Analysis of Large-Scale Scalar Data Using Hixels

Joshua A. Levine\textsuperscript{2}, in collaboration with D. Thompson\textsuperscript{1}, J.C. Bennett\textsuperscript{1}, P.-T. Bremer\textsuperscript{3}, A. Gyulassy\textsuperscript{1}, P.P. Pébay\textsuperscript{4}, V. Pascucci\textsuperscript{1}
HPC Has Lead to Increases in Both Data Size and Complexity

- “Hero” runs
  - Increased spatial resolution
  - Increased number of variables

- Uncertainty Quantification (UQ)
  - Ensembles of runs
  - Polynomial Chaos
  - Stochastic Simulations

- Many analysis methods do not scale with size & complexity of the data

Hixels: A Unified Data Representation

- A **hixel** is a point with an associated histogram of scalar values.
- Hixel samples may represent:
  - Spatial down-sampling
  - Ensemble values
  - Random variables
- Trade data size/complexity for uncertainty
1D Example of Hixels (Block Compression)
Motivation: Feature-Based Analysis

- Characterize and define features
- Segmentation domain by function behavior
- Answer questions:
  - How many features are there?
  - What is the behavior of other variables within these features?
  - How do you define a good threshold value on which to segment the domain?

Data courtesy of: Dr. Jacqueline Chen, SNL
Goal: Extend Topological Methods

- What structures are present?
- How persistent are they?
- How do we visualize features?

Our Contributions:
1. Sampled topology
2. Topological analysis of statistically associated buckets
3. Visualizing fuzzy isosurfaces
Sampled Topology: Algorithm

1. Sample the hixels to construct a scalar field $V_i$
2. Compute the Morse complex for $V_i$
   a) Identify basins around minima & arcs between adjacent basins
   b) Encode arc locations in a binary field $C_i$
      • Boundaries = 1, Rest = 0
3. Construct aggregate $A$ as mean of the $C_i$’s
4. Visualize variability of arc locations

Assumption: hixels are independent
Aggregate Segmentation on Temporal Jet

1 run  
16 runs  
64 runs  
256 runs  
16384 runs

\[ p = 0 \]

\[ p = 0.128 \]

\[ p = 0.008 \]
Convergence of Sampled Topology

Topological convergence for 8x8 blocks
Varying Block Size & Persistence

1x1
1 runs

2x2
512 runs

4x4
2048 runs

8x8
8196 runs

16x16
16384 runs

A

p = 0.064
p = 0.016
p = 0.004
p = 0
p = 0.256
Topological Analysis of Statistically Associated Buckets: Algorithm

- Aimed at recovering prominent features from ensemble data
  - Exploit dependencies between runs
  - Identify regions in space & scalar values consistent with positive association
  - Perform topological segmentation on these regions individually

1. Compute buckets
2. Compute contingency statistics
3. Identify sheets
4. Perform topological analysis on individual sheets
Computing Buckets

- Values of high probability associated with peaks in the histogram
- Identify peaks + range of function values around that peak
- Topological segmentation on histogram
  - Use areal (hypervolume) persistence
  - Weight of interval = area of the histogram
  - Merge until the probability of smallest bucket is above a particular threshold
Persistence Simplification of Buckets

Persistence Pairs
Persistence Simplification of Buckets
Persistence Simplification of Buckets
Persistence Simplification of Buckets
Effect of Persistence on Bucket Count

Persistence Threshold (p)

Number of Buckets

p = 16
p = 32
p = 64
p = 128
p = 256
p = 512
Contingency Tables on Bucketed Hixels

<table>
<thead>
<tr>
<th></th>
<th>e</th>
<th>f</th>
<th>g</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>4</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>b</td>
<td>2</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>c</td>
<td>0</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>d</td>
<td>6</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>h</th>
<th>i</th>
<th>j</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>5</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>b</td>
<td>1</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>c</td>
<td>2</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>d</td>
<td>0</td>
<td>1</td>
<td>5</td>
</tr>
</tbody>
</table>
Pointwise Mutual Information (PMI) Encodes Association Between Hixels

Goal: Identify buckets that co-occur more frequently than if statistically independent

\[
\text{pmi}(x, y) := \log \left( \frac{p(x,y)(x,y)}{p_x(x)p_y(y)} \right)
\]

\[
\text{pmi}(x,y)=0 \implies x \text{ independent } y
\]
Positive PMI Constructs Sheets of Statistically Associated Buckets

Before: Bucketed Hixels
Positive PMI Constructs Sheets of Statistically Associated Buckets

After: Sheets Connecting Buckets
An Ensemble of Mixed Distributions

- 512 x 512 hixels, 128 bins each
- 3200 samples from Poisson distribution
  - $\lambda$ is a 100 at 5 source points in a circle
  - $\lambda$ decreases to 12 $\propto$ distance from source points
- 9600 samples from a Gaussian distribution
  - $\mu$ & $\sigma$ are min & max at 4 points in a circle
  - $\mu$ & $\sigma$ vary $\mu$ distance from source points
An Ensemble of Mixed Distributions
“Simple” Topological Tests Fail!

- Probability that each hixel corresponds to
  - Minimum ~ 20%
  - Maximum ~ 20%
  - Saddle ~ 7%
  - Regular point ~ 53%
Sheets Isolate Prominent Features

Basins of Minima

Basins of Maxima
Sheets for Lifted Ethylene Jet
Visualizing Fuzzy Isosurfaces: Algorithm

1. Compute likelihood function $g$

$$g = \begin{cases} a, & b = 0 \\ -b, & a = 0 \\ \frac{a}{b} - \frac{b}{a}, & \text{otherwise} \end{cases}$$

2. Volume render $g$

- Provides a fuzzy description of the likelihood of where an isosurface exists
Comparison to Downsampling

<table>
<thead>
<tr>
<th></th>
<th>$4^3$</th>
<th>$8^3$</th>
<th>$16^3$</th>
<th>$32^3$</th>
<th>$64^3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fuzzy iso</td>
<td><img src="image" alt="Fuzzy iso $4^3$" /></td>
<td><img src="image" alt="Fuzzy iso $8^3$" /></td>
<td><img src="image" alt="Fuzzy iso $16^3$" /></td>
<td><img src="image" alt="Fuzzy iso $32^3$" /></td>
<td><img src="image" alt="Fuzzy iso $64^3$" /></td>
</tr>
<tr>
<td>Mean</td>
<td><img src="image" alt="Mean $4^3$" /></td>
<td><img src="image" alt="Mean $8^3$" /></td>
<td><img src="image" alt="Mean $16^3$" /></td>
<td><img src="image" alt="Mean $32^3$" /></td>
<td><img src="image" alt="Mean $64^3$" /></td>
</tr>
<tr>
<td>Lower left</td>
<td><img src="image" alt="Lower left $4^3$" /></td>
<td><img src="image" alt="Lower left $8^3$" /></td>
<td><img src="image" alt="Lower left $16^3$" /></td>
<td><img src="image" alt="Lower left $32^3$" /></td>
<td><img src="image" alt="Lower left $64^3$" /></td>
</tr>
</tbody>
</table>
Fuzzy Isosurface of Temporal Jet

Likelihood that isovalue $\kappa = 0.506$ passes through a hixel
Conclusions and Summary

• Unified representations of large scalar fields from various modalities
• 3 proof of concept applications
  • Sampled topology
  • Topological analysis of statistically associated buckets
  • Visualizing fuzzy isosurfaces
Future Work

- Larger ensembles/larger data
- Performance/scaling
- Infer sheets from multivariate hixels
- Issues to study
  - What is preserved by hixels vs. resolution loss
  - Identify appropriate number of bins/hixel
  - Persistence thresholds for bucketing algorithm
  - Balance data storage vs. feature preservation
  - What topological features can/cannot be preserved by hixelation
Acknowledgement

Contact: Joshua A. Levine
jlevine@sci.utah.edu

This work was supported by the Department of Energy Office of Advanced Scientific Computing Research, award number DE-SC0001922. Sandia is a multi-program laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy’s National Nuclear Security Administration under Contract DE-AC04-94AL85000. This work was also performed under the auspices of the US Department of Energy (DOE) by the Lawrence Livermore National Laboratory under contract nos. DE-AC52-07NA27344, LLNL-JRNL-412904L and by the University of Utah under contract DE-FC02-06ER25781. We are grateful to Dr. Jacqueline Chen for the combustion data sets and M. Eduard Göller, Georg Glaeser, and Johannes Kastner for the stag beetle dataset.