ISSN 1392–1541 **Geodezija ir kartografija, 2005, XXXI t., Nr. 3**
Geodesy and Cartography, 2005, Vol XXXI, No 3

UDK 528.73

RANSAC FOR OUTLIER DETECTION

[1]Birutė Ruzgienė, [2]Wolfgang Förstner

[1]Dept of Geodesy and Cadastre, Vilnius Gediminas Technical University,
Saulėtekio al.11, LT-10223 Vilnius-40, Lithuania
e-mail: Birute.Ruzgiene@ap.vtu.lt
[2]Institute of Photogrammetry, Bonn University, Nussallee 15, D-53115, Germany
e-mail: wf@ipb.uni-bonn.de

Received 10 06 2005, accepted 07 07 2005

Abstract. Up-to-date digital photogrammetry involves operations on huge data sets, and with classical image processing procedures it might be time consuming to find out the best solution. One of the key tasks is to detect outliers in given data, eg for curve fitting or image matching. The problem is hard as the number of outliers is usually large, possibly larger than 50 %, thus powerful estimation techniques are needed. We demonstrate one of these techniques, namely *Random Sample Consensus* (*RANSAC*), for fitting a model to sample data, especially for fitting a straight line through a set of given points. Experiments with up to 80 % outliers prove the efficiency of *RANSAC*. The results are representative for image analysis in digital photogrammetry.

Keywords: statistics, uncertainty, probability, parameter estimation, outliers, inliers, matching.

1. Introduction

Progress in digital photogrammetry relies on automatic procedures exploiting the technological possibilities of high resolution image acquisition, fast access to disk space and high computing power. Research mainly deals with conceptual problems in photogrammetric evaluation processes, the goal being a highly efficient software. This requires flexible and robust solutions to estimation and classification problems.

In this paper we discuss the problem of estimation in the presence of high percentages of outliers. This type of problem occurs in all stages of automatic calibration, orientation and surface reconstruction, as automatic image matching procedures are error prone. Statistics and probability theory are indispensable for handling uncertainty and estimating parameters under these conditions. There exist powerful *robust* estimation techniques. However, no unique technique exists which is applicable in all situations.

One of the most attractive techniques is *Random Sample Consensus* (*RANSAC*). It randomly chooses a minimal set of observations and evaluates their likelihood until a good solution is found or a preset number of trials is reached. *RANSAC* is a technique which is best suited for estimation problems with a small number of parameters and a large percentage of outliers. It has regularly been applied for estimation of model parameters in feature matching, detection and registration.

The paper gives a demonstration on the ability of *RANSAC* to correctly estimate parameters, even when the percentage of outliers is far beyond 50 % and the outliers hide the true solution. As an example we take the classical problem of straight line fitting.

Experiments have been made with *MATLAB* 7.0, a development package with a high level language for programming and visualization [1].

2. Model fitting by *RANSAC*

We assume the data contain a certain percentage ε of outliers. The inliers are usually assumed to be distorted by small independent measurement errors following a Gaussian distribution, often with an equal variance.

If no assumption on the distribution of the outliers is made, it might happen that the outliers imitate the model, with the consequence that any estimation procedure would fail if $\varepsilon > 0{,}5$. Fortunately, we often are in a better situation, where higher percentages of outliers might be allowed. For our experiments we assume the outliers to be uniformly distributed in the space of observations.

Given a data set of N observations, in principle each observation may be wrong. In case of N data one would therefore need to check all 2^N possible combinations of observations to find a subset of observations yielding an *optimal solution*. It was the idea of Fischler & Bolles in 1981 ([2], cf also [3, 4]), to see that a much smaller number of trials is necessary to find a *good solution*, which may then be refined by classical procedures.

2.1. The prerequisites

The basic *RANSAC* algorithm assumes the following input:

- The data set $D = \{d_i\}$ with in total $N = |D|$ items. In our case we have a set $\{\boldsymbol{x}_i \; i=1, \ldots, N\}$ of N data points $\boldsymbol{x}_i = [x_i, y_i]'$.

- A function $\boldsymbol{p} = F(S)$, which directly computes the model parameters $\boldsymbol{p}$ from a subset S of data items containing a minimal sufficient number $M = |S|$ of items. *Directly* means that no approximate values are needed. To yield highest efficiency, the number of data items necessary for this determination should be as small as possible. In our case we need $M = 2$ points, say with homogeneous coordinates $\boldsymbol{x_1} = [x_1, y_1, 1]'$ and $\boldsymbol{x_2} = [x_2, y_2, 1]'$ from which the homogeneous line parameters $\mathbf{l} = [a, b, c]'$ can be determined via $\mathbf{l} = \boldsymbol{x} \times \boldsymbol{y}$.
- A cost function $\rho(\boldsymbol{p}, d_i)$ for each data item d_i and for each parameters $\boldsymbol{p}$ gives costs. This cost function is chosen such that outliers have only a limited influence onto the parameter to be estimated. In general, it will depend on some residual

$$e_i = e_i(\boldsymbol{p}, d_i) \tag{1}$$

of the observations referring to a certain model, specified by parameters $\boldsymbol{p}$. If in our example we would want to achieve a least-squares solution, we would use the total costs depending on the distances d_i of the points to the fitted line

$$c_0 = \bar{d} = \frac{\sqrt{\sum_{i=1}^{N} d_i^2}}{N}. \tag{2}$$

For achieving robustness, in the most simplest case we may just count the number of outliers by choosing the function ρ in Fig 1. It is calculated comparing points' distance d_i from the line with given tolerance: if $d_i > k$, these points are outliers and $\rho = 1$; otherwise the points are within the tolerance and are inliers $\rho = 0$. We use the total costs

$$C = \sqrt{\frac{\sum_{i=1}^{N} \rho(d_i)}{N}}, \tag{3}$$

with

$$\rho(d_i) = \begin{cases} 1, \text{ if } |d_i| < k. \\ 0, \text{ else} \end{cases} \tag{4}$$

Observe that by using $\rho(d_i) = d_i^2$ we obtain the least-squares solution.

In order to achieve a least-squares solution in case of only inliers we choose the cost function in Fig 1: if $|d_i| < k$, the points are good and the individual costs are $\rho = d_i^2$; if $|d_i| \geq k$, the points are bad and the individual costs are $\rho = k^2$ (Fig 2), thus

$$\rho(d_i) = \begin{cases} d_i^2, \text{ if } |d_i| < k. \\ k^2, \text{ else} \end{cases} \tag{5}$$

Thus the total costs yield the sum of the number of outliers and the sum of the normalized square distances of the inliers.

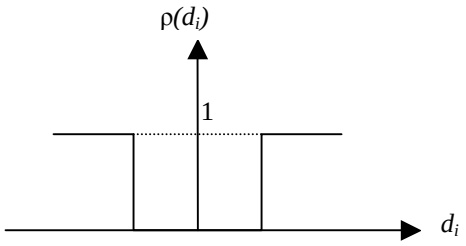

Fig 1. Simple cost function ρ only counting the number of outliers

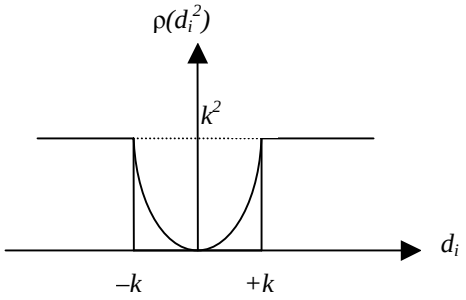

Fig 2. Cost function: for small errors it leads to a least-squares solution and large values have a limited influence on the estimation

2.2. The procedure

The idea of the *RANSAC* algorithm is to repeatedly select a random subset S of the data, to determine a solution $\boldsymbol{p} = F(S)$ and to evaluate it with other data. The algorithm is therefore comprised of the following steps:

1. n times, do repeat steps 2 to 4, where n has to be specified by the user (cf below).
2. **Select a sample S of a minimum number M of data items from the data set.** This selection might take all combinations, which would take $n = \binom{N}{M}$ samples. In our example this would require $n = N(N–1)/2$ samples, and though feasible for small data sets, for large data sets this would require too many samples Therefore, the sample set is chosen randomly. In a first instance, each data item has the same probability of selection. Thus we obtain a subset $S_\nu \subset D$, $|S_\nu| = M$, $\nu \in [1,\ldots, n]$. We might directly exclude subsets which do not allow stable parameter estimation. In our case we select n pairs of points randomly, possibly having excluded pairs of points which are too close to each other to yield a stable line.
3. **Estimate the parameters $\boldsymbol{p_k}$ of the model based on the subset S_k from $\boldsymbol{p_k} = F(S_k)$.** In our context we determine the straight line through the $M = 2$ selected points.
4. **Compute the quality of the model with parameters $\boldsymbol{p}_k$ using the cost function** $C_k = \Sigma_{x \in S} \rho(\boldsymbol{p_k}, x)$. In case we would put a threshold on C_k to accept a solution this would lead to a model

verification. Then we could stop the search for an acceptable solution if C_k would be accepted. However, for stability reasons we evaluate all samples.

5. **Choose the best model, i e the set S_k with parameters p_v where C_k is minimum**. In addition, we might accept the solution only if the number of inliers is larger than a prespecified threshold, say *t*.

2.3. The parameters

RANSAC uses three parameters:

1. the number *n* of trials,
2. the threshold *k* for determining where a data point agrees with model in (4) or (5), and possibly
3. the threshold *t* for the required number of inliers.

ad 1: The number of trials is to be chosen carefully. If not all combinations of data items are tested, one cannot be certain to obtain a valid result. However, one can estimate the probability *P* to obtain at least one good solution in *k* trials from a subset of *M*, if the percentage of errors is ε. This is given by (proof cf below)

$$P = 1 - (1 - (1 - \varepsilon)^M)^n. \qquad (6)$$

If one now requires a minimum probability $P_{\min}$ to obtain at least one good sample, then the minimum number n_{min} of trials is

$$n_{\min}(P_{\min,\varepsilon,M}) = \frac{\ln(1-P_{\min})}{\ln(1-(1-\varepsilon)^M)}. \qquad (7)$$

Table 1. Minimum number of trials n_{min} to find at least one solution with minimum probability P_{min} for an assumed percentage of errors ε

P_{min}	ε						
	0,1	0,3	0,5	0,5	0,7	0,8	0,9
0,95	2	5	***11***	18	32	74	299
0,99	3	7	17	27	49	113	459
0,999	5	11	25	40	74	***169***	688

For our example, the minimum numbers of trials are given in Table 1 for the minimum probability $P_{\min} = 95\,\%$, 99 % and 99,9 %, and for different percentages of assumed outliers, between ε = 10 % and 90 %. For example, if the expected percentage of outliers is 50 % at least 11 trials are necessary, if $P_{\min}$ of finding at least one correct solution is 95 %.

The minimum number $n_{\min}$ of trials thus depends on two parameters, $P_{\min}$ and ε. Whereas the minimum probability may be easily chosen, there might be an uncertainty about the expected percentage of outliers.

These minimum number of trials are independent on the number of data items. So, if, for example, a line has to found in 100 points, of which 50 % are erroneous, one would need 100(100–1)/2 = 4950 trials if all combinations would be tested. Obviously, a large reduction in the number of trials can be achieved.

We will see later that the required minimum probability $P_{\min}$ is not really reached.

Proof of (6): *P (*at least one good in *n* samples) = 1 – *P* (all *n* samples are bad) = 1 – *P* (sample is bad)^n = 1 – (1 – *P* (all *M* entities of a sample are good))^n = 1 – (1 – *P* (one entitiy of a sample good)^M)^n = 1 – (1 – (1 – *P* (one entity of a sample bad)^M) ^{n}.

ad 2: The threshold *k* requires knowledge about the quality of the data items. In case this is not given, one might determine *k* from the residuals e_i, for example, using the median or the N_{in} smallest residuals, where N_{in} = (1–ε)*N* is the expected number of inliers.

ad 3: Here one could just take the number N_{in} of expected number of inliers.

3. Experiments

We developed simulation software to investigate the power of *RANSAC*. The user specifies the number of points, the percentage of outliers, the straight line and the measuring precision used for generating the inlier data. In addition, the user specifies the required minimum probability for success and the expected error rate, independently of the generation in order to investigate the effect of erroneous assumptions.

3.1. Sampling and modelling the input data set

We assume the *outliers* to be uniformly distributed in the square $[-1,+1]^2$.

The *inliers* are assumed to sit on a straight line, except for random errors. The line is specified by its distance *s* to the origin and the direction ϕ of its normal. The positions of the points *along* the line are assumed to be uniformly distributed along the line segment within the unit circle (cf Fig 3). The positions of the points *across* the line are assumed to be normally distributed with standard deviation σ.

Generating the inliers starts with determining the endpoint of the line segment in the unit circle. They are symmetric with respect to the point $[x_f, y_f]$ closest to the origin

$$x_f = s\cos(\varphi),$$
$$y_f = s\sin(\varphi).$$

The starting and the end points of the line segment are given by

$$\begin{aligned} x_e &= x_f + t\cos(\varphi + \pi/2), \\ y_e &= y_f + t\sin(\varphi + \pi/2), \end{aligned} \qquad (8)$$

where $t = \sqrt{1-s^2}$ for the starting and $t = -\sqrt{1-s^2}$ for the end point.

Therefore the true values for the points on the line are

$$\begin{aligned} x_i &= x_f + k\,t\cos(\varphi + \pi/2), \\ y_i &= y_f + k\,t\sin(\varphi + \pi/2), \end{aligned} \qquad (9)$$

where $k \in [-1,+1]$.

The observed point coordinates are:

$$x_i = \tilde{x}_i + e_{x_i},$$
$$y_i = \tilde{y}_i + e_{y_i}, \qquad (10)$$

where e_{x_i} and e_{y_i} are normally distributed random variables with standard deviation σ.

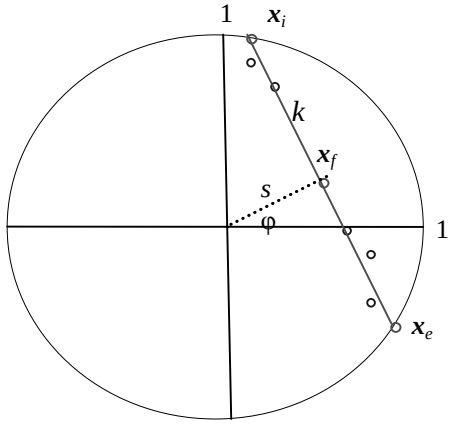

Fig 3. Line segment in the unit circle with inliers. Point x_f is closest to the origin. The inliers are uniformly distributed along the line segment between start and endpoint x_i and x_f

3.2. An example

Fig 4 shows the data generated by the programme for one representative case. The N = 100 points with an outlier rate of 80 % yields a line with 20 inliers having a standard deviation of σ = 0,02. Obviously classical least-squares techniques will not be able to find a correct estimate. Also iteratively eliminating points with large residuals will be not a successful strategy. In spite of a high percentage of outliers a human observer easily can detect the line.

If we now apply *RANSAC* with $P_{\min}$ = 0,999, an expected error rate of ε = 0,8 and a critical value $k = 2\sigma$ = 0,04, we have to try $n_{\min}$ = 169 samples.

The detected line is shown in Fig 5. The true parameters of the line are φ = 0,8 and s = 0,2. The estimated values differed only by approximately 0,01, which corresponds to the expected accuracy.

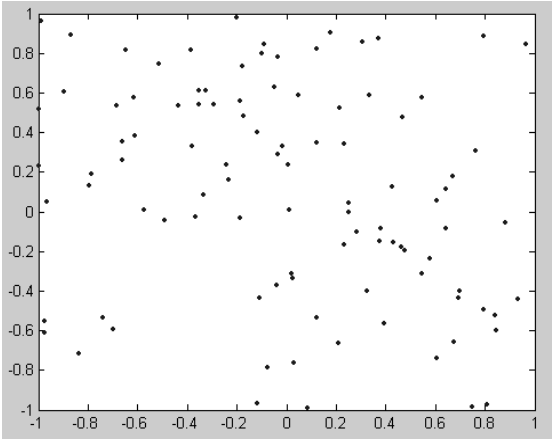

Fig 4. Example for generated data: total number of points N = 100; number of good points on line 20 = 100 (1–80 %); direction angle of normal to given line φ = 0,8 (in rad); distance in *Hessian* normal form of given line to origin s = 0,2; standard deviation of points on line σ = 0,02

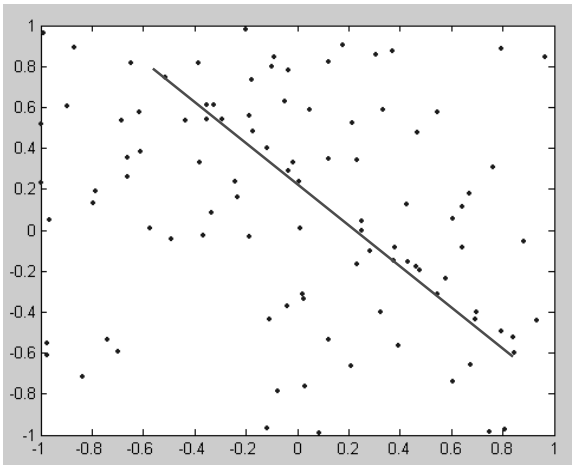

Fig 5. Detected line: true values φ = 0,8; s = 0,2; σ = 0,02

3.3. On the probability to find a good line

The theoretical probability to find a good line in the previous example is expected to be $P_{\min}$ = 0,999. During testing the algorithm we had the impression that the line was not found with this high probability. This has also been found by other researchers (cf [5], [6]).

For further investigation, we automatically performed the line generation and the line detection 10 000 times. We counted the number of cases where the estimated angle φ was closer to the true angle than 6σ and the estimated distance s were closer than 6σ from the true distance.

We varied the standard deviation from σ = 0,0001 to σ = 0,1 in steps of factor $\sqrt{10}$ and obtained the results collected in Table 2.

Table 2. Empirical probability of finding the correct line in N = 100 and 40 points with 80 % outliers using 169 samples in *RANSAC*

σ	$P(N = 100)$	$P(N = 40)$
0,0001	0,988	0,977
0,003	0,990	0,975
0,001	0,990	0,974
0,03	0,991	0,962
0,01	0,996	0,929
0,03	0,997	0,833
0,1	0,976	0,801

The positive result, shown in Table 2, is the high probability of finding a straight line in extremely noisy data.

However, we have never reached the expected probability $P_{\min}$. The reason is simple and can be seen in the extreme: if the inliers are very noisy, then there is a high chance that there is another straight line through the outliers and imitates a good line. The effect is larger if the number of points is smaller.

Obviously, to be successful with *RANSAC* the number of inliers needs to be larger than a minimum. This minimum depends on the total number of data and

the standard deviation of the inliers compared to the density of the outliers. This relation appears to be not known and needs to be investigated.

4. Conclusions

This paper demonstrates the ability of robust estimators to cope with large percentages of blunders. Random sample consensus is an effective tool to detect the true parameters also in presence of highly corrupted data.

Our experiments were performed with uniformly distributed outliers. The quality of the results will be lower if the outliers show regularities. In the extreme, they may imitate another line. Therefore the theoretical breakdown point of all robust estimators, where there is no general guarantee of success, is less than 50 %.

On the other hand, the required probability for finding at least one good line is not reached. The difference is significant. The experiments suggest this probability to be lower in case of small data sets and high measurement noise, and may go down to below 80 %. These results need further investigation in order to explore the limits of robust estimators.

Altogether, *RANSAC* is a powerful tool to cope with large percentages of blunders. It can be successfully used in digital image analysis, feature matching procedures as well as for automatic relative orientation of images particularly in close range photogrammetry

References

1. *Matlab* tutorial. Department of Mathematics, University of Utah. http://www.math.utah.edu/.
2. Fischler, M. A. and Bolles, R. C. Random sample consensus: A paradigm for model fitting with application to image analysis and automated cartography. *Communications of the Association for computing machinery*, 24 (6), 1981, p. 381–395.
3. Manual of Photogrammetry (Ed by Chris McGlone). Fifth Edition. American Society for Photogrammetry and Remote Sensing, USA, 2004. 1151 p.
4. Hlavac, V. RANSAC. Czech Technical University, Faculty of Electrical Engineering, Department of Cybernetics, Center for Machine Perception. http://cmp.felk.cvut.cz/.
5. Torr, P. H. S.; Davidson, C. IMPSAC: Synthesis of Importance Sampling and Random Sample Consensus. *Pattern analysis and machine intelligence*, (25), No 3, March 2003, IEE Computer Society, p. 354–364.
6. Myatt, D. R.; Torr, P. H. S.; Nasuto, S. J.; Bishop, J. M.; Craddock, R. NAPSAC: High Noise, High Dimensional Robust Estimation – it's in the Bag. In: Proceedings of the British Machine Vision Conference, Cardiff, United Kingdom, 2002, p. 485–467.

***RANSAC* TAIKYMAS KLAIDINGIEMS DUOMENIMS APTIKTI**

B. Ruzgienė, W. Förstner

S a n t r a u k a

Nūdienos skaitmeninė fotogrametrija nagrinėja fotografinių vaizdų, kuriuose gausu duomenų, apdorojimo procedūras, todėl automatiškai rasti geriausią sprendimą ilgai trunka, būtina talpi kompiuterinė atmintis. Atliekant fotonuotraukų sugretinimą (*matching*), vienas iš pagrindinių uždavinių yra teisingai identifikuoti duomenų elementus. Sprendžiant šį uždavinį, kyla klaidingų duomenų, kurių paprastai yra daug (gali būti daugiau nei 50 %), eliminavimo problema. Tam tikslui turi būti parinkta tinkama duomenų įvertinimo metodika.

Analizuojama statistinis duomenų įvertinimo metodas *RANSAC* (*Random Sample Consensus*), skirtas sudarytajam modeliui suderinti su parinktaisiais duomenimis, t. y. šiuo atveju nagrinėjama tiesios linijos, einančios per turimą taškų visumą, radimo ypatumai.

RANSAC efektyvumui nustatyti atliktas eksperimentas – įvertintos tiesios linijos generavimo procedūros, kai nurodoma minimali tikimybė bei paklaidos dydis (žr. 5 pav., 2 lentelę). Eksperimento metu nustatyta, kad teisingo sprendimo tikimybė bus mažesnė, jei duomenų modelis bus mažesnis, o matavimų paklaidos didesnės.

Tyrimo rezultatai parodė, kad net ir esant 80 % klaidingų duomenų (*outliers*), taikyti *RANSAC* yra labai efektyvu – įvedus teisingus parametrus, gaunamas optimalus sprendimas.

RANSAC taikymo klaidingiems duomenims aptikti, atliekant automatinį vaizdų sugretinimą, galimybių tyrimas turėtų būti tęsiamas ateityje.

Raktažodžiai: statistika, netikslumas, tikimybė, parametrų įvertinimas, klaidingi duomenys, sugretinimas.

Birutė RUZGIENĖ. Associate Professor, Doctor.

Vilnius Gediminas Technical University, Dept of Geodesy and Cadastre, Saulėtekio al. 11, LT-10223 Vilnius-40, Lithuania (Ph +370 5 2744703, Fax +370 5 2744705), e-mail: birute.ruzgiene@ap.vtu.lt.

A graduate of Vilnius Civil Engineering Institute (engineer of geodesy, 1968). Doctor (Vilnius Gediminas Technical University, 1999). Research training at Moscow Institute of Geodesy, Aerosurveying and Cartography (1986), at Norway AO Fjellanger Widerøe (1995), at Warsaw Institute of Geodesy and Cartography (1998), at Photogrammetry Institute of Bonn University (2000, 2005). Author of more than 30 scientific papers.

Research interests: digital photogrammetric mapping, image interpretation, features extraction from remote sensing data.

Wolfgang FÖRSTNER. Prof Dr-Ing, Director of the Institute of Photogrammetry, University of Bonn. Nussallee 15, D-53115 Bonn, Germany. e-mail: wf@ipb.uni-bonn.de.

A graduate of Stuttgart University (Geodesy, 1971). PhD (Stuttgart University, Photogrammetry, 1976). In 2005 he received the Photogrammetric Award (Fairchild) in honour of his major contributions to the science of photogrammetry, by helping to establish the increasingly important ties between photogrammetry, digital image processing, and computer vision.

Research interests: image interpretation, computer vision, statistics, geometry and image analysis.