Numer. Math. Theor. Meth. Appl. Vol. **3**, No. 2, pp. 195-211
doi: 10.4208/nmtma.2010.32s.5 May 2010

Fast Multilevel CVT-Based Adaptive Data Visualization Algorithm

M. Emelianenko*

Department of Mathematical Sciences, George Mason University, Fairfax, VA 22031, USA.

Received 25 September 2009; Accepted (in revised version) 6 January 2010

Available online 8 March 2010

Abstract. Efficient data visualization techniques are critical for many scientific applications. Centroidal Voronoi tessellation (CVT) based algorithms offer a convenient vehicle for performing image analysis, segmentation and compression while allowing to optimize retained image quality with respect to a given metric. In experimental science with data counts following Poisson distributions, several CVT-based data tessellation algorithms have been recently developed. Although they surpass their predecessors in robustness and quality of reconstructed data, time consumption remains to be an issue due to heavy utilization of the slowly converging Lloyd iteration. This paper discusses one possible approach to accelerating data visualization algorithms. It relies on a multidimensional generalization of the optimization based multilevel algorithm for the numerical computation of the CVTs introduced in [1], where a rigorous proof of its uniform convergence has been presented in 1-dimensional setting. The multidimensional implementation employs barycentric coordinate based interpolation and maximal independent set coarsening procedures. It is shown that when coupled with bin accretion algorithm accounting for the discrete nature of the data, the algorithm outperforms Lloyd-based schemes and preserves uniform convergence with respect to the problem size. Although numerical demonstrations provided are limited to spectroscopy data analysis, the method has a context-independent setup and can potentially deliver significant speedup to other scientific and engineering applications.

AMS subject classifications: 65D99, 65C20

Key words: Centroidal Voronoi tessellations, computational algorithms, Lloyd's method, acceleration schemes, multilevel method, binning, image analysis, visualization, signal-to-noise ratio.

1. Introduction

Centroidal Voronoi tessellations have diverse applications in many areas of science and engineering and the development of efficient algorithms for their construction is a key to their success in practice. Its use in imaging applications is the subject of many

*Corresponding author. *Email address:* `memelian@gmu.edu` (M. Emelianenko)

recent studies (e.g. [2–4] and references therein). This work will demonstrate how this concept can bring significant advantages to applications dealing with data visualization of extremely large data sets.

In particular, we focus on the problem of adaptively binning intensity images in physics applications, such as those coming from spectroscopy data with counting (Poisson) statistics. Hardness ratio maps, temperature maps and other types of images can be analyzed in a much similar manner. Spectroscopy has established itself as an unrivaled technique for a wide variety of scientific problems. However, experiments carried out using such techniques suffer from severe intensity limitations inherent to imperfect measurement instruments. Because of these limitations one usually needs to trade off signal-to-noise ratio (SNR) and instrumental resolution to raise statistical significance of measured data, which leads to the need to bin (or group) spatially close data for a better view of the image as a whole — a process referred to as *binning*. Adaptive techniques, able to relate bin sizes to local signal-to-noise ratios, are of extreme importance, since they prevent loss of resolution in high intensity areas. Even with adaptive binning, the size of the data to be processed is very large, with a typical modern day spectrometer delivering data sets of the order of 10^8 detector pixels. Hence high-speed and high-accuracy computational algorithms are crucial in order to make online data visualization and assessment possible.

This work builds upon several recently developed CVT-based methods for binning scattering data. Capellari and Copin [5] working on astrophysics imaging data were the first to develop a CVT-based adaptive binning method that achieves a homogeneous distribution of SNR across the image. Diehl and Statler [6] improved its performance by utilizing the concept of a weighted Voronoi diagram for added flexibility. Recently, Bustinduy et al. [7] proposed another adaptive algorithm that is able to preserve high-SNR features and avoid blurring common to the previous two approaches. While pursuing slightly different goals, all of the above works share one common drawback, which is high computational complexity associated with constructing optimal centroidal Voronoi tessellations of the given dataset. This fact serves as a motivation for the current study.

We suggest that a multidimensional extension of a multilevel scheme previously developed and extensively analyzed in 1-dimensional quantization context can significantly accelerate these and other data visualization techniques. The method was originally devised in one space dimension in [1], where it was proved to have uniform convergence with respect to the grid size for a large class of densities. Numerical studies of its behavior for some 2-dimensional domains with simplified geometries have been carried out in [1] and [8] with similar results, so the method is conjectured to have superior convergence properties regardless of dimension. Other possible acceleration techniques such an hybrid Lloyd-Newton and quasi-Newton algorithms, were considered in [8] and their applicability to imaging applications are subject to current explorations. In the version adapted for image analysis, the multilevel method is presently tested in the Capellari-Copin setting which aims at minimizing the spread of the signal-to-noise ratio around a target value $(S/N)_T$, but it can be very easily reformulated to fit other scenarios. More importantly, the multilevel algorithm description provided in current work is independent of the problem dimension or the features of the discrete dataset.