On the Use of Perceptual Cues and Data Mining for Effective Visualization of Scientific Datasets

Christopher G. Healey
EECS Department, CS Division, University of California at Berkeley

Abstract

Scientific datasets are often difficult to analyse or visualize, due to their large size and high dimensionality. We propose a two-step approach to address this problem. We begin by using data mining algorithms to identify areas of interest within the dataset. This allows us to reduce a dataset's size and dimensionality, and to estimate missing values or correct erroneous entries. We display the results of the data mining step using visualization techniques based on perceptual cues. Our visualization tools are designed to exploit the power of the low-level human visual system. The result is a set of displays that allow users to perform rapid and accurate exploratory data analysis.

In order to demonstrate our techniques, we visualized an environmental dataset being used to model salmon growth and migration patterns. Data mining was used to identify significant attributes and to provide accurate estimates of plankton density. We used colour and texture to visualize the significant attributes and estimated plankton densities for each month for the years 1956 to 1964. Experiments run in our laboratory showed that the colours and textures we chose support rapid and accurate element identification, boundary detection, region tracking, and estimation. The result is a visualization tool that allows users to quickly locate specific plankton densities and the boundaries they form. Users can compare plankton densities to other environmental conditions like sea surface temperature and current strength. Finally, users can track changes in any of the dataset's attributes on a monthly or yearly basis.

CR Categories: H.5.2 [Information Interfaces and Presentation]: User Interfaces—ergonomics, screen design, theory and methods; I.3.6 [Computer Graphics]: Methodology and Techniques—ergonomics, interaction techniques; J.2 [Physical Sciences and Engineering]: Earth and Atmospheric Sciences

Keywords: colour, computer graphics, data mining, human vision, knowledge discovery, multidimensional dataset, perception, preattentive processing, scientific visualization, texture.

Introduction

This paper describes our investigation of methods for visualizing certain types of large, multidimensional datasets. These datasets are becoming more and more common; examples include scientific simulation results, geographic information systems, satellite images, and biomedical scans. The overwhelming amount of information contained in these datasets makes them difficult to analyse using traditional mathematical or statistical techniques. It also makes them difficult to visualize in an efficient or useful manner.

The size of a dataset can be divided into three separate characteristics: the number of elements in the dataset, the number of attributes or dimensions embedded in each element, and the range of values possible for each attribute. All three characteristics may need to be considered during visualization.

Our approach to this problem combines an initial data filtering step and a perceptual visualization step. Data mining algorithms are used to identify dependencies, to estimate missing or correct erroneous values, and to compress a dataset's size and dimensionality. The results are displayed to the user in a manner that takes advantage of the low-level human visual system. Offloading the majority of the analysis task on the low-level visual system allows users to very rapidly and accurately perform exploratory visualization on large multidimensional datasets. Trends and relationships, unexpected patterns or results, and other areas of interest can be quickly identified within the dataset. These data subsets can then be further visualized or analysed as required.

Oceanography Simulations

Our current visualization testbed for this work is a set of simulations being run in the Westwater Research Centre at the University of British Columbia. Researchers in oceanography are studying the growth and movement patterns of different species of salmon in the northern Pacific Ocean. Underlying environmental conditions like plankton density, sea surface temperature (SST), current direction, and current strength affect where the salmon live and how they move and grow [19]. For example, salmon like cooler water and tend to avoid ocean locations above a certain temperature. Since the salmon feed on plankton blooms, they will try to move to areas where plankton density is highest. Currents will "push" the salmon as they swim. Finally, SST, current direction, and current strength affect the size and location of plankton blooms as they form.

The oceanographers are designing models of how they believe salmon feed and move in the open ocean. These simulated salmon will be placed in a set of known environmental conditions, then tracked to see if their behaviour mirrors that of the real fish. For example, salmon that migrate back to the Fraser River to spawn chose one of two routes. When the Gulf of Alaska is warm, salmon make landfall at the north end of Vancouver Island and approach the Fraser River primarily via a northern route through the Johnstone Strait (the upper arrow in Figure 1). When the Gulf of Alaska is cold, salmon are distributed further south, make landfall on the west coast of Vancouver Island, and approach the Fraser River primarily via a southern route through the Juan de Fuca Strait (the lower arrow in Figure 1). The ability to predict salmon distributions from prevailing environmental conditions would allow the commercial fishing fleet to estimate how many fish will pass through the Johnstone and Juan de Fuca straits. It would also allow more accurate predictions of the size of the salmon run, helping to ensure that an adequate number of salmon arrive at the spawning grounds.

In order to test their hypotheses, the oceanographers have created a database of SSTs and ocean currents for the region 35° north latitude, 180° west longitude to 62° north latitude, 120° west longitude (Figure 1). Measurements within this region are available at $1^\circ \times 1^\circ$ grid spacings. This array of values exists for each month for the years 1956 to 1964, and 1980 to 1989.

Plankton densities have also been collected and tabulated; these are obtained by ships that take readings at various positions in the

Figure 1: Map of the North Pacific; arrows represent possible salmon migration paths as they pass through the either Johnstone Strait (upper arrow) or the Strait of Juan de Fuca (lower arrow)

ocean. Unfortunately, these measurements are much more sparse than the SST and current values. For the years 1956 to 1964, only 1,542 plankton densities are available. This leaves the oceanographers with a number of problems that need to be addressed before their salmon growth and movement models can be tested.

1. A method of estimating plankton densities is required. Currently, spatial interpolation is used to provide the missing values, but this does not work well for months where few (or no) actual densities are available.

2. The oceanographers would like to know how plankton density is related to other environmental conditions like SST, current direction, and current strength. A similar problem will follow: the determination of how salmon growth and open ocean migration patterns are related to underlying environmental conditions.

3. Finally, a method is needed for visualizing the dataset. This method will be used to display both static (*e.g.,* environmental conditions for a particular month and year) and dynamic results (*e.g.,* a real-time display of environmental conditions as they change over time, possibly with the overlay of salmon locations and movement).

Although the first two problems might be thought to lie outside the scope of visualization, we feel that management of the underlying data is an inherent part of the visualization process, particularly for large and complex datasets. The need for data management has been addressed in numerous papers on visualization [16, 20, 21]. Moreover, this problem was cited as an important area of future research in NSF reports from both the database [15] and visualization communities [14]. To this end, we have implemented extended versions of four data mining algorithms that are designed to address the types of problems present in the oceanography datasets.

After using data mining to process the dataset, we must display it on-screen. We have approached the problems of dataset size and dimensionality by trying to exploit the power of the low-level human visual system. Research in computer vision and cognitive psychology provides insight on how the visual system analyses images. A careful mapping of data attributes to visual features (*e.g.,* colour, intensity, and texture) will allow users to perform rapid visual analysis on their data. We must also avoid visual interference effects that can occur when different visual features are combined at the same spatial location. We are currently conducting experiments on the use of colour and texture for multidimensional data visualization [4]. Results from these experiments are used to visualize the oceanography datasets.

Data Mining

Data mining or knowledge discovery, as it is sometimes referred to, is a relatively new area of database research. Data mining is defined as "the nontrivial extraction of implicit, previously unknown, and potentially useful information from data" [3]. This is done by combining a variety of database, statistical, and machine learning techniques. Different data mining algorithms have been developed to perform different types of analysis. Some algorithms search for repeating patterns or trends in a database. Others classify elements into groups based on their attribute values.

We are interested in data mining algorithms that perform classification. We believe that these algorithms can be used to improve the efficiency of visualizing large, multidimensional datasets. Their advantages are twofold. First, they can be used to reduce the amount of data that needs to be displayed. Second, they can be used to "discover" previously unknown and potentially useful information. For example:

- datasets can be filtered by identifying the subset of elements that participate in a particular relationship, reducing a dataset's size,
- attributes that are significant to a given relationship can be identified and displayed, reducing element dimensionality,
- data elements can be grouped or classified; only the classification value (and possibly a confidence measure) can be displayed, reducing element dimensionality, and
- erroneous attribute values can be identified and missing values can be estimated, increasing a dataset's accuracy.

To test our hypothesis, we implemented four existing techniques, then tested them to see if they offered improved efficiency or usefulness compared to visualization without any form of data management. We chose two algorithms based on decision trees [1, 11], one algorithm based on statistical tables [2], and one algorithm based on rough sets [24].

All four data mining algorithms build their classification rules from a user-supplied training set. The decision tree algorithms begin by identifying significant attributes using chi-squared tests. The attribute that provides the largest information gain is used to partition the root of the tree. This process continues recursively using any remaining attributes. Leaves in the tree hold a single classification value. Unclassified elements match their attribute values against each node in the tree (*i.e,* the attribute values define a path from root to leaf through the tree). The leaf node's classification value is assigned to the element.

The statistical table algorithm uses probabilities to perform classification. For each attribute, a table is built containing every possible *(attribute value, classification value)* pair. Probabilities are computed for each pair. A positive probability suggests

that (based on the training set) the given attribute value implies the given classification value; a negative probability means it implies some other classification value. Given an unknown element, the tables are used to compute probabilities for every possible classification value. The classification with the highest positive probability is assigned to the element.

The rough set algorithm uses set theory and equivalence relations to identify a subset of attributes that group classification values in a manner equivalent to the original attributes in the training set. Each attribute in the subset is assigned a coverage value; higher values imply greater importance during classification. The algorithm can then build rules that map combinations of attribute values to a classification value. Unclassified elements match their attribute values to each rule. The rule with the highest total coverage is used to assign a classification value to the element.

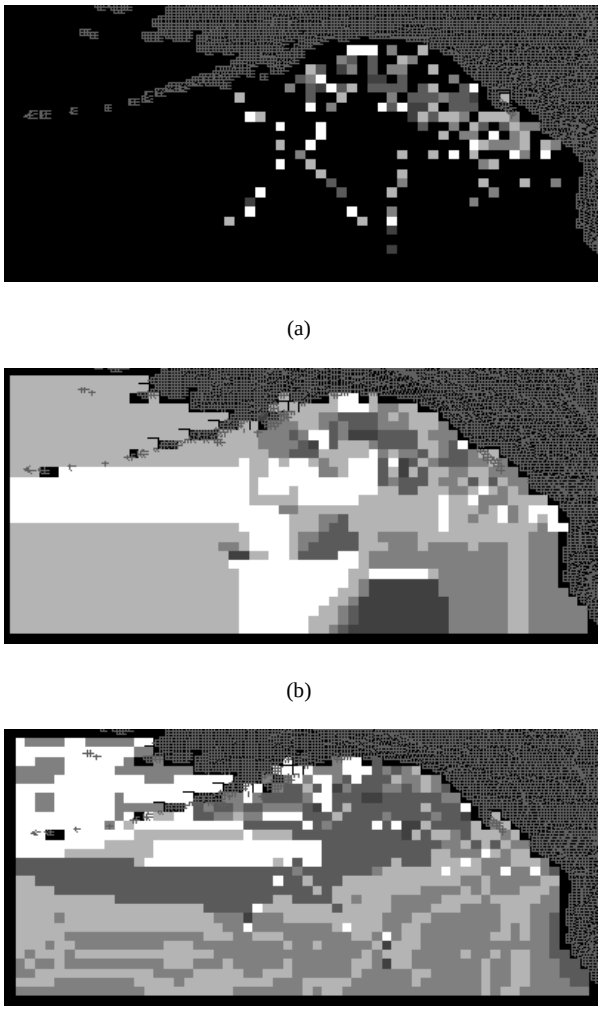
The data mining algorithms are designed to process a training set, then provide classification values for one or more unclassified elements. During visualization, however, users often require more than a simple classification value. We modified and extended the algorithms to provide additional results, in particular, classification confidence weights, the ability to compare different classifications, and the ability to identify attributes that are significant to a specific classification. This allows a user to answer questions like:

- How confident is the algorithm about the classification value it suggests? A confidence weight is returned with each classification value to help answer this question.
- How "good" is the classification value suggested by the data mining algorithm, compared to other potential classification values? Comparing confidence weights shows how the algorithm ranks different classifications.
- Which attributes are significant to the classification and which are not? A significance weight is assigned to each attribute when the classification rules are built; attributes that are ignored during classification have a significance weight of zero.

Results from experiments that tested the extended data mining algorithms were positive. We showed that data mining produced more accurate results than bilinear interpolation on a large environmental dataset. Significance weights identified the most important attributes used during classification. Confidence weights were excellent predictors of classification values that were in error. Data elements with low confidence weights were also used to identify "holes" in the training set. Low confidence weights indicate elements the algorithm did not see during training (and hence is unsure how to classify); once identified, representative elements can be added to the training set, thereby improving the classification power of the algorithm. Finally, we tested each algorithm to see how it performed when errors were introduced into the training set. All of the algorithms continued to perform well with some level of training set error, however, the decision tree algorithms were able to handle significantly more errors while still returning the fewest mistakes as a result. Complete descriptions of the four data mining algorithms, the extensions we developed, and our experimental results are available in [5].

Oceanography Results

Our initial concern for the oceanography datasets was accurate estimation of plankton densities. We created a training set that contained all available density measurements (a total of 1,542 elements) for the years 1956 to 1964. Each of these readings included a latitude, longitude, and the month and year the reading was taken. We added to each element the corresponding SST, current direction, and current strength (these were taken directly from the environment database for the given month, year, latitude, and longitude). Continuous values (SST, current direction and strength, and plankton density) were divided into five equal-width ranges; each value's range was used during classification. Although the data mining algorithms will automatically range continuous data, we found that more accurate results are obtained when a user familiar with the dataset chooses the bounds for each range. The ranges used for SST, current U and V direction, current strength, and plankton density are shown in Table 1.



(a)

(b)

(c)

Figure 2: Known and estimated plankton densities for August 1956, greyscale used to represent density (dark grey for low to white for high): (a) known densities; (b) missing densities estimated using interpolation (note the banding that occurs at the boundaries of the array); (c) missing densities estimated using data mining (patterns within the array correspond to the prevailing SSTs and current strengths)

We started by reading the training set with each of our four data mining algorithms, then using significance weights to identify which attributes were being used to classify (*i.e.,* estimate) plankton density. All four algorithms reported similar results:

SST (C)	SST < 6.34	$6.34 \leq$ SST < 8.98	$8.98 \leq$ SST < 11.82	$11.82 \leq$ SST < 14.80	SST ≥ 14.80
Current U	U < -0.6	$-0.6 \leq$ U < -0.2	$-0.2 \leq$ U < 0.2	$0.2 \leq$ U < 0.6	U ≥ 0.6
Current V	V < -0.6	$-0.6 \leq$ V < -0.2	$-0.2 \leq$ V < 0.2	$0.2 \leq$ V < 0.6	V ≥ 0.6
Strength (cm/s)	Str < 6.087	$6.087 \leq$ Str < 9.015	$9.015 \leq$ Str < 11.567	$11.567 \leq$ Str < 14.542	Str ≥ 14.542
Plankton (g/m^3)	Plk < 10	$10 \leq$ Plk < 28	$28 \leq$ Plk < 53	$53 \leq$ Plk < 114	Plk ≥ 114

Table 1: Boundaries used to divide SST (measured in degrees Celsius), normalized current U and V direction, current strength (measured in centimetres per second), and plankton density (measured in grams per metre cubed) into five equal-width ranges.

month was the most important attribute to use during classification, followed by current strength and SST. Other attributes (current direction and year) had a significance weight of zero. The oceanographers concurred with these results; plankton densities display a seasonal variability, large current upwellings will produce larger plankton blooms, and higher ocean temperatures cause faster plankton production and higher overall densities. These results allowed us to restrict our visualizations to month, SST, strength, and plankton density. The oceanographers searched these displays for temperature and current patterns, and their relationship to the corresponding plankton densities.

Once rules are built from the training set, each data mining algorithm can assign an estimated plankton density to unknown ocean positions based on SST, current strength, and month. This was done for all missing plankton densities for the years 1956 to 1964. We used the interval classification algorithm [1], since it showed the smallest sensitivity to errors in its training set during prior testing [5]. Approximately 11% of the estimated plankton densities exhibited low confidence weights. Although these elements are included during visualization, we plan to examine them in isolation, to try to determine why the data mining algorithm had difficulty assigning them a density value. Initial investigation suggests that elements with certain combinations of month, SST, and current strength were not present in our training set. As a result, the data mining algorithms were uncertain about how to analyse these kinds of elements during classification.

An example of our results is shown in Figure 2. The plankton densities that were actually available are shown in Figure 2a. Figure 2b shows missing values that have been estimated using spatial interpolation. As expected, this technique performs poorly for locations in the ocean where no initial values are present. Most of the northwest and southwest quadrants have been classified to have moderate density; there is almost certainly more variation in this region. Data mining, on the other hand, uses the month, along with the underlying SSTs and current strengths, to estimate plankton density. In Figure 2c, the northwest and southwest quadrants have variability similar to that which exists across the known densities (Figure 2a). Although it is impossible to conclude that the values provided by the data mining algorithm are "more correct" than the interpolated values, our algorithms are not at a disadvantage when no real data values neighbour the value we want to estimate.

Perceptual Visualization

Researchers in computer vision and cognitive psychology are studying how the low-level visual system analyses images. One very interesting result has been the discovery of a limited set of visual features that are processed preattentively, without the need for focused attention. These features can be used to perform certain visual tasks very rapidly and accurately. Examples include searching for elements with a unique visual feature, identifying the boundaries between groups of elements with common features, tracking groups of elements as they move in time and space, and estimating the number of elements with a specific feature. These tasks are preattentive because they can be performed on large multi-element displays in less than 200 msec. Moreover, the time required to complete the tasks is independent of the number of data elements being displayed. Eye movements take at least 200 msec to initiate, and random locations of the elements in the display ensure that attention cannot be prefocused on any particular location, yet subjects report that these tasks can be completed with very little effort. This suggests that certain information in the display is processed in parallel by the low-level visual system.

Our interest is focused on identifying relevant results in the vision and psychology literature, then extending these results and integrating them into a visualization environment. We are currently studying perceptual aspects of colour, orientation, and texture. Results from our experiments allow us to build visualization tools that use these visual features to effectively represent multidimensional datasets. Because our tools take advantage of the low-level visual system, they offer a number of important advantages:

- Visual analysis is rapid and accurate, since preattentive tasks need an exposure duration of 200 msec or less. We have shown that tasks performed on static frames extended to a dynamic environment, where frames are shown one after another in a movie-like fashion [6] (*i.e.*, tasks that can be performed on a single frame in 200 msec can also be performed on a sequence of frames shown at five frames per second).
- Our tasks are insensitive to display size (to the limits of the display device); increasing the number of elements in the display results in little or no increase in the amount of time required to visually analyse the display. Again, this is a direct result of the fact that preattentive tasks are independent of display size.
- Certain combinations of visual features cause interference patterns in the low-level visual system that can mask information in a display. Our experiments were designed to identify these situations. This means our visualization tools can be built to avoid data-feature mappings that might interfere with the analysis task.

We chose to use two well-known features to visualize the oceanography datasets: colour and texture. Research in our laboratory has studied the use of both features during visualization. For colour, we conducted a number of experiments to determine how to choose colours that are equally distinguishable from one another. That is, we want to pick n colours which, when displayed simultaneously, allow the user to identify the presence or absence of any one of the colours. Our results showed that three criteria must be considered during colour selection [4]:

- *colour distance:* the distance from each colour to its nearest neighbour(s) is equal and above a minimum threshold; dis-

tance is measured in a perceptually balanced colour model (in our case, CIE LUV),

- *linear separation:* each colour must be linearly separable from all the other colours, again by a minimum threshold measured in a perceptually balanced colour model, and
- *colour category:* each colour must occupy a uniquely named colour region.

We found that when these rules are satisfied, up to seven isoluminant colours can be displayed simultaneously. A user can quickly determine whether any one of the seven colours is present or absent in a given display. Work in progress is studying how to integrate intensity and fully saturated colours into our model. This will allow users a wider range of colours to choose from, and may also increase the maximum number of colours that can be simultaneously displayed and identified.

Experiments are also being run to study the use of perceptual texture elements (or pexels) for multidimensional data visualization. Texture has been studied extensively in the computer vision and psychology communities [7, 12, 13, 18]. A number of visualization systems that use texture have been described, including EXVIS [10], the use of Wold features [9], the use of Markov random fields [8], and studies of the fundamental dimensions of a texture element [23].

We are interested in using pexels to visualize multidimensional datasets. As opposed to "texture maps" (patterns that are mapped onto regions of a graphical object), perceptual textures are arrays of elements with visual and spatial characteristics that are controlled by the underlying data being displayed. Research results suggest certain perceptual "dimensions" can be varied to control the appearance of the texture formed by the elements, for example:

- *density:* how closely elements are packed together,
- *height:* how tall or short an element is,
- *orientation:* how an element "sits up" or "lies down" or "spins" on the surface it's connected to, and
- *randomness:* whether elements are spatially arranged as a regular grid, or with a random distribution.

Our experiments are testing the use of height, density, and randomness to display multidimensional data. Our pexels look like paper strips; at each data position, a pexel is displayed. The user maps attributes in the dataset to the density (which controls the number of strips in each pexel), height, and randomness of each pexel. Examples of each of these perceptual dimensions are shown in Figure 3. We are also testing for visual interference, feature preference, and target region size (*i.e.,* how many pexels does a region need to contain before it can be rapidly identified).

Figure 4a shows an environmental dataset visualized with texture and greyscale (we used greyscale for printing purposes only; colour is used to display on-screen images). Locations on the map that contain pexels represent areas in North America with high levels of cultivation. Height shows the level of cultivation (75-99% for short pexels, 100% for tall pexels), density shows the ground type (sparse for alluvial, dense for wetlands), and greyscale shows the vegetation type (dark grey for plains, light grey for forest, and white for woods). Users can easily identify lower levels of cultivation in the central and eastern plains. Areas containing wetlands can be seen as dense pexels in Florida, along the eastern coast, and in the southern parts of the Canadian prairies. Figure 4b shows a map of central Japan and the Korean peninsula. As in Figure 4a, height is mapped to cultivation level and greyscale is mapped to

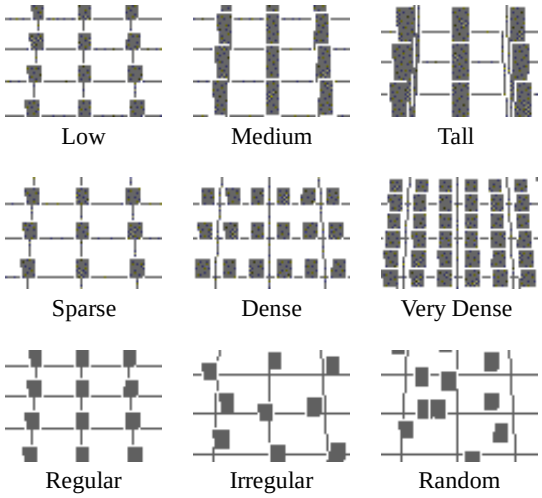


Figure 3: Variation of perceptual texture dimensions height (top row), density (middle row), and randomness (bottom row) across three discrete values

vegetation type. In this image, however, randomness is mapped to ground type: regular for alluvial, and irregular for wetlands. Wetlands (*i.e.,* pexels with random placement) can be seen in the northwestern regions of the peninsula.

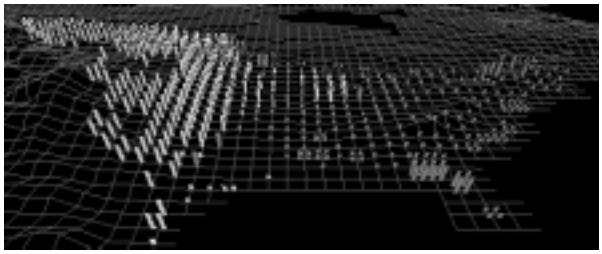
Although the experiments are still being run, preliminary results show that perceptual textures can be used to display multidimensional data. We have also compiled initial information on feature preference, feature interference, and the region size required for rapid identification. These results were used when we designed tools to visualize the oceanography datasets.

Oceanography Visualization

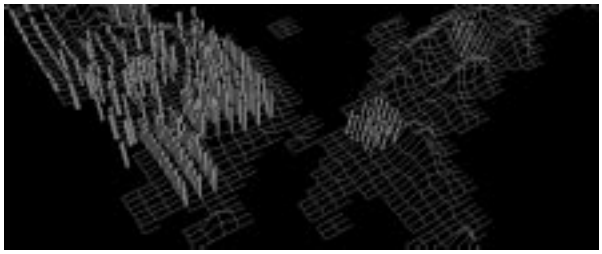
We chose to visualize SST and current strength with plankton density, since these attributes (along with month) were significant during data mining. Displaying the three attributes together allows the oceanographers to search for relationships between plankton density, current strength, and SST. Plankton is displayed using colour; SST and current strength are displayed using texture. Colours for the five plankton ranges were chosen using our colour selection technique [4]. Although other colour scales were available (for example, by Ware [22]), our colours are specifically designed to highlight outliers, and to show clearly the boundaries between groups of elements with a common plankton density. We display the five plankton density ranges from low to high using blue (monitor RGB=36, 103, 151), green (monitor RGB=18, 127, 45), brown (monitor RGB=134, 96, 1), red (monitor RGB=243, 51, 55), and purple (monitor RGB=206, 45, 162),

For the underlying texture, we mapped current strength to height and SST to density. Our choices were guided by results we observed from tests run during the design of our texture experiments, specifically:

- differences in height may be easier to detect, compared to differences in density or randomness,
- variation in height may mask differences in density or randomness; this appears to be due to the occlusion that occurs when tall pexels in the foreground hide short pexels in the background; this will be less important when users can control their viewpoint into the dataset (our visualization tool allows the user to interactively manipulate the viewpoint), and



(a)



(b)

Figure 4: Using pexels to display environmental conditions; (a) a map of North America, pexels represent areas of high cultivation, height mapped to level of cultivation, density mapped to ground type, greyscale mapped to vegetation type; (b) a map of Japan and the Korean peninsula, height and greyscale mapped to cultivation and vegetation, randomness mapped to ground type

- tightly spaced grids can support up to three easily distinguishable density patterns; placing more strips in a single pexel (*e.g.*, arrays of three by three or four by four strips) will either cause the strips to overlap with their neighbours, or make each strip too thin to easily identify.

Because there may be a feature preference for height over density, and because current strength was deemed "more important" than SST during data mining, we used height to represent currents and density to represent SSTs. The five ranges of current strength are mapped to five different heights. We do not use a linear mapping, rather the lower two ranges (corresponding to the weakest currents) are displayed using two types of short pexels, and the upper three ranges (corresponding to the strongest currents) are displayed using three types of tall pexels. This allows a user to rapidly locate boundaries between weak and strong currents, while still being able to identify each of the five ranges. For SSTs, the lower three ranges (corresponding to the coldest SSTs) are displayed with a pexel containing a single strip, while the upper two ranges (corresponding to the warmest SSTs) are displayed with pexels containing arrays of two and four strips, respectively. The densities we chose allow a user to see clearly the boundaries between cold and warm temperature regions. If necessary, users can change the range boundaries to focus on different SST gradients.

The oceanographers want to traverse their datasets in monthly and yearly steps. Experiments run in our laboratory have shown that preattentive tasks performed on static frames can be extended to a dynamic environment, where displays are shown one after another in a movie-like fashion [6]. Our visualization tool was designed to allow users to scan rapidly forwards and backwards through the dataset. This makes it easy to compare changes in the value and location of any of the environmental variables being displayed. The oceanographers can track seasonal changes in current strength, SST, and plankton density as they move month by month through a particular year. They can also see how interannual variability affects the environmental conditions and corresponding plankton densities for a particular month across a range of years.

Figure 5 shows three frames from the oceanography dataset: February 1956, June 1956, and October 1956. Colour shows the seasonal variation in plankton densities. Height and density allow the oceanographers to track current strengths and SSTs. In February (Figure 5a), most plankton densities are less than 28 g/m^3 (*i.e.*, blue and green strips). Currents are low in the north-central Pacific; a region of weak currents also sits off the south coast of Alaska. Most of the ocean is cold (sparse pexels), although a region of higher temperatures can easily be seen as dense pexels in the south. In June (Figure 5b) dense plankton blooms (red and purple strips) are present across most of the northern Pacific. The positions of the strong currents have shifted (viewing the entire dataset shows this current pattern is relatively stable for the months March to August). Warmer SSTs have pushed north, although the ocean around Alaska and northern British Columbia is still relatively cold. By October the plankton densities have started to decrease (green, brown, and red strips); few high or low density patches are visible. Current strengths have also decreased in the eastern regions. Overall a much larger percentage of the ocean is warm (*i.e.*, dense pexels). This is common, since summer temperatures will sometimes last in parts of the ocean until October or November.

Conclusions

This paper described our two-step approach to visualizing complex scientific datasets. We begin by using data mining algorithms to identify significant trends, classify elements, and focus the dataset. The results are then visualized using perceptual features. We demonstrated our techniques by analysing and visualizing an environmental dataset being used to run salmon growth and migration simulations. We used data mining to estimate missing plankton densities, and to identify the attributes significant to this estimation. The resulting sea surface temperatures, ocean current strengths, and plankton densities were visualized using colour and texture. The colours and textures (built as arrays of paper strips with varying height and density) were chosen based on results from perceptual experiments run in our laboratory. We exploit the low-level human visual system with our visualization tools. This makes a large part of the visual analysis automatic; little effort or focused attention is required by the user to perform exploratory tasks like target identification, boundary detection, region tracking, and estimation. These tasks can be carried out on sequences of displays shown one after another at relatively high frame rates (*e.g.*, 100 to 200 msec per frame). This technique allows the oceanographers to scan through their datasets month by month or year by year to view seasonal or interannual changes in environmental conditions.

Experiments studying the use of height, density, and randomness to generate perceptual textures are still in progress. Once completed, we believe the results will allow us to increase the flexibility and effectiveness of our visualization tools. The oceanographers will begin testing salmon growth and migration models in the near future. We plan to use data mining to try to relate environment conditions to the simulated salmon, and to visualize the salmon as they move through the open ocean.

Although our practical example in this paper was an oceanographic dataset, data mining and perceptual visualization can be applied to a wide range of visualization environments. We have

used perceptual colour selection to highlight regions of interest in reconstructed medical volumes [17]. We have also used data mining to estimate sea surface temperatures in an environmental dataset from NASA [5]; we showed that our results were more accurate than estimates produced by bilinear interpolation. We will continue to test the flexibility of our techniques with new visualization problems and datasets.

Acknowledgments

I would like to thank Dr. Peter Rand and Dr. Michael Healey for access to their data and their expertise from the salmon growth simulations. Dr. James Enns, Dr. Vince Di Lollo, and Dr. Kellogg Booth provided valuable technical advice during my research. I would also like to thank Jeanette Lum for coordinating and running our experiment sessions. Maryann Simmons and Randy Keller offered important feedback which improved the organization and presentation of this paper. This research was funded in part by the National Science and Engineering Research Council of Canada, and by the Office of Naval Research (Grant N00014-96-1120) and the Ballistic Missile Defense Organization through the Multiuniversity Research Initiative.

References

[1] AGRAWAL, R., GHOSH, S., IMIELINSKI, T., IYER, B., AND SWAMI, A. An interval classifier for database mining applications. In *Proceedings 18th Very Large Database (VLDB) Conference* (1992), pp. 560–573.

[2] CHAN, K. C. C., AND WONG, A. K. C. A statistical technique for extracting classificatory knowledge from databases. In *Knowledge Discovery in Databases*, G. Piatetsky-Shapiro and W. J. Frawley, Eds. AAAI Press/MIT Press, Menlo Park, California, 1991, pp. 107–123.

[3] FRAWLEY, W. J., PIATETSKY-SHAPIRO, G., AND MATHEUS, C. J. Knowledge discovery in database: An overview. In *Knowledge Discovery in Databases*, G. Piatetsky-Shapiro and W. J. Frawley, Eds. AAAI Press/MIT Press, Menlo Park, California, 1991, pp. 1–27.

[4] HEALEY, C. G. Choosing effective colours for data visualization. In *Proceedings Visualization '96* (San Francisco, California, 1996), pp. 263–270.

[5] HEALEY, C. G. *Effective Visualization of Large, Multidimensional Datasets*. Ph.D. thesis, The University of British Columbia, Canada, 1996.

[6] HEALEY, C. G., BOOTH, K. S., AND ENNS, J. T. Real-time multivariate data visualization using preattentive processing. *ACM Transactions on Modeling and Computer Simulation 5*, 3 (1995), 190–221.

[7] JULÉSZ, B. Experiments in the visual perception of texture. *Scientific American* (April, 1975), 34–43.

[8] LI, R., AND ROBERTSON, P. K. Towards perceptual control of Markov random field textures. In *Perceptual Issues in Visualization*, G. Grinstein and H. Levkowitz, Eds. Springer-Verlag, New York, New York, 1995, pp. 83–94.

[9] LIU, F., AND PICARD, R. W. Periodicity, directionality, and randomness: Wold features for perceptual pattern recognition. In *Proceedings 12th International Conference on Pattern Recognition* (Jerusalem, Israel, 1994), pp. 1–5.

[10] PICKETT, R., AND GRINSTEIN, G. Iconographic displays for visualizing multidimensional data. In *Proceedings of the 1988 IEEE Conference on Systems, Man, and Cybernetics* (Beijing and Shenyang, China, 1988), pp. 514–519.

[11] QUINLAN, J. R. Induction of decision trees. *Machine Learning 1*, 1 (1986), 81–106.

[12] RAO, A. R., AND LOHSE, G. L. Identifying high level features of texture perception. *CVGIP: Graphics Models and Image Processing 55*, 3 (1993), 218–233.

[13] REED, T. R., AND HANS DU BUF, J. M. A review of recent texture segmentation and feature extraction techniques. *CVGIP: Image Understanding 57*, 3 (1993), 359–372.

[14] ROSENBLUM, L. J. Research issues in scientific visualization. *IEEE Computer Graphics & Applications 14*, 2 (1994), 61–85.

[15] SILBERSHATZ, A., STONEBRAKER, M., AND ULLMAN, J. D. The "Lagunita" report of the NSF invitational workshop on the future of database systems research. Tech. Rep. TR-90-22, Department of Computer Science, University of Austin at Texas, 1990.

[16] STONEBRAKER, M., CHEN, J., NATHAN, N., PAXSON, C., SU, A., AND WU, J. Tioga: A database-oriented visualization tool. In *Proceedings Visualization '93* (San Jose, California, 1993), pp. 86–93.

[17] TAM, R., HEALEY, C. G., AND FLAK, B. Volume visualization of abdominal aortic aneurysms. In *Proceedings Visualization '97* (Phoenix, Arizona, 1997), pp. 43–50.

[18] TAMURA, H., MORI, S., AND YAMAWAKI, T. Textural features corresponding to visual perception. *IEEE Transactions on Systems, Man, and Cybernetics SMC-8*, 6 (1978), 460–473.

[19] THOMSON, K. A., INGRAHAM, W. J., HEALEY, M. C., LEBLOND, P. H., GROOT, C., AND HEALEY, C. G. Computer simulations of the influence of ocean currents on Fraser River sockeye salmon (*oncorhynchus nerka*) return times. *Canadian Journal of Fisheries and Aquatic Sciences 51*, 2 (1994), 441–449.

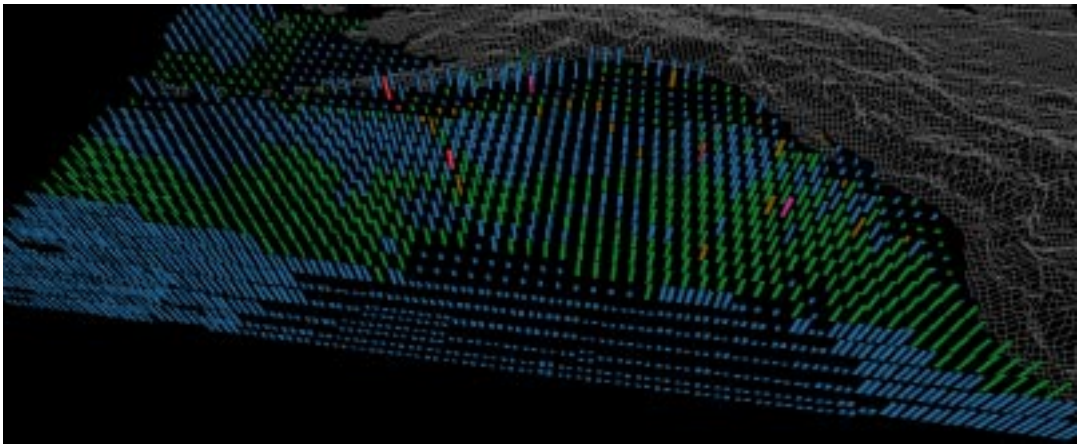
[20] TREINISH, L. A. Unifying principles of data management for scientific visualization. In *Animation and Scientific Visualization*, R. Earnshaw and D. Watson, Eds. Academic Press, New York, New York, 1993, pp. 141–170.

[21] TREINISH, L. A., FOLEY, J. D., CAMPBELL, W. J., HABER, R. B., AND GURWITZ, R. F. Effective software systems for scientific data visualization. *Computer Graphics 23*, 5 (1989), 111–136.

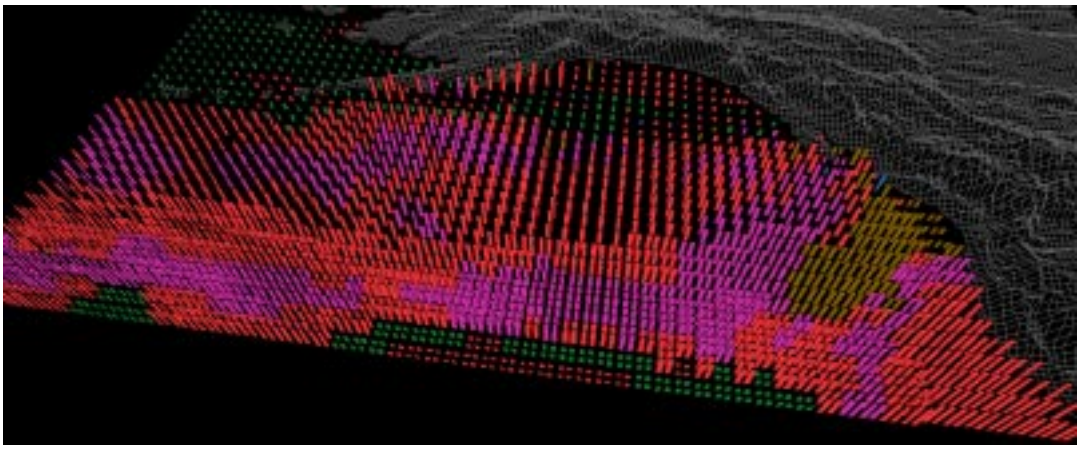
[22] WARE, C. Color sequences for univariate maps: Theory, experiments, and principles. *IEEE Computer Graphics & Applications 8*, 5 (1988), 41–49.

[23] WARE, C., AND KNIGHT, W. Using visual texture for information display. *ACM Transactions on Graphics 14*, 1 (1995), 3–20.

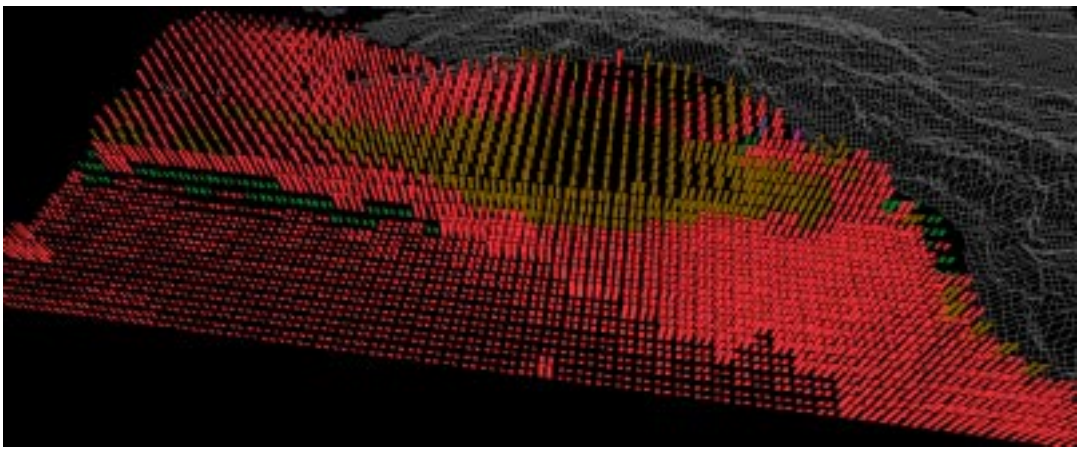
[24] ZIARKO, W. The discovery, analysis, and representation of data dependencies in databases. In *Knowledge Discovery in Databases*, G. Piatetsky-Shapiro and W. J. Frawley, Eds. AAAI Press/MIT Press, Menlo Park, California, 1991, pp. 195–209.



(a)



(b)



(c)

Figure 5: Visualization of the oceanography datasets, colour used to represent plankton density (blue, green, brown, red, and purple represent lowest to highest densities), height used to represent current strength, texture density used to represent SST: (a) February, 1956; (b) June, 1956; (c) October, 1956