Project number IST-25582

CGL

Computational Geometric Learning

D3.1: Work Package 3 [Period 1] Report

STREP

Information Society Technologies

Period covered:	November 1, 2010–October 31, 2011
Date of preparation:	October 30, 2011
Date of revision:	
Start date of project:	November 1, 2010
Duration:	3 years
Project coordinator name:	Joachim Giesen (FSU)
Project coordinator organisation:	Friedrich-Schiller-Universität Jena Jena, Germany

In the following we describe the work done within Work Package 3 within the first period. In the description we follow the structure imposed by the tasks for this work package and period. Wherever there is no deviation and all goals have been met, it is *not* mentioned specifically. We start by restating the objectives for Work Package 3 and conclude with a discussion of the milestones for Period 1.

Objectives

Data sets found in Science and Engineering are often large and complex in the sense that they encode subtle correlations between the variables describing the system of interest. One way to apprehend this complexity consists of decomposing the data into appropriate geometric structures, so as to process these individually in order to unravel these complex interactions. This work-package follows this spirit for three application domains.

Work Package 3.1 is concerned with the investigation of specific hyper-surfaces called landscapes, which are height functions parameterized above high-dimensional spaces. Landscapes are central objects in optimization, and in Sciences for modeling macro-molecules and clusters of atoms. In both realms, one generally has a partial knowledge of the landscape through a point cloud sampled on it. While the landscape topology is trivial, learning features such as the significant minima and their basins, together with the connexions between them raises challenging questions. The main goals in Work Package 3.1 will be to develop, in conjunction with Work package 1, novel multi-scale analysis, partitioning and dimensionality reduction techniques for landscapes. Validation-wise, we expect these contributions to provide novel insights for analyzing and exploring landscapes in optimization and biophysics.

Work Package 3.2 investigates configuration spaces (C-spaces, for short), which are fundamental tools for studying a large variety of systems, and in particular moving objects (robots) evolving amidst obstacles, and molecular systems. While tools for accurate representation of low (2,3) dimensional C-spaces exist, current machinery for higher-dimensional C-spaces often offers only rather crude approximation, significantly limiting their applicability. Our major objective in Work Package 3.2 will be to develop methods for accurately and efficiently representing higher-dimensional C-spaces. We will specifically experiment with the newly developed methods on problems in automatic motion planning and molecular motion simulation.

Finally, Work Package 3.3 aims at, in collaboration with cosmologists, improving models and algorithms to analyze the distribution of matter in the universe and its evolution under the action of gravity, in particular in conjunction with the formation of galaxies. Continuing ongoing co-operation with cosmologists in this area, the geometric inference methods developed in Work Package 1 will be applied to detect these global geometric structures by analyzing cosmological data sets.

Tasks

Task 1.a: Multi-scale analysis of landscapes

The main goal of this task is to perform a multi-scale analysis of a sampled energy landscape, with two types of contributions in mind: a partial construction of the Morse-Smale complex of the height function (the height corresponding to the energy of the system), and a calculation of all transition paths (a path connecting two local minima across an index one saddle).

During this first period, we focused on the latter aspect for two reasons. On the one hand, all transition paths form a subset of the Morse-Smale complex. On the other hand, previous work on transition paths was calling for improvements in two directions. The first goal was to go beyond the stable mode-seeking algorithm developed in high-dimensional geometric data analysis, as this algorithm only identifies selected minima and index one critical points, and selected paths joining them. The second goal was to go beyond the so-called disconnectivity graph developed in biophysics, as it fails to provide a multi-scale analysis of a landscape, and does not encode all the transition paths either.

Both goals have been achieved since we now have:

- an algorithm computing all the transition paths of a sampled landscape;
- an algorithm performing a multi-scale analysis of the landscape, and also maintaining stable transition paths while simplifying the landscape using topological persistence.

The theoretical contributions just outlined are important milestones to reach the main goal. Yet, we need to provide experimental evidence that these concepts and the accompanying algorithms are amenable to effective calculations on high-dimensional data, and that they provide new insights on specific systems. This is ongoing work.

Task 1.b: Dimensionality reduction techniques for landscapes

No work proposed in Period 1.

Task 1.c: Validations in biophysics and optimization

The goal of this task is to apply the algorithms mentioned in Task 1a to macro-molecular systems, and also to energy landscapes encountered in optimization. During the first period, intensive software developments were undertaken to meet these two goals.

In biophysics, setting up an experimental platform has been especially challenging, as we had to learn and choose a simulation system, and to automate a number of tasks. Having chosen Gromacs, see `http://www.gromacs.org/`, we developed three programs completely automating the generation of molecular dynamics simulations. These programs are described in our CGL technical report CGL-TR-07 [2].

In computer science, we had to develop from scratch a C++ environment to analyze sampled energy landscapes. In particular, we now have classes to:

- represent a sampled energy landscape.
- compute nearest neighbor graphs (NNG) on such landscapes, be they exact or approximate. For samples representing Euclidean points, NNG are built using the Euclidean distance. For samples representing molecular conformations, the NNG are built using the so-called least root-mean-square deviation. (We note in passing that the computation of NNG may yield interactions with Work Package 2, see Task 1c.)
- perform the geometric and topological analysis mentioned in Task 1a.

These developments follow the CGAL spirit, that is, each class is parameterized by a traits class providing full flexibility.

The software developed so far provides the basis for the experiments currently being run.

In biophysics, numerical experiments are conducted in collaboration with Charles Robert, a specialist of MD simulations working with the Institut de Biologie Physico-chimique, in Paris.

In optimization, ongoing analysis are performed in collaboration with Christian Mueller, from ETH, who is also involved in CGL.

We are in line with the proposed work. We have started intensive tests, both on data describing molecular systems, and data coming from optimization. Therefore, the second period should be fruitful in terms of insights obtained on these systems thanks to our novel algorithms.

Task 2.a/b: Certified approximation and hybrid representation of configuration spaces

The major goal for the first year of the project was to lay the foundations for hybridizing exact methods from computational geometry together with the sampling-based techniques which are in prevalent use in robotics.

Indeed we devised a general and modular algorithmic framework for path planning of robots, which we call *Motion Planning via Manifold Sample* or MMS for short. Our framework combines geometric methods for exact and complete analysis of low-dimensional configuration spaces, together with practical, considerably simpler sampling-based approaches that are appropriate for higher dimensions. In order to facilitate the transfer of advanced geometric algorithms into practical use, we suggest taking samples that are entire low-dimensional manifolds of the configuration space that capture the connectivity of the configuration space much better than isolated point samples. Geometric algorithms for analysis of low-dimensional manifolds then provide powerful primitive operations. The modular design of the framework enables independent optimization of each modular component.

The framework is presented in the paper [4]: Motion Planning via Manifold Samples, by Oren Salzman, Michael Hemmer, Barak Raveh, and Dan Halperin. Proc of the European Symposium on Algorithms (ESA), pp. 493-505, 2011. See also Technical Report CGL-TR-06 for a full version.

In collaboration between FU and TAU we are now working on extending MMS to include approximate contact surfaces. Preliminary investigation led to new insights about the paramterization of configuration spaces; see more below under Task 2.c.

While the MMS results constitute the highlights of our work for these tasks, we did address related issues. We continued our work on approximating certain configuration-space obstacles past the results that were presented at SoCG11 [1]; the new results appear in the full version of the paper—see Technical Report CGL-TR-03. We also worked on approximation of swept volumes with high-quality surface meshing [7] (see Technical Report CGL-TR-05), and on problems in molecular modeling [3] (see Technical Report CGL-TR-04) whose results will form the basis of some of our second-period work.

Task 2.c: Validations in robotics

To validate the MMS approach mentioned above, we implemented our framework for the concrete and fundamental case of a polygonal robot translating and rotating amidst polygonal obstacles. Following extensive comparative experiments, we demonstrate that the integration of several carefully engineered components leads to significant speedup over the popular PRM sampling-based algorithm (implemented in the Kavraki Lab, which is one of the leaders in the field), which represents the more simplistic approach that is prevalent in practice. In particular we have developed, implemented and optimized a primitive operation for complete and exact combinatorial analysis

of a certain set of manifolds, using arrangements of curves of rational functions and concepts of generic programming.

In addition we have proved a certain desirable property, called probabilistic completeness, of the application of our framework to this motion planning instance. The proof appears in Oren Salzman's M.Sc. thesis, which is uploaded to the CGL site.

We are currently working on extension of the scheme, the probabilistic-completeness proof, and validation for higher-dimensional configuration spaces.

We derived a new *explicit parameterization* of the contact surfaces that arise in the planning problem for a polygonal robot amid polygonal obstacles. The major advantage of this explicit parameterization is that it results in simpler algebraic expressions than common alternative parameterizations. It is already instrumental in our current efforts to apply MMS in the higher-dimensional space of coordinated motion of several robots. We wish to use and extend MMS and integrate it with this new *explicit parameterization*. In particular, we will analyze the intersections of arbitrary (hyper)-planes and the contact surfaces in the configuration space (the current implementation restricts the type of sample planes to either horizontal or vertical).

Task 3.a: Validations in cosmology

We have progressed in our program to use the geometric and topological information contained in the Cosmic Web — the weblike Megaparsec distribution of galaxies marked by prominent filamentary and sheetlike patterns surrounding large near-empty voids — to determine and extract cosmological information from the spatial galaxy distribution. We have made considerable progress in defining clear geometric and topological descriptions of the concepts of walls and filaments (over-dense regions) and voids (under-dense regions). This progress was underlined by a successful workshop "Cosmic Web Morphology and Topology" (Warsaw, July 2011), bringing together for the first time experts from the field of cosmology and computational geometry and topology.

One important application of our work is directed towards determining the nature of dark energy. Dark energy is the mysterious energy driving the accelerated expansion of the Universe, representing some 73% of its energy content. Its discovery is considered to be one of the most important scientific discoveries of the past decades, as has recently been recognized by the awarding of the Nobel prize of physics 2011 to the three leading discoverers. With the recent selection of the Euclid satellite as a cornerstone mission of the European Space Agency (ESA[1]), the search for dark energy has become of key importance for European science in the coming decade.

Technical Report CGL-TR-11, submitted to Astrophysical Letters, October 2011 [5]: R. van de Weygaert, P. Pranav, B. J.T. Jones, E.G.P. Bos, G. Vegter, H. Edelsbrunner, M. Teillaud, W.A. Hellwing, C. Park, J. Hidding, M. Wintraecken Probing Dark Energy with Alpha Shapes and Betti Numbers.

One aspect of this concerns an extensive collaboration on the use of homology for characterizing the cosmic web and analyzing its topological structure. As yet, most emphasis has been put on the determination of Betti numbers from the alpha shapes determined by the galaxy distribution. The first publication of this has been submitted to Astrophysical Journal Letters. In collaboration with H. Edelsbrunner we are deepening the analysis towards including a full persistence analysis. In the first we have compared the homology in tailormade Voronoi clustering models — as yet Betti numbers determined from the alpha shapes of the corresponding mass distribution, as a function of the alpha parameter — of the weblike matter contribution in different cosmic dark energy models.

[1]Notice that the acronym ESA is used to denote two different entities in the proposal, for both of which this abbreviation is used commonly.

A striking result is that we have demonstrated that "Betti signatures" are indeed able to detect subtle differences in clustering, undetectable by other measures, caused by a difference of dark energy dominating the Universe's dynamics and evolution.

Technical Report CGL-TR-12, journal paper [6]: R. van de Weygaert, G. Vegter, P. Pranav, B.J.T. Jones, H. Edelsbrunner et al. : Alpha, Betti and the Megaparsec Universe: on the Homology and Topology of the Cosmic Web. Transactions on Computational Science, XIV:6970, 60-101, 2011.

We have also studied the evolution of Betti numbers in computer simulations of genuine cosmological models. This has provided us with ample new insights into the evolution of the network of voids and tunnels in the Cosmic Web, while it proves that the Betti signature of a cosmological mass distribution offer a sensitive new method for obtaining additional cosmological information. This invited review, following the presentation of our work at the ISVD2010 conference in Quebec, on the use of homology analysis, including persistence, has been completed and published in recent weeks.

Outlook. The collaboration also involves two MSc projects, supervised by G. Vegter and R. van de Weygaert, on the application of methods from computational topology, in particular persistence structures, to Gaussian random fields. This project, also involving M. Wintraecken, is cosmologically highly significant as the primordial Universe was nearly Gaussian. The final results are planned for the second CGL-year.

In addition, we have been working with F. Chazal and D. Cohen-Steiner on the use of Geometric Inference to cosmological data (work package 1.1). This will be used to process and clean the astronomical data, to be followed by a topological analysis using persistence, and provides new insights onto the filamentary network connecting clusters of galaxies. We are currently working on a detailed investigation of the effect of a variation in input parameters on the resulting network.

Milestones

MS11: Decide on classification strategy for energy landscapes depending on preliminary results

Practically, constructing high-dimensional Morse-Smale diagrams may prove difficult, in particular due to under-sampling and/or stability issues. The outcome of these investigations will condition the strategies developed for classifying landscapes, and a fall-back option will consist of using persistent minima and their connexions.

As explained while describing Task 1a and 1c, we are currently consolidating the experimental results to make a firm statement on the added value of transitions paths.

MS12: Choice of dimension for approximation of constraint surfaces

In view of the presence and availability of software that can handle two-dimensional arrangements, we decided for now to compute arrangements of piecewise linear structures in two-dimensional subspaces and not in three-dimensional subspaces.

Bibliography

[1] Eric Berberich, Dan Halperin, Michael Kerber, and Roza Pogalnikova. Deconstructing approximate offsets. In *Symposium on Computational Geometry*, pages 187–196, 2011. A full version appears in Technical Report CGL-TR-03.

[2] Frederic Cazals and Christine Andrea Roth. Investigating the energy landscape of macromolecular systems: Data generation software overview. Technical Report CGL-TR-07, INRIA-ABS, 2011.

[3] Barak Raveh, Nir London, Lior Zimmerman, and Ora Schueler-Furman. Rosetta flexpepdock ab-initio: Simultaneous folding, docking and refinement of peptides onto their receptors. *PLoS ONE*, 6(4), 2011. See also Technical Report CGL-TR-04.

[4] Oren Salzman, Michael Hemmer, Barak Raveh, and Dan Halperin. Motion planning via manifold samples. In *ESA*, pages 493–505, 2011. A full version appears in Technical Report CGL-TR-06.

[5] R. van de Weygaert, P. Pranav, B. J.T. Jones, E.G.P. Bos, G. Vegter, H. Edelsbrunner, M. Teillaud, W.A. Hellwing, C. Park, J. Hidding, and M. Wintraecken. Probing dark energy with alpha shapes and betti numbers. Technical Report CGL-TR-11, University of Groningen, 2011. Submitted to Astrophysical Letters, October 2011.

[6] R. van de Weygaert, G. Vegter, P. Pranav, B.J.T. Jones, and H. Edelsbrunner et al. Alpha, betti and the megaparsec universe: on the homology and topology of the cosmic web. *Transactions on Computational Science*, XIV:60–101, 2011. See also Technical Report CGL-TR-12.

[7] Andreas von Dziegielewski and Michael Hemmer. High quality surface mesh generation for swept volumes. In *EuroCG '11: Abstracts from the 27th European Workshop on Computational Geometry*, pages 143–146, Morschach, Switzerland, March 2011. See also Technical Report CGL-TR-05.