making, in particular, a modified sequential probability ratio test (SPRT). Next, the R-RANSAC with SPRT algorithm is introduced. The algorithm removes the requirement for a priori knowledge of the fraction of outliers and estimates the quantity online. We show experimentally that on standard test data, the method has performance close to the theoretically optimal and is 2 to 10 times faster than standard RANSAC and is up to four times faster than previously published methods.

Index Terms—RANSAC, randomized RANSAC.

1 Introduction

The **RAN**dom **SA**mple **C**onsensus (RANSAC) algorithm introduced by Fishler and Bolles in 1981 [6] is a widely used robust estimator that has become a standard in the field of computer vision [1]. RANSAC and related hypothesize-and-verify methods [4], [11], [12], [13], [14], [15] have been applied to many vision problems.

The RANSAC algorithm proceeds as follows: Repeatedly, subsets of the input data (for example, a set of tentative correspondences) are randomly selected (with replacement), and model parameters fitting these subsets are computed. In the second step, the quality of the parameters is evaluated on the input data. Different cost functions have been proposed [15], the standard being the number of data points (inliers) consistent with the model. The process is terminated when the probability of finding a better model becomes lower than a user-specified probability η_0. The $1-\eta_0$ confidence in the solution holds for all levels of contamination of the input data, that is, for any number of outliers within the input data.

The speed of standard RANSAC depends on two factors: The number of random samples and the number N of the input data points. In all common settings where RANSAC is applied, almost all models whose quality is verified are incorrect with arbitrary parameters originating from contaminated samples. Such models are consistent with only a small number of the data points. In [9], Matas and Chum showed how this property could be exploited to increase the speed of RANSAC. The algorithm, called R-RANSAC, reduces the time needed for the model evaluation step by introducing a two-stage procedure. First, a statistical test is performed on d randomly selected data points ($d \ll N$). Evaluation of the remaining $N-d$ data points is carried out only if the first d data points are inliers. The speed up of R-RANSAC depends on the probabilities of the two types of errors committed in the pretest, the rejection of an uncontaminated model and the acceptance of a contaminated model.

This idea was modified by Nistér to include competitive verification of models. The algorithm performed impressively in a real-time structure from motion system [12]. The main limitation of Nistér's preemptive RANSAC is that a fixed number of models are evaluated, which is equivalent to an a priori assumption that a lower bound on the fraction of inliers is known. This limits the applicability of preemptive RANSAC in problems where the fraction of inliers varies widely, such as wide baseline stereo.

As noted in [9], the two-stage procedure of R-RANSAC is not optimal. *As a main contribution of this paper, we define an optimal hypothesis evaluation procedure*, that is, a method for randomized model quality evaluation that returns, in the fastest average time possible, a solution with the confidence $1-\eta_0$. The derivation of the optimality property is based on Wald's theory of sequential decision making [16].

In Section 2, the concept of randomization of the verification step is described in detail. Section 3 reviews previous work relevant to the topic. In Section 4, we introduce the relevant parts of Wald's decision theory and show how its results can be brought to bear on the problem of minimizing RANSAC runtime. The RANSAC with SPRT algorithm [10] is described in detail in Section 5. In Section 6, the theoretical results are experimentally verified on standard stereo matching problems. The paper is concluded in Section 8.

2 Randomized Consensus

The speed of RANSAC depends on two factors. First, the percentage of outliers determines the number of random samples needed to guarantee a given confidence in the optimality of the solution. Second, the time needed to assess

- *O. Chum is with Czech Technical University, Faculty of Electrical Engineering, Department of Cybernetics, Karlovo námestí 13, 121 35 Prague, Czech Republic and the Visual Geometry Group, Department of Electrical Engineering, University of Oxford. E-mail: chum@cmp.felk.cvut.cz.*
- *J. Matas is with Czech Technical University, Faculty of Electrical Engineering, Department of Cybernetics, Karlovo námestí 13, 121 35 Prague, Czech Republic. E-mail: matas@cmp.felk.cvut.cz.*

Manuscript received 11 Dec. 2006; revised 9 Aug. 2007; accepted 13 Aug. 2007; published online 25 September 2007.
Recommended for acceptance by P. Torr.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-0862-1206.
Digital Object Identifier no. 10.1109/TPAMI.2007.70787.

the quality of a hypothesized model parameters is proportional to the number N of input data points. The total runtime t of RANSAC can be expressed as

$$t = k(\mathrm{t}_M + \overline{m}_S\, t_V), \tag{1}$$

where k is the number of samples drawn, t_M is time needed to instantiate a model hypotheses given a sample, $\overline{m}_S$ is the average number of models per sample, and t_V is the average time needed to evaluate the quality of the sample. We choose the time needed to verify a single correspondence as the unit of time for t_M, t_V, and t. Note that in standard RANSAC, $t_V = N$.

The core idea of the Randomized (hypothesis evaluation) RANSAC is that most evaluated model hypotheses are influenced by outliers. To reject such erroneous models, it is sufficient to perform a statistical test on only a small number of data points. The test can be formulated as follows: The hypothesis generation step proposes a model. It is either "good," that is, it is uncontaminated by outliers and leads to the optimal solution (the solution with maximal support), or it is "bad" (or contaminated) with at least one of the data points in the sample is an outlier.[1] The property "good" is a hidden state that is not directly observable but is statistically linked to observable events. The observable events are "data point (correspondence) is or is-not consistent with the model."

The statistical test has two effects on RANSAC behavior: It 1) reduces the number of verified data points (and thus time complexity of the verification step) and 2) introduces the possibility of rejecting (overlooking) a good sample. The probability α of rejecting a good sample is the significance of the test, and it increases the number of samples drawn before the $1-\eta_0$ confidence is ensured. The correct model parameters are recovered if an uncontaminated sample is drawn and passes the test. This happens with probability

$$P = P_g(1-\alpha),$$

where P_g is a probability of a "good" sample being drawn.

The problem is to find a test that balances the number of correspondences needed for model verification and the increase in the number of samples induced by false rejections so that the total runtime t in (1) is minimized. It was shown in [9] that $k_{\eta_0} \approx -\ln \eta_0 \bar{k}$, where $\bar{k}$ is the average number of samples before a first uncontaminated sample is drawn and k_{η_0} the minimum number of steps needed to guarantee $1-\eta_0$ confidence in the solution. Hence, the average time to find an all-inlier sample and the time to stop with the confidence differ only by a multiplicative factor; minimizing either leads to identical results. The formulas below are derived for $k = \bar{k}$. Since the average time to draw an uncontaminated model that passes the test is $\bar{k} = 1/(P_g(1-\alpha))$, we have

$$t = \frac{1}{P_g(1-\alpha)}(t_M + \overline{m}_S\, t_V). \tag{2}$$

The design of a statistical verification test depends on two probabilities ε and δ, where ε denotes the fraction of inliers within the set of data points, and δ is a probability that a data point is consistent with a model with arbitrary parameters. These probabilities are typically unknown beforehand and have to be either estimated during the course of the algorithm, or the test must be efficient for a very large range of their values.

3 Previous Work

The idea of randomized verification of hypotheses in RANSAC appeared only recently [3], and the related literature is thus limited. We review the original R-RANSAC that relied on the $T_{(d,d)}$ test [3], [9] and the two algorithms it inspired: the preemtive RANSAC [12] and the RANSAC with a bail-out test [2].

3.1 The $T_{d,d}$ Test

R-RANSAC with the $T_{d,d}$ test [3], [9] employs a simple and thus mathematically tractable class of preverification tests, which were defined as follows: The $T_{d,d}$ is passed if all d data points out of d randomly selected are consistent with the hypothesized model.

In the $T_{d,d}$ test, the number of verified correspondences per a test depends on d and is equal to that in [3]:

$$t_V(d) = P_g((1-\alpha)N + \alpha\bar{t}_\alpha) + (1-P_g)\big(\beta N + (1-\beta)\bar{t}_\beta\big). \tag{3}$$

Here, β stands for the probability that a bad sample passes the preverification test. Note that it is important that $\beta \ll 1-\alpha$, so that a bad (contaminated) sample is consistent with a smaller number of data points than a good sample. In [3], the optimal value for d is derived in detail. The constants introduced in the previous section are expressed as

$$\alpha = 1-\varepsilon^d \quad \text{and} \quad \beta = \delta^d,$$

where δ is the probability that a data point is consistent with an arbitrary model.

For a broad range of values of ε and δ, the optimal value of d is $d = 1$. Therefore, without any prior knowledge of ε and δ, the suggested test (from the $T_{d,d}$ class of tests) is $T_{1,1}$.

Note that instead of drawing m data points and verifying only d data points in the test, the following approach can be adopted in the $T_{d,d}$ test. Draw $m+d$ data points randomly, and fit model parameters to the sample. Measure the error on the $m+d$ data points in the sample. If the error is smaller than a predefined threshold Δ, that is, there is a model that fits all the data well, proceed with the verification step; otherwise, generate a new hypothesis. The advantage of such an approach is that a single model can be obtained from $m+d$ data points (in contrary to $\overline{m}_S$ models from m data points only), and also a model of higher accuracy is typically obtained if more points are used. This is specific to the $T_{d,d}$ class of tests, where the consistency of all d data points is required.

1. In the definition of "good" and "bad" models, we ignored the fact that for certain models, for example, the fundamental matrix, the optimal model can be obtained from a sample contaminated by an outlier that is by coincidence consistent with the correct epipolar geometry. We also ignored the fact that an uncontaminated, all-inlier sample may not lead to a "good" model due to noise and poor conditioning of the estimation process. The problem can be overcome with so-called local optimization, for details, see [4]. Neither simplification has an impact on the derivations below.

Despite the simplicity of $T_{1,1}$ test, experiments have shown that preverification can reduce RANSAC runtime significantly [9].

3.2 The Bail-Out Test

The bail-out test (Algorithm 1) introduced by Capel [2] exploits the following idea. Whenever the probability of the event "the currently evaluated hypothesis contains as many inliers as the so-far-the-best hypothesis" falls under a fixed threshold P_{cf} (1 percent), the evaluation is terminated, and the hypothesis is discarded as incorrect. Since the exact probability

$$P_{\mathrm{cf}} = P(I > I^*) = \sum_{I=I^*}^{N} P(I|I_n, n, N)$$

is computationally expensive to evaluate, an approximation is suggested. The number I_n of data points consistent with a hypothesis within n tested data points follows a hypergeometric distribution

$$I_n \sim \mathrm{HypG}().$$

The binomial distribution gives a lower bound on the hypergeometric distribution and can be approximated by a normal distribution:

$$I_n \sim \mathrm{N}(\mu, \sigma^2) = \mathrm{N}\left(n\varepsilon_i, n\varepsilon_i(1-\varepsilon_i)\left(\frac{N-n}{N-1}\right)\right). \quad (4)$$

The size of minimal support $I_n^{\min}$ out of n verified correspondences for a hypothesis to proceed to further verification is expressed as $I_n^{\min} = \lfloor n\varepsilon_i - z_c\sigma \rfloor$, where z_c is the quantile associated with the confidence P_{cf} for a zero mean unit variance normal distribution.

How often is a "good" hypothesis rejected using such a test? In [2], this possibility is ignored, and the probability α of this event is neglected. If the test on $I_n^{\min}$ was applied only once as in the $T_{d,d}$, then $\alpha = P_{\mathrm{cf}}$. However, the test is applied after each verified data point, so P_{cf} is clearly only a lower bound on α. The exact evaluation of α for this type of test is intractable, since the results of the individual tests are not independent, which means that it is not possible to devise an optimal test based on $I_n^{\min}$.[2]

Algorithm 1: The bail-out test.
Set $n = 1$ and $I_n = 0$.

Although $n \leq N$
  if $\mathbf{x}_n$ is inlier $I_n{+}{+}$
  if $I_n < I_n^{\min}$ abort evaluation
  $n{+}{+}$
end
Return I_N

3.3 Preemptive RANSAC

The aim of preemptive RANSAC by Nistér [12] is to efficiently select the best hypothesis from a fixed number of generated hypotheses. The algorithm was designed for a real-time structure from motion applications, where it is necessary to find a result within a scheduled time. The fixed number of hypotheses restricts the level of outlier contamination, since there is a one-to-one relationship between the number of necessary steps (sample draws) and the level of data contamination. When the fraction of inliers is too low for the prespecified number of hypothesis, the method simply fails.

Preemptive RANSAC is based on comparing hypotheses after evaluating just a fraction of the data points and rejecting those with worse scores. The algorithm uses a nonincreasing function $t(i)$, $i \in 0 \ldots N$, that determines how many top ranked hypotheses are further considered (whereas the other are rejected) after the cost function was evaluated on i data points. The data points to be verified are selected at random with replacement. The algorithm proceeds as follows: First, $k = t(0)$ hypotheses are generated. Then, one data point is chosen for each active hypothesis, and the cost function is updated using that data point. After i data points have been verified for each hypothesis, the $t(i)$ best hypotheses are kept, and the others are rejected. The procedure stops when only a single hypothesis is kept, that is, $t(i) = 1$, or after verifying all N data points for each surviving hypothesis.

Algorithm 2: The structure of preemptive RANSAC.
1 Generate all $k = t(0)$ hypotheses
2 For $i = 1$ to N
  - evaluate the cost function for randomly chosen (with replacement) data point for all $t(i-1)$ hypotheses left
  - keep $t(i)$ hypotheses with the highest value of the cost function over i data points.

Note, there is a significant difference between the *objectives* of standard (randomized) RANSAC and preemptive RANSAC. The former attempts to find, with predefined confidence, the best of *all* model parameters in the shortest possible time; the latter seeks the best model within a set of hypotheses of *fixed size*. The preemptive RANSAC formulation is natural in a time-constrained setting. If a good solution is not among the fixed number of models that can be evaluated in the time available, it is simply out of reach. For low-contamination problems where only a few samples are needed to find a solution, preemptive RANSAC can be much slower than standard RANSAC. However, it is well known that not all inlier solutions are equally precise. Preemptive RANSAC in this case spends remaining available time finding the best model in the maximum likelihood sense performing effectively preemptive MLESAC [15]. Preemptive RANSAC has an almost constant runtime, a desirable property in online systems, and it puts fewer constraints on the selection of the cost function at the price of not being able to handle problems with different proportion of inliers. In short, standard and preemptive RANSAC solve different but related problems and are thus applicable in different contexts.

4 The Optimal Sequential Test

In many applications, such as industrial inspection, the problem of sequential testing is to decide whether a model—typically a batch of products—is "good" or "bad" by making the smallest possible number of observations while satisfying the predefined bounds on the probabilities of the two possible errors—accepting a "bad" model as "good," and *vice versa*. Wald proposed the sequential probability ratio test (SPRT) and showed [16] that, given

2. In our simulations, $\alpha \approx 0.014$ for $P_{\mathrm{cf}} = 0.01$.

bounds on the errors of the first and second kind, it minimizes the number of observations (time to decision).[3]

Wald's SPRT test is a solution of a *constrained optimization* problem. The user supplies the acceptable probabilities of the errors of the first and the second kind, and the resulting optimal test is a trade-off between the time to decision (or cost of observations) and the errors committed.

However, when evaluating RANSAC, the situation is different. First of all, a "good" model is always evaluated for all data points (correspondences) since the number of inliers is one of the outputs of the algorithms. Therefore, the only error that can be committed is an early rejection of a "good" model (error of the first kind). However, this only means that more samples have to be drawn to achieve the required confidence $1-\eta_0$ of finding the optimal solution. Therefore, unlike in the classical setting, we are solving a *global optimization* problem, minimizing a single real number—the time to decision—since the consequence of an error is also a loss of time.

The model evaluation step of the optimal R-RANSAC proceeds as Wald's SPRT with the probability α of rejecting a "good" sample set to achieve maximum speed of the whole RANSAC process. To understand the operation of R-RANSAC with SPRT, some familiarity with Wald's decision theory is required. We therefore introduce its relevant parts. Some of the results are presented in a form that is not fully general but sufficient for the derivation of the R-RANSAC with SPRT algorithm. Some of Wald's terminology is modified in order to make the exposition more accessible.

In the model evaluation step, our objective is to decide between the hypothesis H_g that the model is "good," and the alternative hypothesis H_b that the model is "bad." A "good" model is computed from an all-inlier sample. The Wald's SPRT is based on the likelihood ratio [16]

$$\lambda_j = \prod_{r=1}^{j} \frac{p(x_r|H_b)}{p(x_r|H_g)} = \lambda_{j-1} \cdot \frac{p(x_j|H_b)}{p(x_j|H_g)}, \tag{5}$$

a ratio of two conditional probabilities of the observation x_r under the assumptions of H_g and H_b, respectively. In RANSAC, x_r is equal to 1 if the rth data point is consistent with a model with parameters θ and 0 otherwise. For example, a correspondence is consistent with (that is, supporting) an epipolar geometry represented by a fundamental matrix F if its Sampson's error is smaller than some predefined threshold [7]. The probability $p(1|H_g)$ that any randomly chosen data point is consistent with a "good" model is approximated by the fraction of inliers ε among the data points.[4] The probability of a data point being consistent with a "bad" model is modeled as a probability of a random event with Bernoulli distribution with parameter $\delta : p(1|H_b) = \delta$. The process of estimation of δ and ε is discussed in Section 5.

3. Precisely speaking, the SPRT is only approximately optimal. However, the approximation has been shown by Wald to be so close to the optimum that, for practical purposes, it is considered the optimal test.

4. The probability ε would be exact if the data points were selected with replacement. Since the objective of the verification is to count the size of the support of the model, the correspondences are drawn without replacement. However, the approximation is close.

Algorithm 3: The adapted SPRT.

Output: model accepted/rejected, number of tested data points j, a fraction of data points consistent with the model

Set $j = 1$

1 Check whether jth data point is consistent with the model
2 Compute the likelihood ratio λ_j (5)
3 If $\lambda_j > A$, decide the model is "bad" (model "rejected"), else increment j
4 If $j > N$, where N is the number of correspondences, decide model "accepted" else go to Step 1.

After each observation the standard Wald's SPRT makes one of three decisions: Accept a "good" model, reject a "bad" model, or continue testing. Since in RANSAC, the total number of inliers is needed to decide on termination, nothing is gained by an early decision in favor of a "good" model. Therefore, the option of an early acceptance of the model has been removed in the Adapted SPRT (Algorithm 3). The full SPRT is described by Wald [16] and, in a more accessible form, by Lee [8].

4.1 The Optimal Value of the Decision Threshold

The decision threshold A is the only parameter of the Adapted SPRT. We show that it can be set to achieve optimal performance with minimal average RANSAC runtime given the probabilities δ and ε. We use the following theorems (for proofs, see [16]).

Theorem 1. *The probability α of rejecting a "good" model in SPRT $\alpha \leq 1/A$.*

Proof. Wald's theorem [16, p. 41] states $\alpha \leq (1-\beta)/A$, where β stands for the probability that a "bad" model is incorrectly accepted as "good." In the adapted SPRT, since the only decision of the test can be "reject," $\beta = 0$, and thus, $\alpha \leq 1/A$. □

The approximation $\alpha \approx 1/A$ is close and is often used.

Theorem 2 (Wald's lemma). *The average number of observations (checked data points) carried out while testing a "bad" model is $C^{-1}\log A$, where*

$$C = p(0|H_b)\log\frac{p(0|H_b)}{p(0|H_g)} + p(1|H_b)\log\frac{p(1|H_b)}{p(1|H_g)}. \tag{6}$$

Proof. According to [16, p. 53]

$$C = \mathbb{E}\left(\log\frac{p(x|H_b)}{p(x|H_g)}\right). \tag{7}$$

The value of x is from {0, 1}. The expectation $\mathbb{E}$ is a sum of two terms weighted by probability $p(x|H_b)$. Equation (6) follows. □

In the particular case of RANSAC, $p(1|H_b) = \delta$, $p(0|H_b) = 1-\delta$, $p(0|H_g) = 1-\varepsilon$, and $p(1|H_g) = \varepsilon$. Therefore, the average number of verified correspondences per "bad" model is

$$C^{-1}\log A = \left((1-\delta)\log\frac{1-\delta}{1-\varepsilon} + \delta\log\frac{\delta}{\varepsilon}\right)^{-1}\log A. \tag{8}$$

The value of A influences the total runtime in two opposing ways. The larger the value of A, the smaller the probability of rejection of a "good" model. On the other hand, the number of correspondences verified per model increases with $\log A$ (8). We wish to set A to achieve the minimal average time needed to find the solution.

The average time-to-solution in R-RANSAC is $t = \bar{k}\bar{t}_s$, where $\bar{k}$ is the average number of samples drawn until a "good" model, and $\bar{t}_s$ is the average testing time per sample. In the following, the time unit will be the time needed to check one data point. The probability P_g of drawing a "good" model is $P_g = \varepsilon^m$, where m is the number of data points in the RANSAC sample. The number of tested samples before a "good" one is drawn and not rejected is a random variable with geometric distribution and mean $\bar{k} = 1/(P_g(1-\alpha)) \approx 1/(P_g(1-1/A))$. The average time $\bar{t}_s$ of processing a sample consists of two components: time t_M needed to instantiate a model hypotheses given a sample,[5] and the average time of testing each hypothesis. Let $\overline{m}_S$ be the number of models that are verified per sample[6] and $C^{-1}\log A$ be the average number of tests in the SPRT (Theorem 2). The average time to the solution expressed as a function of A is

$$t(A) = \frac{1}{P_g(1-1/A)}\left(t_M + \overline{m}_S \frac{\log A}{C}\right). \tag{9}$$

Formula (9) can be simplified to

$$t(A) = \frac{K_1 + K_2 \log A}{1 - 1/A},$$

where $K_1 = t_M/P_g$ and $K_2 = \overline{m}_S/(P_g C)$. We are interested in the optimal value of A,

$$A^* = \arg\min_A t(A).$$

The minimum is found by solving

$$\frac{dt}{dA} = -\frac{K_1 + K_2 - K_2 A + K_2 \log A}{(A-1)^2} = 0.$$

After rearrangement, we have

$$A^* = \frac{K_1}{K_2} + 1 + \log A^* = \frac{t_M C}{\overline{m}_S} + 1 + \log A^*. \tag{10}$$

Equation (10) has two real solutions for positive K_1/K_2, $A_1^* < 1 < A_2^*$. Since $\delta < \varepsilon$, the contribution to the likelihood ratio (5) of a correspondence that is not consistent with the model is greater than 1, therefore, the solution of interest is $A^* > 1$. This solution can be obtained as $A^* = \lim_{n\to\infty} A_n$, where $A_0 = K_1/K_2 + 1$ and $A_{n+1} = K_1/K_2 + 1 + \log(A_n)$. The series converges rapidly, typically within four iterations.

4.2 Extension to MLE Framework

In RANSAC, we directly observe whether or not a data point supports the model. In practice, an error function ρ is evaluated, and data points with the error function under a threshold are thought to support the model. In MLESAC [13], [15], it is assumed that the error $x = \rho(\theta, \mathbf{x})$ of a data point $\mathbf{x}$ with respect to a model with parameters θ is distributed as a mixture of Gaussian error distribution for inliers and uniform error distribution for outliers for a "good" model and as a uniform distribution for a "bad" model:

$$\begin{aligned} p(x|H_g) &= \varepsilon\left(\frac{1}{\sigma\sqrt{2\pi}e^{\frac{x^2}{2\sigma^2}}}\right) + (1-\varepsilon)\frac{1}{Z}, \\ p(x|H_b) &= \frac{1}{Z} \end{aligned}.$$

Then, the likelihood ratio λ_j is expressed as

$$\lambda_j = \lambda_{j-1}\left(Z\varepsilon\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{x_j^2}{2\sigma^2}} + (1-\varepsilon)\right)^{-1}, \tag{11}$$

where x_j is an error of jth data point.

The term C defining the average number $C^{-1}\log A$ of observations carried out while testing a "bad" model is derived, following (7), as follows:

$$C = \int_0^Z p(x|H_b)\log\frac{p(x|H_b)}{p(x|H_g)}dx,$$

and finally

$$C = \frac{1}{Z}\int_0^Z -\log\left(Z\varepsilon\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{x^2}{2\sigma^2}} + (1-\varepsilon)\right)dx. \tag{12}$$

The integral in (12) has to be either approximated or evaluated numerically. The rest of the derivation is identical with RANSAC.

5 R-RANSAC WITH SPRT

The R-RANSAC with SPRT algorithm is outlined in Algorithm 4. To fully the specify details of the algorithm, two issues have to be addressed. First, the estimation of parameters δ and ε; second, the termination criterion guaranteeing $1-\eta_0$ confidence in the solution has to be derived.

Algorithm 4: The structure of R-RANSAC with SPRT.

Initialize ε_0, δ_0, calculate A_0, and set $i = 0$.

Repeat until the probability η (15) of finding a model with support larger than $\hat{\varepsilon}$ falls under a user defined value η_0:

1. Hypothesis generation

- Select a random sample of minimum size m from the set of data points.
- Estimate model parameters θ fitting the sample.

2. Verification

Execute the SPRT (Algorithm 3) and update the estimates if

a. Model rejected: re-estimate δ. If the estimate $\hat{\delta}$ differs from δ_i by more than 5 percent design $(i+1)$th test $(\varepsilon_{i+1} = \varepsilon_i, \delta_{i+1} = \hat{\delta}, i = i+1)$

b. Model accepted and the largest support so far: design $(i+1)$th test $(\varepsilon_{i+1} = \hat{\varepsilon}, \delta_{i+1} = \hat{\delta}, i = i+1)$. Store the current model parameters θ.

5. Computing model parameters from a sample takes the same time as verification of t_M data points.

6. In the 7-point algorithm for epipolar geometry estimation, 1 to 3 models have to be verified.

Algorithm 4 is similar to the standard RANSAC algorithm, with the modification that with standard RANSAC, all data points are checked in the model verification step, whereas in R-RANSAC, data points are evaluated sequentially so that hypotheses with low support can be rejected before all points are considered. After a hypothesis is rejected, δ is reestimated (Algorithm 4, Step 2a). Accepted hypotheses are candidates for the RANSAC outcome (see below). The overhead of the evaluation of the likelihood ratio λ_j derived in (5) is negligible compared to the evaluation of the model versus the data point error function.

The optimal test derived in Section 4 requires the knowledge of two parameters, ε and δ. These probabilities are different for different data sets, and we assume they are unknown. The proposed algorithm uses values of ε and δ that are estimated during the sampling process, and the test is adjusted to reflect the current estimates.

If the probabilities ε and δ are available a priori in some standard setting where the algorithm is run repeatedly, they can be used in the initialization of the algorithm.

Estimation of δ. Since almost all tested models are "bad,"[7] the probability δ can be estimated as the average fraction of consistent data points in rejected models. When current estimate δ differs from the estimate used to design the SPRT (by more than 5 percent, for example), new $(i+1)$th test is designed. The initial estimate δ_0 is obtained by geometric considerations, that is, as a fraction of the area that supports a hypothesized model (a strip around an epipolar line in case of epipolar geometry) to the area of possible appearance of outlier data (the area of the search window). Alternatively, a few models can be evaluated without applying SPRT in order to obtain an initial estimate of δ.

Estimation of ε. In general, it is not possible to obtain an unbiased estimate of ε, since this would require the knowledge of the solution to the optimization problem we are solving. The tightest lower bound on ε is provided by the size of the largest support so far. It was shown in [9] that a sample with the largest support so far appears $\log k$ times, where k is the number of samples drawn. When such a sample (with support of size I_{i+1}) appears, a new test is designed for $\varepsilon_{i+1} = I_{i+1}/N$. Throughout the course of the algorithm, a series of different tests with

$$\varepsilon_0 < \ldots < \varepsilon_i < \ldots < \varepsilon$$

are performed. The initial value of ε_0 can be derived from the maximum time the user is willing to wait for the algorithm to terminate.

The termination criterion. The algorithm is terminated when the probability η of missing a set of inliers larger than the largest support found so far falls under a predefined threshold η_0. In standard RANSAC, where the probability of rejection of a "good" model is zero, the probability is equal to

$$\eta_R = \big(1 - P_g\big)^k.$$

7. RANSAC verifies, on the average, $-\log(\eta_0)$ "good" models. For the typical $\eta_0 = 0.05$, a "good" model is hypothesized three times prior to the termination of the algorithm.

In R-RANSAC, the probability of hypothesizing and not rejecting a "good" model is $P_g(1-\alpha)$, and the probability η becomes as

$$\eta = \big(1 - P_g(1-\alpha)\big)^k.$$

In R-RANSAC with SPRT, the SPRT is adjusted to the current estimates of δ_i and ε_i, so α is no more constant. Theorem 1, which gives the probability α of rejecting a "good" model for *the test designed for optimal value of* ε, does not cover this situation. The following theorem is needed:

Theorem 3. *The probability of rejecting a "good" model with fraction of inliers ε in a SPRT designed for ε_i and δ_i with threshold A_i is*

$$\alpha_i = A_i^{-h_i}, \tag{13}$$

where h_i is given by

$$\varepsilon\left(\frac{\delta_i}{\varepsilon_i}\right)^{h_i} + (1-\varepsilon)\left(\frac{1-\delta_i}{1-\varepsilon_i}\right)^{h_i} = 1. \tag{14}$$

Proof. For proof, see [16, p. 50]. □

Equation (14) has two solutions, one being $h_i = 0$. Since $\varepsilon_i < \varepsilon$, $h_i > 1$, holds for the other solution. This solution is found numerically.

For each of l tests, the following values are stored: the expected fraction of inliers ε_i, the SPRT threshold A_i, the number of samples k_i processed by the test, and h_i satisfying (14). Then, the probability η is

$$\eta(l) = \prod_{i=0}^{l} \big(1 - P_g(1 - A_i^{-h_i})\big)^{k_i}. \tag{15}$$

The number k_l of samples that are needed to be drawn for the current (that is, lth) SPRT follows from (15) as

$$k_l = \frac{\log \eta_0 - \log(\eta(l-1))}{\log(1 - P_g A_l^{-1})}. \tag{16}$$

Implementation note: since $\eta > \eta_R$, (16) does not have to be evaluated before $\eta_R < \eta_0$ is satisfied.

6 Experiments

Several experiments were performed comparing the proposed R-RANSAC with SPRT to three other RANSAC algorithms:

1. standard RANSAC that verifies all correspondences for every model,
2. R-RANSAC with the $T_{d,d}$ test [9] that rejects the model after the first checked correspondence is inconsistent with it $(d = 1)$,
3. R-RANSAC with the a priori SPRT, that is, the R-RANSAC with SPRT designed for the true values of ε and δ (labeled SPRT*), and
4. RANSAC with the bail-out test proposed by Capel [2].

The results achieved with an a priori SPRT show the best achievable performance of RANSAC with a randomized

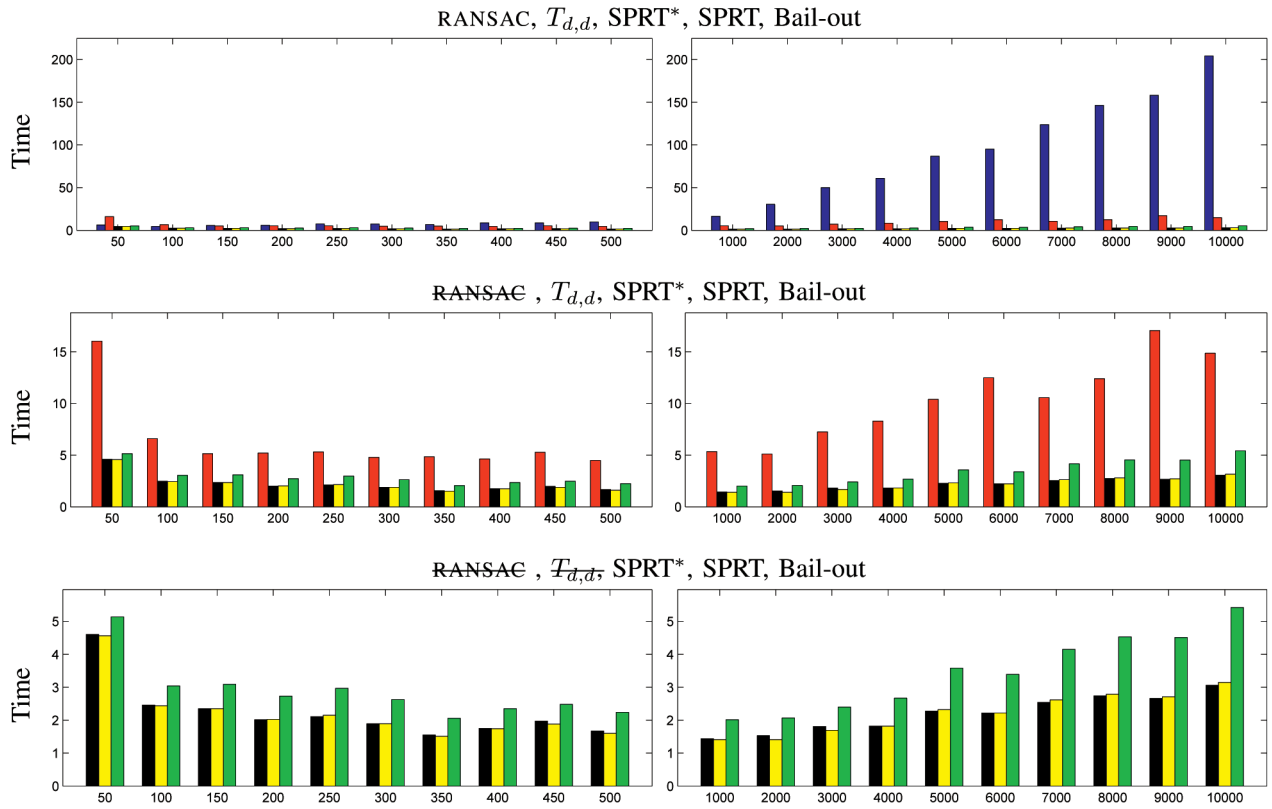

Fig. 1. Dependence of the runtime on the number N of correspondences for the fraction of inliers $\varepsilon = 0.3$. All three rows show differently scaled results of the same experiment.

verification step for a problem characterized by a given δ and ε.

For epipolar geometry estimation, the time needed to compute model parameters $t_M = 200$ was set within the range observed in a large number of experiments (that is, in our implementation, checking whether a correspondence is consistent with a fundamental matrix is 200 times faster than estimating the matrix). The exact value depends on many factors including the CPU speed and type. The constant $\overline{m}_S = 2.38$ was set to the experimentally observed average of the number of models generated by the 7-point algorithm per sample [5].[8] The initial values of δ and ε were set to $\delta_0 = 0.05$ and $\varepsilon = 0.2$, respectively.

For homography estimation, the values were set as follows: $t_M = 200$, $\overline{m}_S = 1$, $\delta_0 = 0.01$, and $\varepsilon_0 = 0.1$.

6.1 Synthetic Data

An experiment with a fixed fraction of inliers $\varepsilon = 30$ percent, and a varying number N of correspondences was undertaken. The average number of samples drawn (Fig. 2), correspondences verified (Fig. 3), and the average runtime (Fig. 1) was recorded over 50 executions. The following observations can be made: 1) The performance of RANSAC with SPRT is close to optimal (see Fig. 1), outperforming the Bail-out test. 2) For the values of ε and N in this test, neither standard RANSAC nor the $T_{d,d}$ test are competitive (see Figs. 2 and 3).

8. It is known that the 7-point algorithm produces 1 to 3 potential models. In experiments, the average number of models per sample equal to 2.38 has been observed consistently in a number of scenes. No theoretical justification on the stability of this average is known to the authors.

6.2 Real Data

The experimental image pairs are displayed in Fig. 4. The number N of correspondences and the true values of ε and δ estimated by evaluating 100,000 verifications of random models are summarized in Table 1. The results of compared algorithms are shown in Table 2.

As a consequence of the randomization of model verification that erroneously rejects some "good" models, on the average, the randomized algorithms must draw a larger number of samples than standard RANSAC. This is confirmed in the first column in Table 2. This small increase is more than compensated by the reduction in the number of data points (correspondences) checked on the average per model. The runtime of RANSAC is reduced by factors ranging between 1.9 and 9.5. In all experiments, the SPRT outperforms the $T_{d,d}$ test. In all but one of the tests, SPRT outperformed the Bail-out test; the difference was more prominent for demanding problems with lower fraction of inliers. The one scene, CORRIDOR, where the Bail-out test outperformed SPRT is the simplest one, where only about 100 models had to be tested. The a priori setting of the Bail-out test was suitable (by coincidence) for this problem. SPRT estimates ε and δ on the fly starting from conservative estimates, initially verifying needlessly many data points. Note that the time needed to solve the problem was negligible for all algorithms, ranging between 0.1 and 0.2 milliseconds.[9]

9. At this speed, the timer precision might have influenced the result.

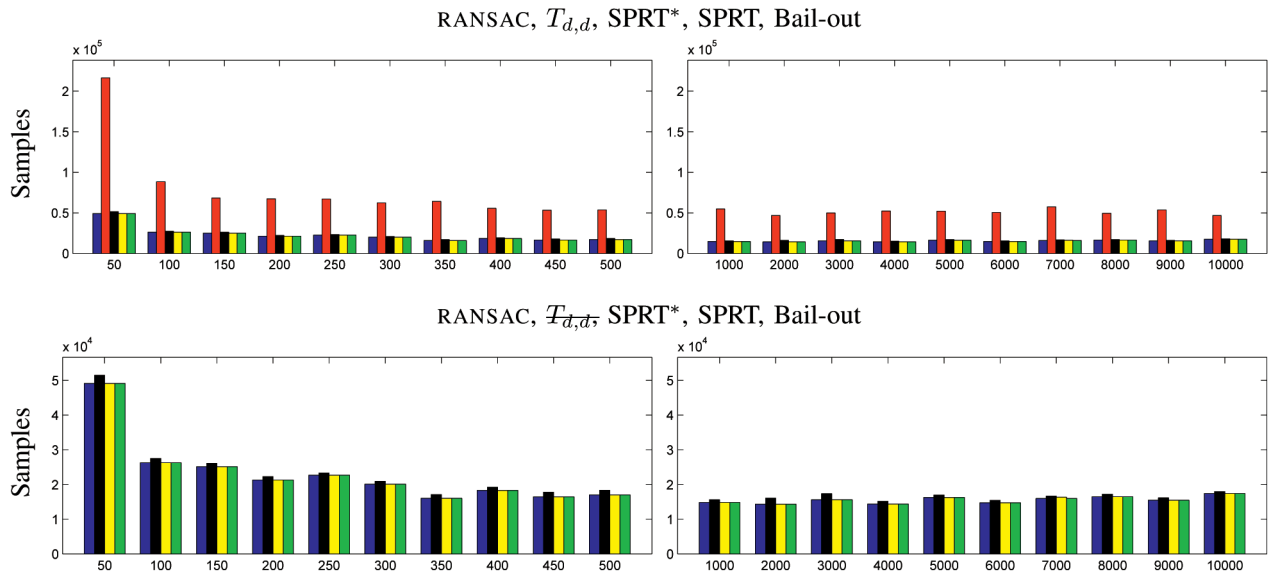

Fig. 2. Dependence of the number of samples on the number N of correspondences for the fraction of inliers $\varepsilon = 0.3$. Both rows show differently scaled results of the same experiment.

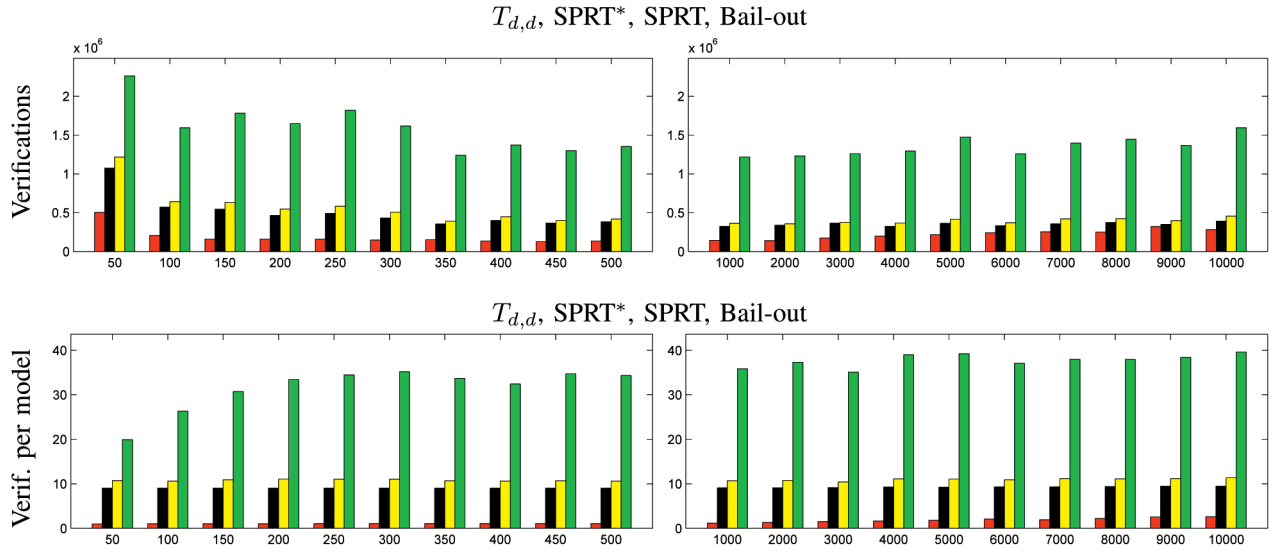

Fig. 3. Dependence of the number of verified data points on the number N of correspondences for the fraction of inliers $\varepsilon = 0.3$.

7 Notes and Discussion

The Bail-out test. The optimal R-RANSAC strategy, based on SPRT [16] derived in this paper, attempts to find the threshold for hypothesis rejection such that the overall runtime is minimized. Note that the bail-out test [2] can be thought of as a special case, when the threshold (the probability that the score of current model exceeds the best score so far) is fixed at 1 percent.

Number of samples. Table 1 shows the predicted number of samples for standard RANSAC, according to

$$k^* = \frac{\log \eta_0}{\log(1 - P_g)},$$

for $\eta_0 = 0.05$. Note that the predicted number of samples is two to three times lower than the actual number of samples drawn. This is due to an error on inlier measurements, which causes some all-inlier samples not to be consistent with all inliers. The phenomenon is in agreement with previous observations [4], [13]. Since the discrepancy is of a multiplicative character in the number of samples and thus the runtime, all parameters and thresholds derived to minimize the runtime are valid. Note that the results of the synthetic experiments are exact, since no inlier noise was added.

Epipolar geometry for narrow baseline stereo. For narrow baseline stereo matching, the tentative correspondences are selected only from a small disparity window. Therefore, even a mismatch is never too far from the correct correspondence. When an EG is estimated from a "bad" sample of 7 such correspondences, one of the up to three EGs is likely to be close to the correct EG. Hence, such a model will have a different distribution of errors, leading to imprecise estimates of δ. However, the problem is specific to a narrow baseline stereo and epipolar geometry estimation. This is beyond the scope of this paper, covering general use of Wald's theory in randomizing the verification step of RANSAC.

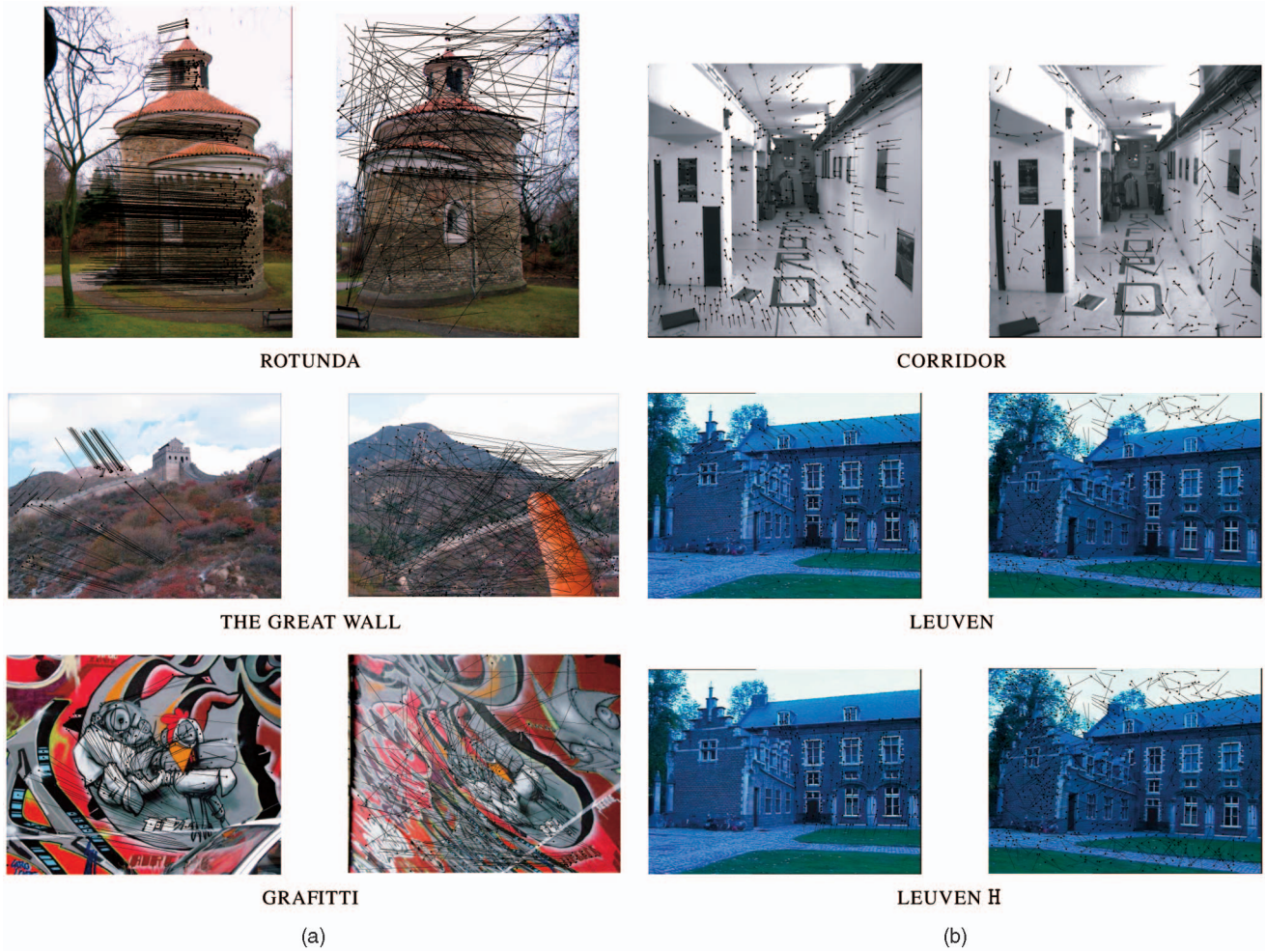

Fig. 4. The experimental image pairs with (a) inliers and (b) outliers superimposed. Two wide-baseline epipolar geometry experiments ROTUNDA and the GREAT WALL; two narrow-baseline EG experiments LEUVEN and CORRIDOR; two homography experiments GRAFITTI and LEUVEN H.

Predicted number of verifications per model. The average number of models verified per "bad" model is given by (8). However, some of the samples drawn in the course of RANSAC are "good." On the average, every $1/P_g$ sample is "good." Out of good samples, one of the $\overline{m}_S$ models is "good." Therefore, the predicted number of verifications per model is

$$\text{vpn} = (1 - P_g)v_b + P_g\left(\frac{N + (\overline{m}_S - 1)v_b}{\overline{m}_S}\right), \qquad (17)$$

TABLE 1
Number of Correspondences (Corr), Fraction of Inliers (ε), the Probability of a Correspondence Being Consistent with a Bad Model (δ), and Predicted Number of Samples k^* in Standard RANSAC

	corr	ε	δ	k^*
LEUVEN	793	0.49	0.043	453
CORRIDOR	607	0.67	0.174	49
ROTUNDA	619	0.33	0.014	7536
GREAT WALL	514	0.28	0.015	24711
LEUVEN H	793	0.26	0.010	668
GRAFITTI H	409	0.41	0.010	107

where $v_b = C^{-1} \log A$ from (8). The prediction therefore holds tight for difficult scenarios with low fractions of inliers, see Table 3 for an example. There, the vast majority of verified models are indeed "bad," and the numbers are statistically significant. The number of verifications per model is underestimated for narrow baseline experiments, since many "bad" models are considered further in the verification, as discussed above.

8 Conclusions

An optimal sequential strategy for randomized evaluation of model quality in RANSAC was derived. A method for the estimation of two probabilities characterizing the problem and critically influencing the design of the optimal strategy was proposed and experimentally verified. A termination criterion derived in the paper guarantees that the solution found is correct with confidence $1 - \eta_0$, as in standard RANSAC.

The properties of R-RANSAC with SPRT were tested on a diverse set of standard data. The test problems included homography estimation and epipolar geometry estimation in both wide and narrow baseline setting.

R-RANSAC with SPRT was 2 to 9 times faster than standard RANSAC and up to 3.5 times faster than R-RANSAC

TABLE 2
The Comparison of RANSAC, R-RANSAC with $T_{d,d}$ Test, a priori SPRT^* and SPRT: The Number of Samples (k), the Number of models (Models), the Number of Checked Correspondences per Model (VPM), Time in ms (Time), and Relative Speed-Up with Respect to Standard RANSAC (SPD-Up) Averaged over 500 Runs

The LEUVEN experiment					
	k	models	vpm	time	spd-up
RANSAC	1276	2504	786.0	6.6	1.0
R-RANSAC	2935	5761	37.6	2.1	3.1
Wald opt	1339	2629	19.3	1.0	6.5
Wald	1292	2537	23.9	1.0	6.3
Bail out	1276	2505	25.0	1.0	6.4

The CORRIDOR experiment					
	k	models	vpm	time	spd-up
RANSAC	104	183	600.0	0.4	1.0
R-RANSAC	165	291	115.9	0.2	2.0
Wald opt	108	190	44.2	0.1	3.2
Wald	104	185	132.2	0.2	1.9
Bail out	104	183	54.8	0.2	2.8

The ROTUNDA experiment					
	k	models	vpm	time	spd-up
RANSAC	15500	28012	612.0	56.7	1.0
R-RANSAC	50410	91116	9.2	26.1	2.2
Wald opt	16712	30204	10.5	9.6	5.9
Wald	15809	28570	12.5	9.6	5.9
Bail out	15521	28050	27.4	11.1	5.1

The GREAT WALL experiment					
	k	models	vpm	time	spd-up
RANSAC	56072	111961	507.0	190.3	1.0
R-RANSAC	216611	432509	8.6	113.5	1.7
Wald opt	60869	121540	11.2	36.1	5.3
Wald	57263	114338	13.5	36.3	5.2
Bail out	56136	112089	33.5	44.4	4.3

The LEUVEN H experiment					
	k	models	vpm	time	spd-up
RANSAC	1876	1876	789.0	51.0	1.0
R-RANSAC	9058	9058	2.8	19.6	2.6
Wald opt	2025	2025	14.9	5.4	9.5
Wald	1898	1898	19.9	5.4	9.5
Bail out	1880	1880	41.6	6.8	7.5

The GRAFITTI H experiment					
	k	models	vpm	time	spd-up
RANSAC	205	205	405.0	3.1	1.0
R-RANSAC	584	584	13.2	1.6	2.0
Wald opt	213	213	31.8	0.7	4.2
Wald	206	206	39.5	0.8	4.0
Bail out	205	205	34.0	0.7	4.2

with $T_{d,d}$ test. The Bail-out test had a similar performance on simple problems (where the solution of the optimization problem was found almost immediately by all algorithms) and was outperformed on more challenging tasks. Experiments on a large set of synthetic problems confirms the quasi-optimality of R-RANSAC with SPRT.

ACKNOWLEDGMENTS

The authors are grateful for support form the Czech Science Foundation Project 102/07/1317 and an EPSRC Platform grant.

TABLE 3
Number of Verified Correspondences per Model: Predicted Number for "Bad" Models by (8), Predicted Number over All Models (17), and Observed Number in Optimal Wald

	vpn 'bad'	vpn	vpn Wald
LEUVEN	7.7	10	19.3
CORRIDOR	7.4	22.5	44.2
ROTUNDA	10.4	10.5	10.5
GREAT WALL	12.4	12.4	11.2
LEUVEN H	12.7	16.2	14.9
GRAFITTI H	7.9	19.3	31.8

REFERENCES

[1] *Proc. IEEE Int'l Workshop "25 Years of RANSAC" in conjunction with CVPR '06 (RANSAC25 '06),* http://cmp.felk.cvut.cz/cvpr06-ransac, 2006.
[2] D. Capel, "An Effective Bail-Out Test for RANSAC Consensus Scoring," *Proc. British Machine Vision Conf. (BMVC '05),* pp. 629-638, 2005.
[3] O. Chum and J. Matas, "Randomized RANSAC with T(d, d) Test," *Proc. British Machine Vision Conf. (BMVC '02),* vol. 2, BMVA, pp. 448-457, 2002.
[4] O. Chum, J. Matas, and J. Kittler, "Locally Optimized RANSAC," *Proc. Ann. Symp. German Assoc. for Pattern Recognition (DAGM '03),* 2003.
[5] O. Chum, T. Werner, and J. Matas, "Epipolar Geometry Estimation via RANSAC Benefits from the Oriented Epipolar Constraint," *Proc. Int'l Conf. Pattern Recognition (ICPR '04),* vol. 1, pp. 112-115, Aug. 2004.
[6] M. Fischler and R. Bolles, "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography," *Comm. ACM,* vol. 24, no. 6, pp. 381-395, June 1981.
[7] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision,* second ed. Cambridge Univ., 2003.
[8] P.M. Lee, *Sequential Probability Ratio Test.* Univ. of York, www.york.ac.uk/depts/maths/teaching/pml/ais/sprt.ps, 2007.
[9] J. Matas and O. Chum, "Randomized RANSAC with $T_{d,d}$ Test," *Image and Vision Computing,* vol. 22, no. 10, pp. 837-842, Sept. 2004.
[10] J. Matas and O. Chum, "Randomized RANSAC with Sequential Probability Ratio Test," *Proc. Int'l Conf. Computer Vision,* vol. 2, pp. 1727-1732, Oct. 2005.
[11] D. Myatt, P. Torr, S. Nasuto, J. Bishop, and R. Craddock, "NAPSAC: High Noise, High Dimensional Robust Estimation—It's in the Bag," *Proc. British Machine Vision Conf. (BMVC '02),* vol. 2, pp. 458-467, 2002.
[12] D. Nister, "Preemptive RANSAC for Live Structure and Motion Estimation," *Proc. Int'l Conf. Computer Vision (ICCV '03),* vol. 1, pp. 199-206, Oct. 2003.
[13] B. Tordoff and D. Murray, "Guided Sampling and Consensus for Motion Estimation," *Proc. Seventh European Conf. Computer Vision (ECCV '02),* vol. 1, pp. 82-96, 2002.
[14] P. Torr, A. Zisserman, and S. Maybank, "Robust Detection of Degenerate Configurations while Estimating the Fundamental Matrix," *Computer Vision and Image Understanding,* vol. 71, no. 3, pp. 312-333, Sept. 1998.
[15] P. Torr and A. Zisserman, "MLESAC: A New Robust Estimator with Application to Estimating Image Geometry," *Computer Vision and Image Understanding,* vol. 78, pp. 138-156, 2000.
[16] A. Wald, *Sequential Analysis.* Dover, 1947.

Ondřej Chum received the MSc degree in computer science from Charles University, Prague, in 2001 and the PhD degree from the Centre for Machine Perception, Czech Technical University, Prague, in 2005. From 2005 to 2006, he was a research Fellow at the Centre for Machine Perception, Czech Technical University. Recently, he is with the Visual Geometry Group, University of Oxford, UK. His research interests include object recognition, large-scale image and particular-object retrieval, invariant feature detection, and RANSAC-type optimization. He is a reviewer for the *IEEE Transactions on Pattern Analysis and Machine Intelligence* and other journals. He is regularly a program committee member at major computer vision conferences (for example, ICCV, CVPR, and ICPR). He has co-organized the "25 years of RANSAC" Workshop in conjunction with CVPR 2006. He was the recipient of the Best Paper Prize at the British Machine Vision Conference in 2002. He is a member of the IEEE.

Jiří Matas received the MSc degree (with honors) in cybernetics from the Czech Technical University, Prague, in 1987 and the PhD degree from the University of Surrey, UK, in 1995. From 1991 to 1997, he was a research Fellow at the Centre for Vision, Speech and Signal Processing, University of Surrey, working with J. Kittler. In 1997, he joined the Center for Machine Perception, Czech Technical University. Since 1997, he has held various positions at these two institutions. His research interests include object recognition, sequential pattern recognition, invariant feature detection, and Hough Transform and RANSAC-type optimization. He has published more than 100 papers in refereed journals and conference proceedings. His publications have more than 800 citations in the Science Citation Index. He is on the editorial board of the *IEEE Transactions on Pattern Analysis and Machine Intelligence* and the *International Journal of Computer Vision*. He has served in various roles at major international conferences (for example, ICCV, CVPR, ICPR, and NIPS) and co-chairing ECCV 2004 and CVPR 2007. He received the Best Paper Prize at the British Machine Vision Conferences in 2002 and 2005. He is a member of the IEEE.