Multidimensional Compression with Pattern Matching

Olivia Del Guercio[†], Rafael Orozco[*], Alex Sim[‡] and Kesheng Wu[‡]

[†]Scripps College   [‡]Lawrence Berkeley National Laboratory   [*]Bucknell University

delgur@gmail.com   kwu@lbl.gov   rao010@bucknell.edu

This work utilizes statistical pattern matching to reduce the storage requirement for sequences of numerical values. The work employs multidimensional similarity measures including Minimum Jump Cost (MJC) and Dynamic Time Warp (DTW) to find similar data blocks and to create a dictionary of "patterns." These multidimensional similarity measures extend the overall data reduction technique from one-dimensional sequences to multidimensional sequences. We expected the compression algorithm with MJC as similarity measure to use less CPU time, while sacrificing quality to a certain extent. We indeed observe that using MJC the compression algorithm is 2× as fast as using DTW, what was surprising was that using MJC would produce even lower compression errors. In this extended abstract, we will only show performance results with MJC as the similarity measure.

Next, we show a comparison of compression with MJC to the leading compression algorithm named SZ[1]. At compression ratios greater than 80, MJC has lower error, and the differences in runtime are also reducing as the compression ratio increases.

MJC visually deviates much less from the original data than SZ. So while there is a sacrifice in efficiency MJC preserves many desirable features in reconstructed data–the value of which is often dependent on the set. SZ relies on nearby points to perform reconstruction, leading to inaccurate decompression in highly variable data, such as output from scientific sensors.
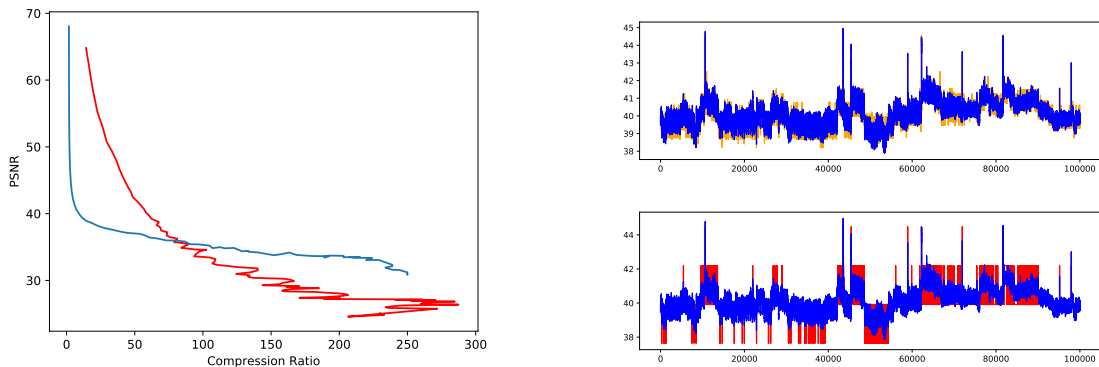

Figure 1: Compression ratio plotted against PSNR for MJC and SZ (left). MJC and SZ compression quality plotted against original single stream of current data over time at 100 CR. MJC has PSNR of 35.4 and SZ has a PSNR of 33.5 (right).

---

[1]https://collab.cels.anl.gov/display/ESR/SZ