Data Representation and Exploration with Geometric Wavelets

Eric E. Monson*
Duke Visualization Technology Group

Guangliang Chen
Duke Mathematics

Rachael Brady
Duke Visualization Technology Group

Mauro Maggioni†
Duke Mathematics and Computer Science

Abstract

Geometric Wavelets is a new multi-scale data representation technique which is useful for a variety of applications such as data compression, interpretation and anomaly detection. We have developed an interactive visualization with multiple linked views to help users quickly explore data sets and understand this novel construction. Currently the interface is being used by applied mathematicians to view results and gain new insights, speeding methods development.

Index Terms: H.5.2 [Information Interfaces and Presentation]: User Interfaces—Graphical user interfaces (GUI); I.5.1 [Pattern Recognition]: Models—Geometric

1 Geometric Wavelets

Data sets such as images, documents or gene expression data may be modeled as point clouds in high-dimensional Euclidean space. In the case of images, each pixel can be thought of as one coordinate in a vector with a length equal to the number D of pixels in the image, and the intensity of each pixel corresponds to the coordinate magnitude in that pixel's direction. Real data points often have structure which has dimension d much smaller than the ambient space dimension D, for example under the well-studied case when they lie near a low-dimensional manifold $\mathscr{M}$. Discovering and characterizing this lower-dimensional structure can dramatically affect the performance in tasks such as data compression, interpretation, outlier detection, classification and clustering.

If $\mathscr{M}$ is just a linear subspace, Principal Component Analysis (PCA) can discover a dictionary of d vectors which describe the data well at low computational cost. However, when $\mathscr{M}$ is nonlinear it is usually necessary to use random dictionaries or black box optimization, which are much more costly and in general do not yield interpretable features of the data. Geometric Wavelets [2] are multi-scale dictionary elements which are constructed directly from the data, adapt to arbitrary nonlinear manifolds, and have guarantees on the computational cost, the number of elements in the dictionary and the sparsity of the representation (as a function of an approximation error parameter). In particular they provide feature sets that may be particularly useful for data exploration, and tasks such as anomaly detection and classification.

The mathematical details of the construction can be found in [1]. It proceeds in several steps: first, relationships between data points are computed with respect to a given similarity function. At the coarsest scale all data points are considered one group and global PCA is performed, yielding a d-dimensional plane fit to the data with axes in the directions of maximum variance, which we think of as a parallel of "scaling functions" in wavelet analysis. The projection of the data points onto this plane is the coarsest-scale approximation of the data. Next, the graph is split into two groups (e.g. by using METIS [3]). On each of these finer scale groups, PCA is again performed and the projection of the data points onto these two new planes will more accurately approximate $\mathscr{M}$. To form a compact representation for the data at this finer scale, as in a wavelet decomposition, we only encode the differences between the original coarse projections of the data and the points projected onto the planes at the finer scale. In order to do this an efficient scheme is derived based on the construction of a minimal space spanning this set of differences. The axes of this difference space are called "geometric wavelets", and the projections of the finer-scale corrections to the data points onto the plane spanned by these axes are called the "wavelet coefficients". The process is continued, forming a binary tree of parents and children at finer and finer scales until no further details are needed to approximate the data up to a pre-specified precision. Geometric wavelets provide a dictionary or feature set of the data that efficiently captures coarse-to-fine structure in the data, and the data may be transformed back and forth between its original representation and a geometric wavelet representation via fast algorithms. Figure 1 shows a schematic of the coarsest and the first finer scale of this decomposition. (Distances are exaggerated for clearer viewing.)

*e-mail: emonson@duke.edu
†e-mail: mauro.maggioni@duke.edu

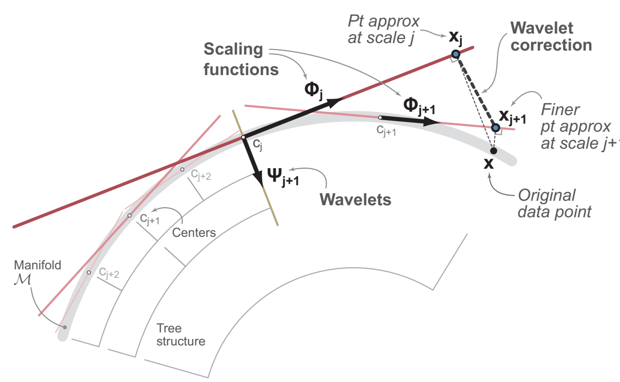


Figure 1: Geometric Wavelets schematic.

2 Visualization GUI

The mathematical methods behind this representation are being developed in Matlab (*Mathworks*), and the researchers often end up with numerous static plots on the screen while viewing results. One goal for the visualization is to allow users to quickly see and navigate the representation. A second goal is to begin developing a platform onto which we could build more specialized applications as others start using these techniques with their own data. A nice byproduct of the visualization is its usefulness for explaining Geometric Wavelets and helping people gain more intuition about this representation.

The application itself is implemented in Python, using PyQt4 to glue together the views, which themselves are constructed using wrapped classes and customized variants from VTK [7]. During this development stage, the Geometric Wavelets are not computed directly in the GUI – Matlab output is loaded from files.

Figure 2 shows the GUI layout and describes some of its features. For this example, a data set consisting of 1000 ones and 1000 twos chosen at random from the MNIST handwritten digits database is being used [4]. Initially only the icicle view of the bi-

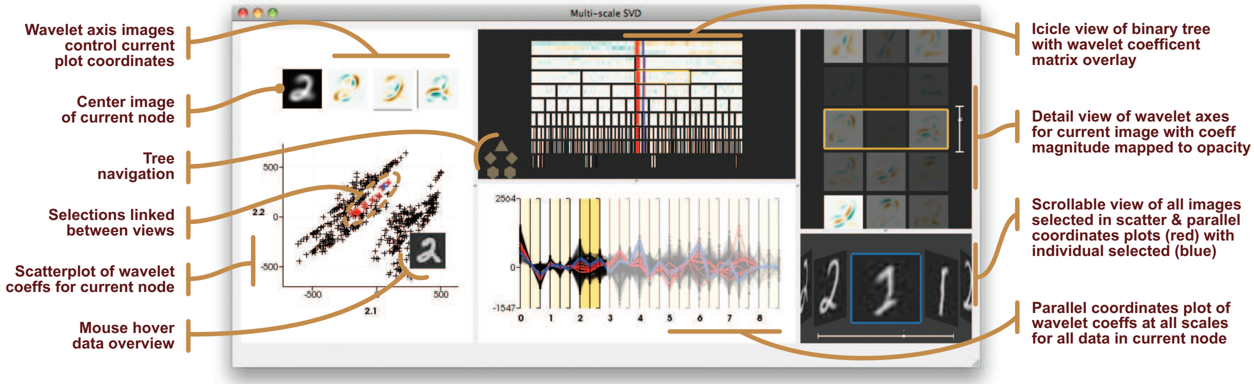
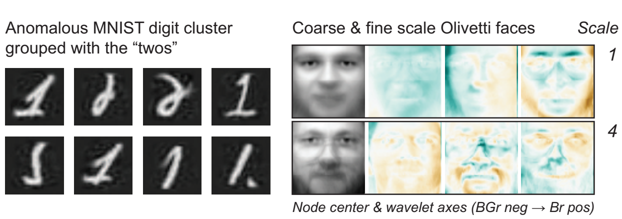


Figure 2: Annotated GUI for viewing and navigating the multi-scale Geometric Wavelets data representation.

nary splits (tree) of data point groups is shown (coarsest scale at the top and finer below). Overlaid on this tree is the matrix of wavelets coefficients for all data points at all scales. When a node of the tree is selected, the wavelet coefficients are plotted in a scatter plot on the left. Above the scatter plot are the node center, which is a quick reference of the "average" data point in that node, along with the axes which define the wavelets. Clicking on the wavelet images switches which dimensions are plotted, and hovering over a scatterplot point displays the original data as a "tooltip". This is an easy way to get an overview of what clusters and outliers in the plots represent. Below the tree, a parallel coordinates plot of the wavelet coefficients for the data in the selected node at all scales is shown. The current scale is highlighted in gold, and finer scales have a semi-opaque overlay, indicating that these values are not strictly comparable since child nodes lie in different spaces.

Groups of data can be selected in either the scatter or parallel coordinates plots, and red highlights will show up for that data in those plus the icicle view. After this type of subset selection is made, the original data (images) associated with these points appears in a scrollable "flow" view in the lower right. This allows easy group data comparisons. If an individual image is clicked on, the detail view above it shows the multi-scale characteristics of that image through the wavelet images at each scale (and that point is highlighted in blue on the two plots and the icicle view). The opacity tracks the absolute value of the wavelet coefficient in that direction, so with a glance you can see the primary components that represent that data point. By clicking on different scales in this detail view, it is possible to navigate through the tree in the icicle view along the path defined by this individual. This helps especially when trying to see what groups an individual is a member of at different scales for finding other similar data points (like outliers).

3 Data Exploration and Methods Development



We note here some observations from exploring various data sets and their representations. At coarser scales you get generalized approximations of the data, with readily interpretable node centers and wavelet directions. With the MNIST digits (e.g. ones and twos, as above), there is good separation of categories at coarse scales. At finer scales it is easy to find anomalous data (either mis-categorized or strangely shaped digits), by finding extreme wavelet coefficients or wavelet axis images which are messy mixtures of shapes rather than variations on recognizable digits. When viewing the Olivetti faces (400 images of 40 people [6]), it is clear that coarser-scale wavelets contain information which could be ignored for classification tasks, but finer-scale wavelets encode more specific features which cluster and characterize people and expressions.

The visualization GUI has been developed in close collaboration with the Applied Math methods developers. While viewing scatterplots of the wavelet coefficients, they noticed that many nodes have coefficients which are clustered along lines that do not necessarily correspond to the directions of the wavelet space axes. This means that a much more sparse representation will be possible if the wavelet directions are optimized for the natural directions of the data. The ability to quickly view the large and complex space of results is already streamlining methods development.

4 Future Work

At this stage the dimensionality d is fixed at the beginning of the analysis, but in the future it will be variable and adapted to the local dimensionality of the data [5]. Methods are also under development for pruning the tree to obtain an even more compact data representation. For the visualization we will be labeling data points and defining groups for semi-supervised learning and analysis tasks, as well as adapting the GUI to other data types and integrating it with the wavelet construction.

Acknowledgements

This work was supported in part by the NSF FODAVA (Mathematical Foundations of Visual Analytics) program.

References

[1] W. K. Allard, G. Chen, and M. Maggioni. Multiscale Geometric Methods for Data Sets II: Geometric Wavelets (in preparation).

[2] G. Chen and M. Maggioni. Multiscale Geometric Wavelets for the Analysis of Point Clouds. *Proc. CISS*, 2010.

[3] G. Karypis and V. Kumar. Metis: Unstructured graph partitioning and sparse matrix ordering system. *University of Minnesota*, 1995.

[4] Y. LeCun and C. Cortes. The MNIST database of handwritten digits: http://yann.lecun.com/exdb/mnist/.

[5] A. V. Little, M. Maggioni, and L. Rosasco. Multiscale Geometric Methods for Data Sets I: Intrinsic dimension (in preparation).

[6] F. Samaria and A. Harter. http://cs.nyu.edu/∼roweis/data.html.

[7] W. Schroeder, K. Martin, and B. Lorensen. *The Visualization Toolkit, Fourth Edition*. Kitware Inc., 2006.