RANSAC
Robust Model Estimation
From Data Contaminated By Outliers

Lecturer: Jan Čech

Authors: Ondřej Chum, Jiří Matas, Ondřej Drbohlav
Czech Technical University in Prague
http://cmp.felk.cvut.cz
What is RANSAC?

- RANSAC = RANdom SAmple Consensus

**Example:** Finding a line in 2D data.
- Not all input points are on a line.
- Finding a line also implicitly divides points to **inliers** (= those on a line) and **outliers** (= those not on a line).
Line Fitting: Line Parametrization

- Line parametrization (homogeneous):

\[ ax + by + c = 0, \quad (a \neq 0 \lor b \neq 0) \]  \hspace{1cm} (1)

\[ a, b, c \in \mathbb{R} : \text{line parameters} \]  \hspace{1cm} (2)

\[ (x, y) : \text{point coordinates} \]  \hspace{1cm} (3)

- Line parametrization (radial):

\[ x \cos \phi + y \sin \phi = r, \]  \hspace{1cm} (4)

\[ \phi \in [0, \pi[, \ r \in \mathbb{R} : \text{line parameters} \]  \hspace{1cm} (5)
Line Fitting: Line Parametrization and Residuals

- Line parameters: $\phi \in [0, \pi], \ r \in \mathbb{R}$

- Point $x = (x, y)$ on the line:

\[
x \cos \phi + y \sin \phi = r
\]

$\Leftrightarrow x \cdot (\cos \phi, \sin \phi) = r$

- Point $x = (x, y)$ not on the line:

\[
x \cdot (\cos \phi, \sin \phi) \neq r
\]

- Signed distance $\rho(x)$ from line:

\[
\rho(x) = x \cdot (\cos \phi, \sin \phi) - r
\]

Note: $n = (\cos \phi, \sin \phi)$ (thus $\|n\| = 1$)
Line Fitting, Inliers Only: Easy!

Data points
\[ \mathcal{X} = \{ x_j, j = 1, 2, \ldots, N_p \} \]
\[(x_j \in \mathbb{R}^2)\]

Find the line which “best fits” these points.
Line Fitting, Inliers Only: Easy!

Data points
\[ \mathcal{X} = \{x_j, j = 1, 2, \ldots, N_p\} \quad (x_j \in \mathbb{R}^2) \]

Find the line which “best fits” the points.

As optimization: Find best line with parameters \( \theta^* \) as

\[ \theta^* = \arg\min_{\theta} \sum_{x \in \mathcal{X}} f(x, \theta) \]

For \( f_{LSQ}(x, \theta) = [\rho(x)]^2 \) this is easily solvable by Singular Value Decomposition (SVD).
General Case with Outliers, Example 1

Example 1

Least squares fit
Example 2

Least squares fit
General Case with Outliers, Robust Cost Function

- + set of data points

Find:

\[ \theta^* = \arg \min_{\theta} \sum_{x \in X} f(x, \theta) \]

\[ \theta = (r, \phi) \]

- No outliers: \( f_{LSQ}(x, \theta) = [\rho(x)]^2 \)

- Use instead:

\[ f_{RANSAC}(x, \theta) = \begin{cases} 
0, & \text{if } \rho(x) \leq \text{threshold } \sigma \\
\text{const}, & \text{otherwise}
\end{cases} \]

- Such cost function is non-convex

- How to find optimal line parameters?
Random Sample Consensus - RANSAC

Select sample of $m$ points at random (here $m=2$)
Select sample of $m$ points at random

Estimate model parameters from the data in the sample
RANSAC

Select sample of $m$ points at random

Estimate model parameters from the data in the sample

Evaluate the error (residual) for each data point
RANSAC

Select sample of $m$ points at random

Estimate model parameters from the data in the sample

Evaluate the error (residual) for each data point

Select data that support the current hypothesis
RANSAC

Select sample of $m$ points at random

Estimate model parameters from the data in the sample

Evaluate the error (residual) for each data point

Select data that support the current hypothesis

Repeat sampling
RANSAC

Select sample of \( m \) points at random

Estimate model parameters from the data in the sample

Evaluate the error (residual) for each data point

Select data that support the current hypothesis

Repeat sampling
**RANSAC**

Select sample of $m$ points at random

Estimate model parameters from the data in the sample

Evaluate the error (residual) for each data point

Select data that support the current hypothesis

Repeat sampling
RANSAC [Fischler and Bolles 1981]

Input: $\mathcal{X} = \{x_j\}_{j=1}^N$ data points

$e(S) = \theta$ estimates model parameters $\theta$ given sample $S \subseteq \mathcal{X}$

$f(x, \theta) = \begin{cases} 
0, & \text{if distance to model } \leq \text{ threshold } \sigma \\
1, & \text{otherwise}
\end{cases}$

$\Rightarrow J(\theta) = \sum_{x \in \mathcal{X}} f(x, \theta)$ is #outliers

$\eta$ – required confidence in the solution, $\sigma$ – outlier threshold

Output: $\theta^*$ parameter of the model minimizing the cost function

1: $\text{iter} \leftarrow 0, J^* \leftarrow \infty$
2: repeat
3: Select random $S \subseteq \mathcal{X}$ (sample size $m = |S|$) SAMPLING
4: Estimate parameters $\theta = e(S)$
5: Evaluate $J(\theta) = \sum_{x \in \mathcal{X}} f(x, \theta)$ VERIFICATION
6: If $J(\theta) < J^*$ then SO-FAR-THE-BEST
   $\theta^* \leftarrow \theta, J^* \leftarrow J(\theta)$
7: $\text{iter} \leftarrow \text{iter} + 1$
8: until $P(\text{better solution exists}) = f(|\mathcal{X}|, J^*, \text{iter}) < \eta$
9: Compute $\theta^*$ from all inliers $\mathcal{X}_{in}$: $\theta^* \leftarrow \text{LocalOptimization}(\mathcal{X}_{in}, \theta^*)$
RANSAC – how many samples?

- $N$: Number of points
- $Q$: Number of inliers, $Q = N - J^*$
- $m$: Size of sample
- $\frac{Q}{N}$: Inlier ratio

Probability of all-inlier (uncontaminated) sample:

$$P(\text{inlier sample}) = \frac{Q}{N} \approx \epsilon^m$$

Mean time for hitting all-inliers sample is proportional to $1/P$. 
RANSAC – how many samples?

- How about this formulation:
  - Set the number of samples $k$ such that **at least one** pair of points from the line has been hit with probability larger than $\eta$
  - Equivalently … such that **no** pair of points from the line has been hit with probability lower than $1 - \eta$

- $Q$ Number of inliers, $Q = N - J^*$
- $m$ Size of sample
- $\epsilon^2 = \frac{Q}{N}$ Inlier ratio

Probability of all-inlier (uncontaminated) sample:

$$P(\text{inlier sample}) = \frac{\binom{Q}{m}}{\binom{N}{m}} = \frac{Q(Q-1)\ldots(Q-m+1)}{N(N-1)\ldots(N-m+1)} \approx \epsilon^m$$

We require:

$$P(\text{bad pair } k \text{ times}) = (1 - P(\text{inlier sample}))^k < 1 - \eta$$

Finding the solution with confidence $\eta$ therefore requires at least:

$$k \geq \log(1 - \eta) / \log \left(1 - \epsilon^m\right)$$
### RANSAC termination - How many samples?

Inlier ratio $^2 = Q/ N$ [%]

<table>
<thead>
<tr>
<th>Size of the sample $m$</th>
<th>15%</th>
<th>20%</th>
<th>30%</th>
<th>40%</th>
<th>50%</th>
<th>70%</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>132</td>
<td>73</td>
<td>32</td>
<td>17</td>
<td>10</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>5916</td>
<td>1871</td>
<td>368</td>
<td>116</td>
<td>46</td>
<td>11</td>
</tr>
<tr>
<td>7</td>
<td>$1.75 \cdot 10^6$</td>
<td>$2.34 \cdot 10^5$</td>
<td>$1.37 \cdot 10^4$</td>
<td>1827</td>
<td>382</td>
<td>35</td>
</tr>
<tr>
<td>8</td>
<td>$1.17 \cdot 10^7$</td>
<td>$1.17 \cdot 10^6$</td>
<td>$4.57 \cdot 10^4$</td>
<td>4570</td>
<td>765</td>
<td>50</td>
</tr>
<tr>
<td>12</td>
<td>$2.31 \cdot 10^{10}$</td>
<td>$7.31 \cdot 10^8$</td>
<td>$5.64 \cdot 10^6$</td>
<td>$1.79 \cdot 10^5$</td>
<td>$1.23 \cdot 10^4$</td>
<td>215</td>
</tr>
<tr>
<td>18</td>
<td>$2.08 \cdot 10^{15}$</td>
<td>$1.14 \cdot 10^{13}$</td>
<td>$7.73 \cdot 10^9$</td>
<td>$4.36 \cdot 10^7$</td>
<td>$7.85 \cdot 10^5$</td>
<td>1838</td>
</tr>
<tr>
<td>30</td>
<td>$\infty$</td>
<td>$\infty$</td>
<td>$1.35 \cdot 10^{16}$</td>
<td>$2.60 \cdot 10^{12}$</td>
<td>$3.22 \cdot 10^9$</td>
<td>$1.33 \cdot 10^5$</td>
</tr>
<tr>
<td>40</td>
<td>$\infty$</td>
<td>$\infty$</td>
<td>$\infty$</td>
<td>$2.70 \cdot 10^{16}$</td>
<td>$3.29 \cdot 10^{12}$</td>
<td>$4.71 \cdot 10^6$</td>
</tr>
</tbody>
</table>

computed for $\eta = 0.95$
RANSAC – Time Complexity

Repeat $k$ times ($k$ is a function of $\eta$, $Q$, $N$)

1. Hypothesis generation
   - Select a sample of $m$ data points
   - Calculate parameters of the model(s)

2. Model verification
   - Find the support (consensus set) by
   - verifying all $N$ data points

Total running time:

$$t = k \left( t_H + t_V N \right)$$

$$k = \frac{\log(1-\eta)}{\log(1-\epsilon^m)}$$
RANSAC Notes

Pros:
- extremely popular (> 17000 citations in Google Scholar)
- used in many applications
- percentage of inliers not needed and not limited
- a probabilistic guarantee for the solution
- mild assumptions: \( \frac{3}{4} \) known

Cons:
- slow if inlier ratio low
- It was observed experimentally that RANSAC takes several times longer than theoretically expected. This is due to noise – not every all-inlier sample generates a good hypothesis:

\[
P(\text{inlier sample}) \neq P(\text{good model estimate})
\]
RANSAC Issues, Variants

- **Cost function**: MLESAC, Huber loss, …
- **Outlier threshold** $\sigma$: Least median of Squares, MINPRAN, …

- **Correctness of the results. Degeneracy.**
  Solution: DegenSAC.

- **Accuracy** (parameters are estimated from minimal samples).
  Solution: Locally Optimized RANSAC

- **Speed**: Running time grows with
  1. number of data points,
  2. number of iterations (polynomial in the inlier ratio)

  Addressing the problem:
  RANSAC with SPRT (WaldSAC), PROSAC
Locally Optimized RANSAC (LO-RANSAC): Problem Intro

Data: 200 points
LO-RANSAC: Problem Introduction

Data: 200 points
Model, 100 inliers
For simplicity, consider only points belonging to the model (100 points)
For simplicity, consider only points belonging to the model (100 points)

RANSAC

Hypothesis generation from 2 points

Will every two points generate the whole inlier set?

This sample:
YES. 100 inliers.
For simplicity, consider only points belonging to the model (100 points)

RANSAC

Hypothesis generation from 2 points

Will every two points generate the whole inlier set?

This sample:
NO. 45 inliers.
LO-RANSAC: Problem Introduction

For simplicity, consider only points belonging to model (100 points)

RANSAC
Hypothesis generation from 2 points
Will every two points generate the whole inlier set?

The distribution of the number of inliers obtained while randomly sampling points pairs
**LO-RANSAC**

**Input:** \( \mathcal{X} = \{x_j\}_{j=1}^N \) data points

\[ e(S) = \theta \]

estimates model parameters \( \theta \) given sample \( S \subseteq \mathcal{X} \)

\[ f(x, \theta) = \begin{cases} 
0, & \text{if distance to model } \leq \text{ threshold } \sigma \\
1, & \text{otherwise} 
\end{cases} \]

\[ \Rightarrow J(\theta) = \sum_{x \in \mathcal{X}} f(x, \theta) \text{ is } \# \text{outliers} \]

\( \eta \) – required confidence in the solution, \( \sigma \) – outlier threshold

**Output:** \( \theta^* \) parameter of the model minimizing the cost function

1: \( \text{iter} \leftarrow 0, \ J^* \leftarrow \infty \)
2: \( \textbf{repeat} \)
3: Select random \( S \subseteq \mathcal{X} \) (sample size \( m = |S| \))
4: Estimate parameters \( \theta = e(S) \)
5: Evaluate \( J(\theta) = \sum_{x \in \mathcal{X}} f(x, \theta) \)
6: If \( J(\theta) < J^* \) then
   \[ \theta^* \leftarrow \theta, \ J^* \leftarrow J(\theta) \]
7: \( \text{iter} \leftarrow \text{iter} + 1 \)
8: \( \textbf{until} \ P(\text{better solution exists}) = f(|\mathcal{X}|, J^*, \text{iter}) < \eta \)
9: Compute \( \theta^* \) from all inliers \( \mathcal{X}_{in} \): \( \theta^* \leftarrow \text{LocalOptimization}(\mathcal{X}_{in}, \theta^*) \)

12/03/2018

J. Matas, MPV Course

30/93
LO-RANSAC

Input: $\mathcal{X} = \{x_j\}_{j=1}^N$ data points

$e(S) = \theta$ estimates model parameters $\theta$ given sample $S \subseteq \mathcal{X}$

$f(x, \theta) = \begin{cases} 0, & \text{if distance to model } \leq \text{ threshold } \sigma \\ 1, & \text{otherwise} \end{cases}$

$\Rightarrow J(\theta) = \sum_{x \in \mathcal{X}} f(x, \theta)$ is #outliers

$\eta$ – required confidence in the solution, $\sigma$ – outlier threshold

Output: $\theta^*$ parameter of the model minimizing the cost function

1: $\text{iter} \leftarrow 0$, $J^* \leftarrow \infty$
2: repeat
3: Select random $S \subseteq \mathcal{X}$ (sample size $m = |S|$)
4: Estimate parameters $\theta = e(S)$
5: Evaluate $J(\theta) = \sum_{x \in \mathcal{X}} f(x, \theta)$
6: If $J(\theta) < J^*$ then
   $\theta^* \leftarrow \text{LocalOptimization}(\mathcal{X}_{in}, \theta)$, $J^* \leftarrow J(\theta^*)$
7: $\text{iter} \leftarrow \text{iter} + 1$
8: until $P(\text{better solution exists}) = f(|\mathcal{X}|, J^*, \text{iter}) < \eta$
9: gone

12/03/2018 J. Matas, MPV Course 31/93
LO-RANSAC: Example

inliers count = 60
LO-RANSAC: Example

inliers count = 60

Init
Iteration 1
LO-RANSAC: Example

inliers count = 59

Init
Iteration 1
Iteration 2
LO-RANSAC: Example

inliers count = 67

Init
Iteration 1
Iteration 2
...
Iteration 7
LO-RANSAC: Example

Init
Iteration 1
Iteration 2
...
Iteration 7
...
Iteration 15

inliers count = 97
LO-RANSAC: Example

Comparison with model (100 inliers):
Locally Optimized RANSAC

Estimation of (approximate) models with lower complexity (less data points in the sample) followed by LO step estimating the desired model speeds the estimation up significantly.

The estimation of epipolar geometry is up to 10000 times faster when using 3 region-to-region correspondences rather than 7 point-to-point correspondences.

Simultaneous estimation of radial distortion and epipolar geometry with LO is superior to the state-of-the-art in both speed and precision of the model.

Fish-eye images by Braňo Mičušík

Chum, Matas, Obdržálek: Enhancing RANSAC by Generalized Model Optimization, ACCV 2004
It was observed experimentally that RANSAC takes several times longer than theoretically expected. This is due to the noise – not every all-inlier sample generates a good hypothesis.

By applying local optimization (LO) to the-best-so-far hypotheses:
(i) a near perfect agreement with theoretical performance
(ii) lower sensitivity to noise and poor conditioning.

The LO is shown to be executed so rarely, \( \log(\text{iter}) \) times, that it has minimal impact on the execution time.