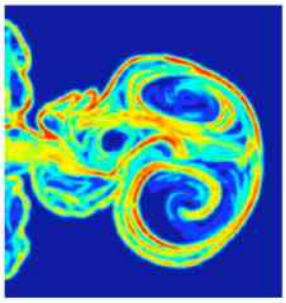
Project Closeout Report

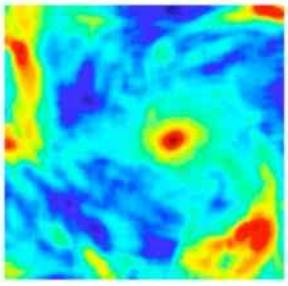
An Information-Theoretic Framework for Enabling Extreme-scale Science Discovery

The Ohio State University

Award #: DE-FC02-10ER26018/DE-SC0005036

Principle Investigator: Han-Wei Shen





Research Goal

The goal of this project is to develop an information-theoretic framework for large-scale scientific data analysis. The framework includes efficient computation of information entropy, and applying entropy measures to evaluate the quality of the visualization output. With the quantitative measurement results, it is possible to optimize the algorithm parameters so that salient information can be more effectively extracted. The target data of our analysis framework include scalar, vector, and multivairate spatial data, both in three- and four-dimensional (time varying) space. The problems that have been tackled under this project include efficient histogram and distribution query for entropy computation, optimizing flow line seed selection, salient isosurface selection, key time step selection for time-varying data, and analysis of multi variate data sets.

Major Activities

In this section and sections that follow, we highlight the major activities at the Ohio State University, divided into the following categories:

1. Histogram computation, representation, and compression

Because computing various information measures requires distributional information from the data, it is crucial to have the ability to compute distributions from an arbitrary region of the domain efficiently and effectively. Toward this goal, a major research direction of this project is to develop algorithms to support fast histogram queries and to store histograms compactly to minimize the computation and I/O cost. We have developed a novel idea called integral histogram to achieve the goals. Integral histograms extend the concept of the Summed Area Table and store each histogram in an axis-aligned region bounded by the origin and every grid point. By pre-computing the integral histograms and arranging them hierarchically, histogram queries for arbitrary spatial intervals can be answered efficiently. While integral histograms can support fast queries for regions of arbitrary sizes at arbitrary locations, the storage cost can be high for large data sets. To address this issue, we have developed several algorithms that can reduce the cost but still ensure fast query speed. We have developed a wavelet base compression algorithm for integral histograms by taking advantage of the fact that values stored in the integral histogram for each bin is a monotonically increasing function. Realizing that most of scientific data sets have a large degree of spatial coherence, i.e., many regions have similar distributions, we developed a method that can extract the most common distributions as the templates which are then used to index all regions based on histogram similarity. Finally, to handle very large data sets, we developed parallel algorithms to compute histograms for distributed memory supercomputers.

2. Information theory based data summarization, reduction, and triage

The distributions generated from the algorithms described above can be used to compute various entropy measures used for data summarization, reduction, and triage. Toward this goal, we have developed a myriad of scientific data analysis and visualization algorithms for extracting salient regions, controlling levels of detail, and grouping data based on their statistical signatures and complexity. Specifically, we developed a flow particle seeding algorithms based on conditional entropy measures. The goal of our method is to evaluate flow visualization results

generated from streamlines. We use the conditional entropy to measure the correlation between the input data and the visualization output, and to refine the streamlines until satisfactory visualizations are obtained. The idea of the conditional entropy was also used in a salient isosurface selection algorithm. In the algorithm, conditional entropies are used to evaluate the statistical similarity between true isosurfaces and isosurface approximated by the level set method. Isosurfaces that cannot be easily approximated by the level set method are selected as the salient isosurfaces for the underlying scalar field. To allow efficient run time level of detail selection, we developed a novel algorithm called histogram spectra which stores the differences between histograms generated from data at different resolutions. With the histogram spectra, users can specify different importances for different value ranges at run time and the most suitable level of detail with the minimum error is selected by our algorithm. Finally, to group data by their statistical significance and similarity, we have developed a streamline classification method based on the histograms of curvatures and torsions computed along the sample points of selected streamlines. Using the histogram of geometric measures allows us to capture the intrinsic geometric natures of streamlines modeled as space curves with very efficient similarity comparison.

3. Time-Varying multi-variate data analysis

Even with fast distribution query and data summarization and reduction, scientists still need assistance during run time exploration of time-varying multivariate data sets. This is of particular challenge because a typical scientific simulation can produce output that includes many variables with a large number of time steps. It is often difficult to understand the relationships between different variables and decide what time steps should be used to perform more detailed analysis. To address the issues, we have developed a salient time step selection algorithm based on the concept of dynamic time warping (DTW). DTW requires fast distance computation between data at two time steps and we have shown that distribution-based measures can serve this goal very well. To guide the users to explore multivariate data sets, we use the concept of specific mutual information to analyze the correlation between values of two variables co-occurring at same grid locations. With the specific mutual information measure, we are able to identify isosurfaces of one variable where the other variable has the minimum and maximum uncertainty.

Specific Findings

1) W**avelet based integral histogram compression:**
We developed a new algorithm called WaveletSAT, which utilizes integral histograms, an extension of the summed area tables (SAT), and discrete wavelet transform (DWT). Similar to SAT, an integral histogram is the histogram computed from the area between each grid point and the grid origin, which can be pre-computed to support fast query. Nevertheless, because one histogram contains multiple bins, it will be very expensive to store one integral histogram at each grid point. To reduce the storage cost for large integral histograms, WaveletSAT treats the integral histograms of all grid points as multiple SATs, each of which can be converted into a sparse representation via DWT, allowing the reconstruction of axis-aligned region histograms of arbitrary sizes from a limited number of wavelet coefficients. We developed an efficient wavelet transform algorithm for SATs that can operate on each grid point separately in logarithmic time complexity, which can be extended to take advantage of parallel GPU-based implementation.

2) **Template-based histogram storage and indexing**
In this work, we addresses the issue of accelerating range distribution query, which returns the

distribution of an axis-aligned query region. Maintaining the interactivity when performing such query is a challenging task because the workload, which affects the response time, of such queries, is difficult to scale up with the data and the query size. We developed a framework for answering range distribution queries for any arbitrary region in near constant time, regardless of the data and query size. We adopt an integral histogram based data structure to bound the workload which is a combination of computation, I/O and communication cost. We devised two novel transformations of this data structure – a decomposition and a similarity driven indexing to reduce the storage cost.

3) **Conditional entropy for seed selection in flow visualization**
We developed an information-theoretic framework for flow visualization with a special focus on streamline generation. In our framework, a vector field is modeled as a distribution of directions from which Shannon's entropy is used to measure the information content in the field. The effectiveness of the streamlines displayed in visualization can be measured by first constructing a new distribution of vectors derived from the existing streamlines, and then comparing this distribution with that of the original data set using the conditional entropy. The conditional entropy between these two distributions indicates how much information in the original data remains hidden after the selected streamlines are displayed. The quality of the visualization can be improved by progressively introducing new streamlines until the conditional entropy converges to a small value.

4) **Salient isosurface selection**
We developed an information-theoretic approach to evaluate the representativeness of a given isosurface set. The main idea is that given two isosurfaces that enclose a subvolume, if the intermediate isosurfaces in the subvolume can be generated by smoothly morphing from one isosurface to the other, no additional isosurfaces are needed since the geometry of the true isosurfaces within the subvolume can be easily inferred. To realize this idea, given a pair of isosurfaces, to determine if such a smooth condition in the enclosed region is satisfied, we use a level-set approach to generate the intermediate surfaces. On each intermediate surface, we sample the values from the scalar field and exam the distribution. If the entropy of the distribution is low, this intermediate surface is aligned well with a true isosurface in the scalar field. For the intermediate surfaces generated by the level-set method from the boundary isosurfaces, the distributions of scalar values from the level-set surfaces form a 2D distribution, called isosurface information map. This information map can be used as an indicator of the representativeness of the boundary isosurfaces for the data in the subregion, allowing a quantitative measurement of information representable by the input isosurfaces. Based on this information-theoretic approach, the isosurface selection algorithm that can automatically pick salient isosurfaces for more effective visualization of scalar fields.

5) **Histogram spectra for level of detail selection**
Level of detail techniques are widely applied to minimize sampling error subject to working set size constraints. Typical large data sets being produced today have many variables sampled across time-varying volumes. Visualization of these multivariate volumes is commonly phrased in terms of conditional expressions such as "show variable A where variable B is between B1 and B2." The bounds, B1 and B2, tend to be specified during the interactive portion of the workflow. Thus, to maximize quality over the salient interval, level of detail selection should also be interactive. We introduce the concept of histogram spectra to quickly and compactly quantify the statistical sensitivity of volumes to sampling. Salient interval volumes of one or more variables are used to select which parts of the histogram spectra are important. The level of detail selection problem, over a time-varying, multivariate, multiresolution volume, is then posed as an integer programming problem using the histogram spectra. We developed

an efficient solution enabling interactive LOD selection on large, out-of-core volumes and showed its efficacy on real data sets from different problem domains.

6) **Distribution-based query and classification of flowlines**
Streamline-based techniques are designed based on the idea that properties of streamlines are indicative of features in the underlying field. In this work, we showed that statistical distributions of measurements along the trajectory of a streamline can be used as a robust and effective descriptor to measure the similarity between streamlines. With the distribution-based approach, we developed a framework for interactive exploration of 3D vector fields with streamline query and clustering. Streamline queries allow us to rapidly identify streamlines that share similar geometric features to the target streamline. Streamline clustering allows us to group together streamlines of similar shapes. Based on user's selection, different clusters with different features at different levels of detail can be visualized to highlight features in 3D flow fields.

7) **Information-aware multi-variate data analysis**
In this work, we developed an algorithm towards building an exploration framework based on information theory to guide the users through the multivariate data exploration process. In our framework, we compute the total entropy of the multivariate data set and identify the contribution of individual variables to the total entropy. The variables are classified into groups based on a novel graph model where a node represents a variable and the links encode the mutual information shared between the variables. The variables inside the groups are analyzed for their representativeness and an information based importance is assigned. We exploit specific information metrics to analyze the relationship between the variables and use the metrics to choose isocontours of selected variables. For a chosen group of points, parallel coordinates plots (PCP) are used to show the states of the variables and provide an interface for the user to select values of interest. Experiments with different data sets reveal the effectiveness of our proposed framework in depicting the interesting regions of the data sets taking into account the interaction among the variables.

Publications

Chaudhuri, A. and Tzu-Hsuan Wei and Teng-Yok Lee and Han-Wei Shen and Peterka, T., Efficient Range Distribution Query for Visualizing Scientific Data. IEEE Pacific Visualization Symposium 2014.

Teng-Yok Lee and Han-Wei Shen. Efficient Local Statistical Analysis via Integral Histograms with Discrete Wavelet Transform. *IEEE Transactions on Visualization and Computer Graphics*. 19 (12), 2693-2702.

Ayan Biswas and Soumya Dutta and Han-Wei Shen and Jonathan Woodring. An Information-Aware Framework for Exploring Multivariate Data Sets. *IEEE Transactions on Visualization and Computer Graphics*. 19 (12), 2683-2692.

Kewei Lu and Abon Chaudhuri and Teng-Yok Lee and Han-Wei Shen and Pak Chung Wong, Exploring Vector Fields with Distribution-based Streamline Analysis. IEEE Pacific Visualization Symposium 2013.

Martin, Steve and Shen, Han-Wei, Transformations for Volumetric Range Distribution Queries. IEEE Pacicic Visualization Symposium 2013.

Tzu-Hsuan Wei and Teng-Yok Lee and Han-Wei Shen. Evaluating Isosurfaces with Level-set-based Information Maps. *Computer Graphics Forum*. 32 (3), 1-10.

Martin, S. and Han-Wei Shen, *Interactive transfer function design on large multiresolution volumes*. Proceedings of 2012 IEEE Symposium on Large Data Analysis and Visualization (LDAV).

CHEN, C.-M., XU, L., LEE, T.-Y., AND SHEN, H.-W. A flow-guided file layout for out-of-core streamline computation. In PacificVis '12: Proceedings of the IEEE Pacific Visualization Symposium 2012, pp.145–152.

Xin Tong and Teng-Yok Lee and Han-Wei Shen, Salient time steps selection from large scale time-varying data sets with dynamic time warping. Proceedings of 2012 IEEE Symposium on Large Data Analysis and Visualization (LDAV).

Martin, Steven and Shen, Han-Wei. *Histogram spectra for multivariate time-varying volume LOD selection*. Proceedings of 2011 IEEE Symposium on Large Data Analysis and Visualization (LDAV).

A. Chaudhuri, T.-Y. Lee, B. Zhou, C. Wang, T. Xu, H.-W. Shen, T. Peterka, and Y.-J. Chiang, "Scalable Computation of Distributions from Large Scale Data Sets." Submitted to IEEE Symposium on Large Data Analysis and Visualization (LDAV '12), 2012.

Kun-Chuan Feng, Chaoli Wang, Han-Wei Shen, and Tong-Yee Lee, Coherent Time-Varying Graph Drawing with Multifocus+Context Interaction, IEEE Transactions on V

NOUANESENGSY, B., LEE, T.-Y., AND SHEN, H.-W. Load-balanced parallel streamline generation on large scale vector fields. IEEE Transactions on Visualization and Computer Graphics 17, 12 (2011), 1785–1794.

Chaoli Wang and Han-Wei Shen, Information Theory in Scientific Visualization Entropy (Special Issue on Advances in Information Theory), 13(1):254-273, January 2011

Teng-Yok Lee, Oleg Mishchenko, Han-Wei Shen, and Roger Crawfis, View point evaluation and streamline filtering for flow visualization, IEEE Pacific Visualization 2011, pp 83-90, March 2011

Load-balaned Parallel Streamline Generation on Large Scale Vector Fields, IEEE Transactions on Visualization and Computer Graphics, 17(12), 2011

Lijie Xu, Teng-Yok Lee, and Han-Wei Shen, An information-theoretic framework for flow visualization. IEEE Transactions on Visualization and Computer Graphics, 16(6): 1216-1224, Nov./Dec. 2010