Analysis of 3D Point Clouds using a Parallel DBSCAN Clustering Algorithm

For decades now, scientists have collected huge amounts of data to be analyzed. Machine learning algorithms, which find important information in the data, have become universal tools in data science today. Still, analyzing large and high-dimensional data collections exceeds the capability of default machine leaning implementations on standard computers. The High Productivity Data Processing Research Group at the Forschungszentrum Jülich (FZJ) works on parallel and scalable machine leaning software. This enables a data analysis that is able to leverage the powerful capabilities of modern High Performance Computing (HPC) environments. Driven by the needs of scientific users, their newest parallel implementation of a clustering algorithm, named HPDBSCAN, has reached the state-of-the-art performance in terms of memory usage and speed up.

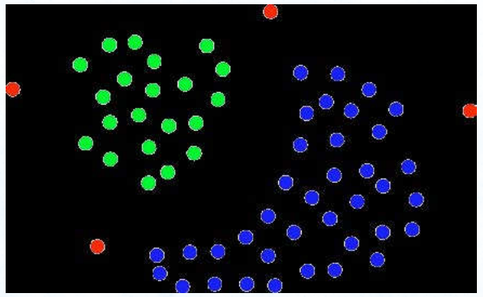

Figure 1: Example of points clustered by DBSCAN. The algorithm found two clusters (blue and green) and 4 noise points (red).

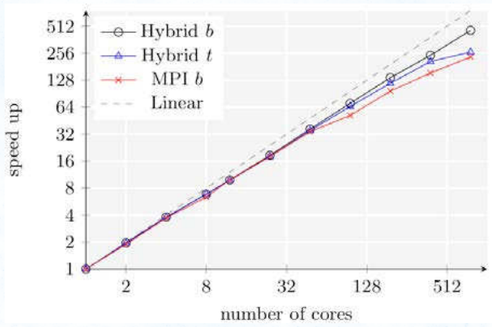

Figure 2: Speed up of HPDBSCAN, tested on different data sets: - b (Bremen data set), t (Twitter data set).

DBSCAN

DBSCAN – or density based spatial clustering for applications with noise – is the original serial clustering algorithm formulated in 1996 by [1] et al. at the University of Munich. Over the years it became, according to Microsoft Research, the most cited machine learning algorithm [2]. Its core idea is rather simple. While iterating through a dataset, the algorithm tries to find dense areas, which are defined based on the number of neighboring points. These form cluster cores, which, through recursive expansion, are enlarged in the process. Points that cannot be assigned in that fashion are considered to be noise within the dataset.

Parallelization Strategy

The parallelization strategy of the algorithm entails a divide-and-conquer approach. This means that each parallel processor locally clusters a subset of the data, which it then merges with its spatial neighbors. The biggest challenges in the parallelization required the spatial decomposition, the load balancing of skewed datasets, as well as the lock-free and communication-optimized merging. Using these techniques, the group was able to achieve scalable performance (see Fig. 2) outperforming previous parallelization attempts of DBSCAN by an order of magnitude in terms of computation time and memory consumption using different datasets [2]. This Highly Parallel HPDBSCAN is implemented as Hybrid MPI/OpenMP application and uses HDF5 files for parallel I/O.

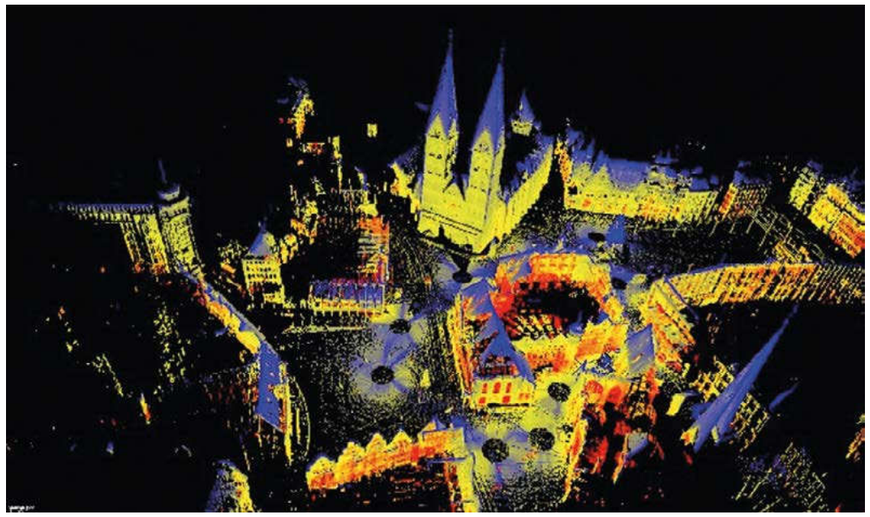

Figure 3: Point cloud of Bremen's old town.

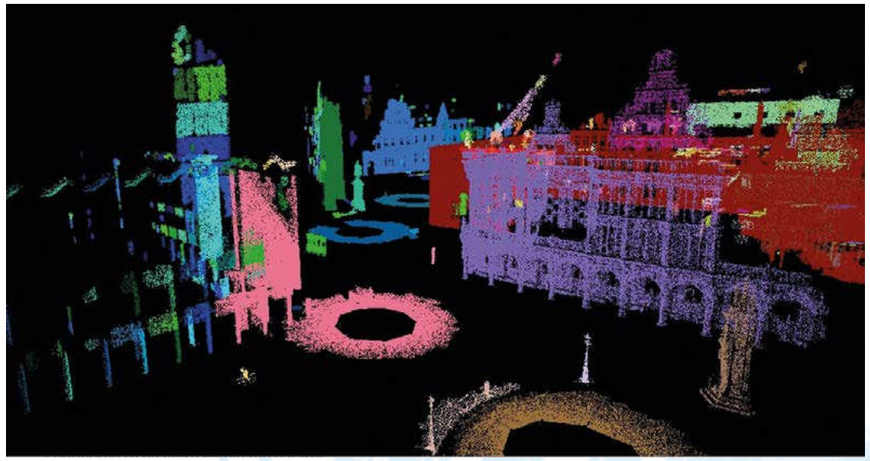

Figure 4: Denoised and clustered point cloud of the old city of Bremen. Each color represents another cluster.

Point Cloud Analysis

One of the application domains where DBSCAN can be used is point cloud analysis. A point cloud is set of three-dimensional points that represent an object or environment. These point clouds are taken by specialized cameras and are comparable to an image in 3D.

Figure 3 shows an example using the Bremen data [3]. Using these point clouds, engineers can reconstruct and model the scanned objects and subsequently use them to e.g. search for leaks, deformations and so forth in the object. This is used in, e.g., industry plant safety monitoring, automatic map creations and large-scale archeological excavations. HPDBSCAN can be used for two tasks supporting research questions. First, it can be used to denoise the point clouds from, e.g., false readings, especially those from small objects moving through the scenery. In a second step, it can be used to cluster together these points in order to identify individual objects and distinguish them from others in an automated fashion. These segmented objects can then be classified either using human experts or other machine learning algorithms. An example of such a point cloud analysis step, based on data shown in Figure 3, is presented below in Figure 4 where the old town of Bremen has been scanned for thermal leakage.

Summary and Outlook

HPDBSCAN is a highly scalable implementation of a widely used clustering algorithm called DBSCAN. It is open source and be obtained through source code repositories and compiled. Currently, HPDBSCAN is being deployed on XSEDE resources and evaluated for its permanent installation value. In the future, we will be using HPDBSCAN is foreseen to be used from partners from the Netherlands, the UK and France in order to do archeological data analysis of roman ruins. HPDBSCAN is also actively used in another research project that is worked on in collaboration with the University of Gothenburg, where we automatically detect water mixing events in the Koljoefjords in Sweden. A wider collection of other scalable machine learning algorithms, by the name of JuML, is currently under development, and is going to be available soon as an open-source package.

References

[1] **Ester, M. e. a.**
A density-based algorithm for discovering clusters in large spatial databases with noise, Kdd. Vol. 96. No. 34, 1996

[2] **Patwary, M., Palsetia, D., Agrawal, A., Liao, W.-k., Manne, F., Choudhary, A.**
A new scalable parallel DBSCAN algorithm using the disjoin-set data structure, in High-Performance Computing, Networking, Storage and Analysis (SC), 2012

[3] **Borrmann D., Nüchter, A.**
Robotic 3D Scan Repository, 18 03 2015. [Online]. Available: http://kos.informatik.uni-osnabrueck.de/3Dscans

contact:
Christian Bodenstein,
c.bodenstein@fz-juelich.de
Markus Götz,
m.goetz@fz-juelich.de
Morris Riedel,
m.riedel@fz-juelich.de

- Christian Bodenstein
- Markus Götz
- Morris Riedel

Jülich Supercomputing Centre (JSC), Germany

Directing the Morphology of amphiphilic Molecules

Amphiphilic molecules contain at least two structural units that thermodynamically repel each other. Since the two incompatible blocks are covalently bonded into a single molecule, they cannot macroscopically phase separated but, instead, self-assemble into spatially modulated structures whose characteristic length scale is dictated by the molecular extension. Typical examples include the self-assembly of lipid molecules, which are comprised of a hydrophilic, polar head and a hydrophobic tail, into bilayer membranes or synthetic block copolymers, which are comprised of two incompatible flexible chain molecules that are joined at their ends and self-assemble into periodic microphases. Despite the differences in the chemical nature of the constituents and the type of interactions, both – biologically relevant lipids as well as synthetic block copolymers – spontaneously form similar structures (e.g., lamellar sheets or wormlike micelles). The structure formation is dictated by the universal competition between the free-energy cost of the interface between the incompatible components and the entropy loss of arranging these molecules uniformly in space.

This delicate balance gives rise to minuscule free-energy differences between different morphologies (on the order of a fraction of the thermal energy scale kT per molecule), and there exist many competing metastable structures (alternate periodic arrangements, defect structures like dislocations, or localized structures like hydrophobic bridges between lipid membranes). This feature is corroborated by the protracted annealing times required to observe well-ordered morphologies in block copolymers or the requirement of specialized proteins that provide the free energy required to overcome the barrier in e.g., pore formation, fusion and fission of membranes. In fact, the complex, rugged free-energy landscape of self-assembling amphiphiles has been likened to that of glass-forming materials. The morphology often does not reach the thermodynamically stable state of lowest free energy but, instead, becomes trapped in a metastable state. By exploring these metastable states and the free-energy barriers that separate them, one can either reproducibly trap the system in desired non-equilibrium morphologies [1] or accelerate equilibration of block copolymer structures [2] or control collective changes of membrane topology involved in cellular transport processes [3,4].

Morphological transformations of amphiphilic, soft-matter systems involve the cooperative rearrangement of many molecules on time and length scales ranging from milliseconds to minutes and nanometers to micrometers for lipids and polymers, respectively. These scales are challengingly small for experimental imaging techniques yet too large for atomistic modeling. Since the transformations often involve highly bent interfaces or strongly stretched molecular conformations, also phenomenological continuum models cannot accurately capture them. In turn, coarse-grained models that only incorporate the relevant degrees of freedom are well suited to explore the universal behavior of amphiphilic structure formation and provide direct insights into