COSC342 Tutorial

RANSAC

Random Sample and Consensus is a method for dealing with outliers. In this tutorial we'll look at an example of RANSAC for fitting a line, using dice to generate random numbers and an Excel spreadsheet to help visualise what is going on.

1 RANSAC Basics

Often we have a set of data we want to fit a model to, such as fitting a homography to a set of corresponding points between two images. In this tutorial we'll use a simpler example – fitting a straight line to a set of (x, y) co-ordinates.

Real measurements are noisy, so often we make a least squares fit. Suppose we have a set of points, $(x_1, y_1), (x_2, y_2), \ldots, (x_k, y_k)$, and we want to fit a line to this data. The basic formula for a line is

$$y = mx + c,$$

and a least-squares fit finds values for m and c that minimises the sum of squared errors between the measured y values and those predicted by the model,

$$\epsilon = \sum_{i=1}^{k} \|y_i - (mx_i + c)\|^2.$$

If our measurements are basically correct, but have small (and independent) errors then this usually works just fine. However, if some of our measurements are fundamentally wrong (rather than just inaccurate) things go awry, as shown in Figure 1. These wrong measurements are called *outliers*, and the correct ones are called *inliers*.

A simple solution to this problem is called Random Sample and Consensus, or RANSAC. RANSAC works by repeatedly picking a small sample of the points at random and fitting the model to them. The basic idea is that eventually we'll pick a sample set that doesn't contain outliers. This is the 'Random Sample' part of the algorithm.

The 'Consensus' part of the algorithm allows us to determine when we have a good sample. We look at all of the data points, and see how many of them more-or-less agree with the model. This is known as the *consensus set*, and we use the sample with the largest consensus set as our solution.

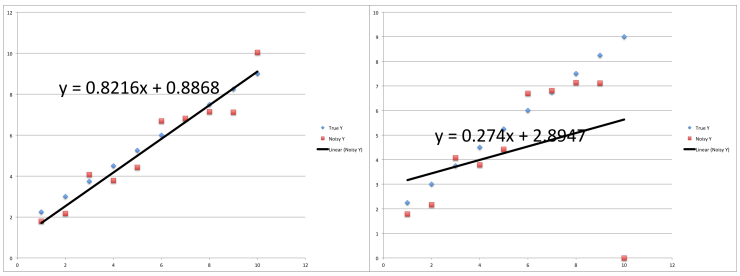

Figure 1: We often can't observe the true data points (blue), but instead make uncertain measurements (red). If this uncertainty is just small random errors, then a least-squares fit works well (left), however, if some measurements are outliers the result can be badly skewed (right).

For the example of fitting a line to some data, we only need 2 points to make a line. We can keep on picking 2 points at random, until we find a line that agrees with most of the points.

2 The Spreadsheet

The Excel spreadsheet has 36 points in it, numbered so that you can generate random points with two rolls of a 6-sided die. For example, rolls of 5 then 4 would be point 54, which is row 29 in the spreadsheet. The points have random x values from 0 to 10 and the true model that we are seeking is $y = 1.5x + 1.5$. The actual measurements we make have some noise added, which is Gaussian noise with mean 0 and variance 1. 25% of the points are outliers, and have measurements which are uniformly and randomly chosen from the range -15 to 15. As a result, the best fit line to the data is skewed slightly.

The Excel spreadsheet automatically computes a lot of the RANSAC properties we need. The only thing that needs to be edited is the 'Sample' column. You should put a 1 and a 2 in this column to indicate the chosen random sample. **Note:** The 1 must go above the 2 in the sample column for Excel's `LOOKUP` function to work.

1. Put a 1 by the first point (Point 11) and a 2 by the last point (Point 66). These are both inliers so you should get a good fit. In particular:
   - The Red Squares are the sample, so you can see what points have been chosen.
   - The Green Triangles are the inliers, and the sample line is shown in green also. A point is an inlier if it is within 2 units (standard deviations) of the line.
   - Below the chart there are cells giving the number of inliers.

2. Put a 1 by the second and third points (Points 12 and 13), which are both inliers. Observe what happens to the fitting line and the number of inliers.

3 How Many Trials?

We could just do lots of trials, and pick the one with the largest number of inliers. However, it is hard to know how many trials to do. An alternative is to pick some probability of success that we're happy to accept, call this p. Typically we set p quite high, say $p = 0.99$, indicating we want to be correct 99% of the time. From this we can figure out how many trials to do assuming we know the number of inliers.

Suppose we have a set of N items, n of which are inliers. To get an outlier-free sample we need to carry out t trials where

$$t \approx \frac{\log(1-p)}{\log\left(1-\left(\frac{n}{N}\right)^s\right)}$$

We can derive this value of t using the following steps

1. If I choose one of the N items at random, what is the probability that it is an inlier?

2. Suppose I need to sample s points to fit my model. For example, to fit a line $s = 2$. If I choose s points at random, what is the probability that they are all inliers?

3. If I choose s points at random, what is the probability that at least one is an outlier (i.e. not all are inliers)?

4. Now suppose that I do t trials, each time choosing s points at random. What is the probability that all of the trials contain at least one outlier?

5. This is the probability that we don't have a correct solution after t trials, so we can set this equal to $1 - p$ and solve for t. To do this you will need to use logarithms, making use of the fact that

$$\log(a^b) = b\log(a).$$

This formula is a little tricky to compute by hand, but is easy enough in the computer. For our line fitting example, $N = 36$, $s = 2$, and we can take $p = 0.99$. Table 3 gives values of t for the possible values of n. Since it is not possible to do 'half a trial', t has been rounded up to the next highest integer.

The final complication is that we don't know how many inliers there are, so don't know n. We overcome this by starting off by assuming that n is very small, say $n = s$. In the case of the line fitting, this means we'll need to do 1490 trials.

Whenever we do a trial and find out that it has more than n inliers, we update n to this new value. Suppose we pick two points at random, and get

Table 1: Number of trials needed for a given number, n, of inliers.

n	1	2	3	4	5	6	7	8	9	10	11	12
t	5966	1490	661	371	237	164	120	91	72	58	47	40
n	13	14	15	16	17	18	19	20	21	22	23	24
t	33	29	25	21	19	17	15	13	12	10	9	8
n	25	26	27	28	29	30	31	32	33	34	35	36
t	7	7	6	5	5	4	4	3	3	3	2	1

12 inliers. This means that we can find at least 12 points that roughly lie on a line, and so n is at least 12. Looking up Table 3 (or, more generally, applying the formula) gives us a new value of 40 trials.

As we find larger inlier sets, the number of trials required drops. Eventually we will have done more trials than we need to, and can return the best result as the outcome.

4 RANSAC Line Fitting

To perform RANSAC we need to keep track of

- The number of trials we have done
- The total number of trials we need to do
- The size of the largest inlier set, and the formula of the best-fit line

1. Generate two points at random, and observe how many points are in the consensus set.
2. If this consensus set is the biggest you've seen:
   - Update the biggest consensus set size
   - Update the number of trials to be done
   - Fit a least-squares line to the inlier points as the best model
3. If you've done at least as many trials as you need to then stop, otherwise go to Step (1).