Random Sample Consensus (RANSAC)

H. Cantzler
Institute for Perception, Action and Behaviour,
Division of Informatics, University of Edinburgh,
Edinburgh, EH1 2QL, UK
helmutc@dai.ed.ac.uk

The RANSAC algorithm is an algorithm for robust fitting of models. It was introduced by Fischler and Bolles in 1981 [2]. It is robust in the sense of good tolerance to outliers in the experimental data. It is capable of interpreting and smoothing data containing a significant percentage of gross errors. The estimate is only correct with a certain probability, since RANSAC is a randomised estimator. The algorithm has been applied to a wide range of model parameters estimation problems in computer vision, such as feature matching, registration or detection of geometric primitives.

1 Subsampling of the input data

The structure of the RANSAC algorithm is simple but powerful. First, samples are drawn uniformly and at random from the input data set. Each point has the same probability of selection (uniform point sampling). For each sample a model hypothesis is constructed by computing the model parameters using the sample data. The size of the sample depends on the model one wants to find. Typically, it is the smallest size sufficient to determine the model parameters. For example, to find circles in the data set, one has to draw three points, since three points are required to determine the parameters of a circle. Drawing more than the minimal number of sample points is inefficient, since the probability of selecting a sample consisting only of inlying data points (*i.e.* all data points belonging to the same model), that gives a good estimate and at random, decreases with respect to the increas-

ing sample size. Thus the minimal sample set maximises the probability of selecting a set of inliers from which later a good estimate will be computed.

2 Hypotheses evaluation

In the next step, the quality of the hypothetical models is evaluated on the full data set. A cost function computes the quality of the model. A common function is to count the number of inliers (*i.e.* data points which agree with the model within an error tolerance). The hypothesis which gets the most support from the data set gives the best estimate. Typically, the model parameters estimated by RANSAC are not very precise. Therefore, the estimated model parameters are recomputed by for example a least-squares fit to the data subset which supports the best estimate. The input data may support several distinct models. In this case, the model parameters for the first model are estimated, the data points supporting the model are removed from the input data and the algorithm is simply repeated with the remainder of the data set to find the next best model. The strength of the algorithm is that it is likely to draw at least one set of points which consists only of inliers and thus results in a good estimate of the model parameters.

3 Process variables

The RANSAC technique uses three variables to control the model estimation process. The first determines whether or not a data point agrees with a model. Typically, this is some error tolerance that determines a volume within which all compatible points must fall in. The number of model hypotheses that are generated is the second variable. It depends on the probability to draw a sample including only inlying data points. As the proportion of outliers and the minimal sample set size increase the number of model hypotheses must be increased to obtain a good estimate of the model parameters. The proportion of outliers depends on the noise level and on how many models are supported by the data set. Furthermore, one tolerance variable is needed to determine if a correct model has been found. An extracted model is deemed valid if there is sufficient support from the data points for this model. A valid circle has been found in the data if for example at least 20 data points are found which lie close enough to the circle. In case of multi-

ple models in the data, more models are extracted until there is insufficient support for any more models.

4 Runtime improvements

The computational efficiency of the algorithm can be improved significantly in several ways. The speed depends on two factors: firstly, the number of samples which have to be drawn to guarantee a certain confidence to obtain a good estimate; and secondly, the time spent evaluating the quality of each hypothetical model. The latter is proportional to the size of the data set.

Typically, a very large number of hypotheses are created from contaminated samples (*i.e.* samples containing outliers). Such models are consistent with only a small fraction of the data. The evaluation of the models can be computationally optimised by randomising the evaluation [1]. Every hypothetical model is first tested only with a small number of random data points from the data set. If a model does not get enough support from this random point set, then one can assume with a high confidence that the model is not a good estimate. Models passing the randomised evaluation are then evaluated on the full data set.

The performance of the algorithm degrades with increasing sample size or in case multiple models are supported by the data due to the decreasing probability of sampling a set that is composed entirely of inliers. A common observation is that outliers possess a diffuse distribution. In contrast, inliers will tend to be located closely together. Therefore, the uniform sampling of points is replaced by selection of sample sets based on proximity taking spatial relationships into account [3]. The first initial sample point is selected randomly. The rest of the points are random points lying within a hypersphere centred on the first point. The selection of sample sets of adjacent points can significantly improve the probability of selecting a set of inlying points and thus drastically reduce the number of samples required to find a good model estimate.

References

[1] O. Chum and J. Matas. Randomized ransac with t(d,d) test. *British Machine Vision Conference, Cardiff, UK*, pages 448–457, 2002.

[2] M.A. Fischler and R.C. Bolles. Random sample consensus: A paradigm for model fitting with application to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[3] D.R. Myatt, P.H.S. Torr, S.J. Nasuto, J.M. Bishop, and R. Craddock. Napsac: High noise, high dimensional robust estimation - it's in the bag. *British Machine Vision Conference, Cardiff, UK*, pages 458–467, 2002.