Fusing 100's of 3D Point Clouds of Objects

R. B. Fisher, S. McDonagh
University of Edinburgh
rbf@inf.ed.ac.uk, s.g.mcdonagh@sms.ed.ac.uk

This abstract presents an approach to fusing hundreds of 3D point clouds to make complete models of 3D objects[1]. This topic first arose in the mid-1990s, but has now become particularly relevant because of the availability of consumer video-rate 3D sensors, such as the Kinect. These sensors can easily acquire hundreds or thousands of 3D images of objects or scenes in a few seconds, which has led researchers to ask again how to fuse the individual 3D datasets.

This project focused on fusing hundreds of views of the exterior of an object, in contrast to hundreds of views of a scene, as in the KinectFusion [2] or SLAM approaches. The core motivating problem was how to manage registration error. For example, incremental image to image registration allows the registration errors to accumulate, and so the shape gradually distorts. The SLAM approach reduces the error by 'loop closure'.

The intuition was that one could create an alignment error minimization approach, whereby each individual scan minimized its registration error with all of the other scans. As each scan was attempting to do this independently, with the right registration error formulation, the scans would settle into a minimum error position.

Previous approaches have incorporated similar alignment update schemes wherein transforms and point correspondences are alternatively optimized while keeping the other fixed ([3], [4], [5], [6]). By alternating the update of view transforms and error estimation, the two solutions tend to mutually improve one another during the process and converge to appropriate solutions.

1. Density estimation for large view set fusion

The proposed method used a non-parametric kernel density estimation to formulate the registration error across hundreds of views. A density function was defined, reflecting the likelihood that a point sample $\mathbf{x} \in \mathbb{R}^3$ lies on the unknown true surface $\mathcal{S}$ as observed in the view collection through point samples $\mathcal{P}$. This surface estimate was then used to guide view registration in the sensor transform space as view pose positions and model surface estimate are alternatively refined. Surface estimation naturally extends to arbitrarily large view counts and makes no assumptions about the form of $\mathcal{S}$. The key idea involves independently looking at each point x at which we want to estimate the density and then deciding which of the available local data observations should contribute to this estimation.

In practice a least-squares plane was fitted to the points in the spatial neighborhood of $\mathbf{x}$ as dictated by the kernel bandwidth distance h. Plane fitting utilized a simple reciprocal Euclidean distance weighting, providing a monotonically decreasing weight function based on spatial distance. If point normals were provided by the scanning device these could be used instead of fitted estimates. Estimation point $\mathbf{x}$ could then be orthogonally projected onto this plane to find the local registration error contribution of the point sample. Object surface structure can be considered locally planar for sufficiently close proximity and surface points in well registered positions will therefore lie on or near locally planar regions.

The common kernel density estimation assumption that the influence of contributing samples diminishes with increasing distance was incorporated using monotonically decreasing weight functions that reduced point influence as distance increased. The second kernel-component followed [9] and made use of a trivariate anisotropic Gaussian function, adapted to the local sample distribution for this task. This allowed the kernel *shape* to adapt to the local point distribution.

In summary, density estimates provide a means to infer where physical surfaces exist that (1) improve in confidence with additional data and more views, (2) have an intrinsic ability to account for outliers and sample measurement noise and (3) provide typically smooth gradients for an iterative pose optimization process.

2. Adaptive bandwidths for estimating surfaces from point clouds

Bandwidth parameters, dictating kernel width, are often fixed. In regions of high data density, a large value of h may lead to over-smoothing and a washing out of structure that might otherwise be extracted. However, reducing h may lead to noisy estimates elsewhere in the data space

where the density is smaller. Thus the optimal choice for h is dependent on location within the data space. Adaptively defining a unique bandwidth for each kernel addresses this problem.

In this work adaptive kernels were instantiated using *balloon*-like estimators that make use of nearest-neighboring data samples. The k-nearest-neighbor kernel density estimate provides a kernel bandwidth defined as the Euclidean distance between the query point $\mathbf{x}$ and the k-th nearest-neighbor of point $\mathbf{x}$ among the available point samples. See [7] for further detail.

Adaptive bandwidth control is motivated by the observation that large multi-view registration tasks commonly contain data sets that exhibit varying levels of measurement redundancy in surface sampling locations. Physical surface areas may be sampled at varying densities (eg. areas of interest may be intentionally sampled with higher density to improve accuracy). In this problem domain, washing out of structure tends to manifest as over-smoothing of distinctive surface features and detail that might prove useful during the registration process. Over-reduction of bandwidth terms can conversely result in fitting (and fabricating) unwanted surface structure to small outlying depth measurements caused by sensor noise. In practice this method of bandwidth selection was advantageous in conjunction with kernel construction as the *shape* and *size* of local kernels was influenced by the neighboring point sample observations.

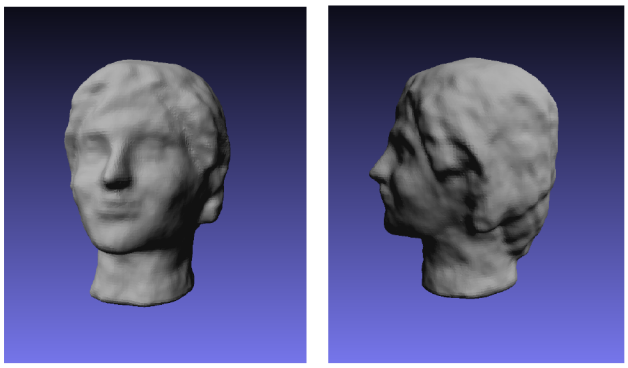
Sensor views can be aligned to this surface approximation essentially defining an implicit *soft* correspondence between sample points.

Kernel bandwidth size initially starts large enough to ensure that all views fuse without cliques, and bandwidth is later reduced to produce tighter fusion. As spatial transform iterations tighten the view alignment, k-nearest-neighbor kernel bandwidth sizes naturally reduce as neighbor point distances reduce in correlation with tighter view fusion and view sample locations. The technique is therefore capable of fitting surface structure to emerging object detail as viewpoints move into positions of better registration.

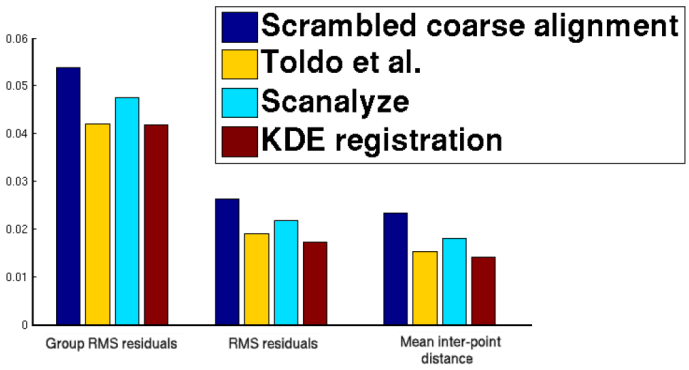
Each scan found the pose that minimized the sum of this density estimate error over all of its data points. To make the approach practical, the algorithm alternates between independently optimizing each scan's position relative to the other pointsets in the previous iteration, and then simultaneously updating all of the pointset positions.

An example of fusion from 220 views can be seen here. This is an example intensity image and associated depth image:

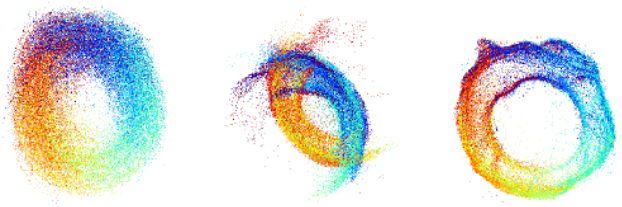
and here are two views of the surface reconstructed from the fusion of 220 views:



We compute standard RMS residual point pair and mean inter-point (IPD) distances of the converged alignment poses. The RMS residuals are computed as the root mean square distances between the points of every view and the single closest neighboring point from any of the other $M-1$ views. This gives a measure of the compactness of the multi-view registration. For the collection of M views, the additional Group RMS metric forces each sample point to identify the closest neighboring point in every other viewpoint. This allocates $M-1$ distance values to each sample point in the combined data set. This secondary RMS measure is useful in addition to the first as it penalizes the previously discussed view clique problem where scans may exhibit good local registration yet poor inter-clique registration. This metric attempts to provide an evaluation measure of how tightly a group of viewpoints has been registered. Experiments with datasets with 220 to 512 views show successful fusion, and substantially better performance than sequential Iterated Closest Point and the approaches of Toldo [8], Pulli [10].



Although the statistics for the Toldo approach are only slightly worse than ours, we can see in the figure below (left) the initial coarse alignment as seen from above, (middle) the result of Toldo's algorithm and (right) the result from our algorithm.



The prototype Matlab implementation requires considerable computation and memory, limiting the total number of data points to a few million. On the other hand, the approach allows an easy semi-synchronized parallel implementation.

References

[1] S. McDonagh. Building models from multiple point sets with kernel density estimation. PhD Thesis, University of Edinburgh, 2015.

[2] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman and A. Davison. KinectFusion: real-time 3D reconstruction and interaction using a moving depth camera. Proc. 24th Ann. Symp. on User interface software and technology, 2011.

[3] P. Besl and N. McKay. A method for registration of 3-D shapes. IEEE. Trans. on Pattern Analysis and Machine Intelligence, 1992.

[4] J. Feldmar and N. Ayache. Rigid, affine and locally affine registration of free-form surfaces. International Journal of Computer Vision, 1996.

[5] W. M. Wells. Statistical approaches to feature-based object recognition. International Journal of Computer Vision, 1997.

[6] H. Chui and A. Rangarajan. A new algorithm for non-rigid point matching. Computer Vision and Pattern Recognition, 2000.

[7] S. McDonagh, R. B. Fisher. Simultaneous registration of multi-view range images with adaptive kernel density estimation. Proc. IMA 14th Mathematics of Surfaces, 2013.

[8] R. Toldo, A. Beinat, F. Crosilla. Global registration of multiple point clouds embedding the Generalized Procrustes Analysis into an ICP framework. Proc. Int. Symp. 3D Data Processing, Visualization and Transmission (3DPVT), 2010.

[9] O. Schall, A. Belyaev, H. Seidel. Robust filtering of noisy scattered point data. IEEE Eurographics Symposium on Point-Based Graphics, 2006.

[10] K. Pulli. Multi-view registration for large data sets. 3D Digital Imaging and Modelling, 1999.