Final Year Thesis (HUA3)

Visual Analysis of Air Pollution Problem in Hong Kong

CHAN Wing Yi, Winnie

Supervised by Professor Huamin QU

Department of Computer Science and Engineering
The Hong Kong University of Science and Technology

May 11, 2007
Contents

• Introduction
  ▫ Background and Motivations
  ▫ Weather Data
  ▫ Challenges

• Related Work

• System Overview

• Visualization Techniques

• Experimental Results

• Conclusion and Future Work
Introduction (1)

- We are now experiencing information explosion
- Hard for knowledge discovery when data-sets are too large using solely plain text and tables
- Information visualization
  - Presents abstract and non-physically based data visually and interactively
  - Helps users to detect the expected and gain insight into the unexpected
  - Harnesses human visual perception capabilities
Introduction (2)

• Multivariate data visualization
  ▫ Visualizes data containing multiple attributes

• Weather data visualization
  ▫ A concrete type of multivariate data visualization
  ▫ Visualizes environmental / weather data

• Visual analysis / visual analytics
  ▫ Visual way for data mining and decision making
  ▫ Analytical reasoning facilitated by interactive visual interfaces
Background and Motivations (1)

- Hong Kong air quality decreasing tremendously
- Air pollution problem becomes one of the biggest social issues
- Causes still unknown - Many hypotheses proposed without any formal proof yet

Hong Kong on a better day already. The spectacular harbor view has been increasingly crippled by massive haze.
Background and Motivations (2)

• Institute for the Environment of HKUST
  ▫ One of the major efforts in studying air pollution
  ▫ Developed a comprehensive atmospheric and environmental database on Hong Kong and surrounding regions
  ▫ Found correlations with classical analysis techniques
  ▫ Failed to obtain convincing results for high-level correlations
  ▫ Demanded visualization techniques for analysis
Weather Data

- Recorded by automatic monitoring stations located in representative regions at regular time intervals

- Special features:
  - Time-series (hourly-based)
  - Contains inherited geographic information
  - Multivariate (typically more than 10 dimensions)
  - Important vector field - wind speed and direction
Challenges

• Visualization desirable but not trivial to do so:
  ▫ People too familiar with existing tools to represent the wind profile
    • E.g. polar coordinates and orientated arrows
    • Constraints the design of visualization tool
  ▫ Large data size of high dimensionality
    • Not easy for effective and efficient visual analytic
  ▫ How to handle multivariate time-series data
    • Need to support comparisons across time and station
    • Could have time delays
    • Different stations may exhibit similar patterns at different points in time
Contents

• Introduction
• Related Work
• System Overview
• Visualization Techniques
• Experimental Results
• Conclusion and Future Work
Related Work

- Rarely considered as a standalone problem
- Studied in multivariate data visualization
- Uniqueness of weather data sometimes overlooked
  - Vector value lost
  - Geographic information ignored
  - Time-series properties represented rather tediously by showing a number of plots
Related Work (1) - Treinish

- More on simulating the weather condition, rather than visualizing the data
Related Work (2) - Textures

- Maps each attribute to individual visual channel, e.g.
  - Wind $\rightarrow$ Orientation
  - Temperature $\rightarrow$ Luminance
  - Pressure $\rightarrow$ Scale
- Low scalability: at most 4 dimensions

[ Tang et. al ]

[ Healey et. al ]
Related Work (3)

- General multivariate application

[ Wilkinson et. al ]

[ Luo et. al ]

[ Guo et. al ]
Contents

• Introduction
• Related Work
• System Overview
  ▫ Data Collection
  ▫ Visualization Tasks
• Visualization Techniques
• Experimental Results
• Conclusion and Future Work
Data Collection

- By the Environment Facility Center (ENVF) of HKUST
  - Contains more than 13 dimensions
  - Spans more than 10 years
## Different Stations and Their Data

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>North</td>
</tr>
<tr>
<td>2.</td>
<td>Yuen Long</td>
</tr>
<tr>
<td>3.</td>
<td>Tuen Mun</td>
</tr>
<tr>
<td>4.</td>
<td>Tai Po</td>
</tr>
<tr>
<td>5.</td>
<td>Tsuen Wan</td>
</tr>
<tr>
<td>6.</td>
<td>Sha Tin</td>
</tr>
<tr>
<td>7.</td>
<td>Kwai Tsing</td>
</tr>
<tr>
<td>8.</td>
<td>Wong Tai Sin</td>
</tr>
<tr>
<td>9.</td>
<td>Sham Shui Po</td>
</tr>
<tr>
<td>10.</td>
<td>Sai Kung</td>
</tr>
<tr>
<td>11.</td>
<td>Kwun Tong</td>
</tr>
<tr>
<td>12.</td>
<td>Kowloon City</td>
</tr>
<tr>
<td>13.</td>
<td>Yau Tsim Mong</td>
</tr>
<tr>
<td>14.</td>
<td>Eastern</td>
</tr>
<tr>
<td>15.</td>
<td>Wan Chai</td>
</tr>
<tr>
<td>16.</td>
<td>Central &amp; Western</td>
</tr>
<tr>
<td>17.</td>
<td>Southern</td>
</tr>
<tr>
<td>18.</td>
<td>Islands</td>
</tr>
</tbody>
</table>

### Data Categories
- Precipitation
- Wind Direction
- Air Temperature
- Wind Speed
- Dew Point
- Relative Humidity
- Sea Level Pressure
- Respirable Suspended Particulates (RSP)
- Nitrogen oxide (NO)
- Nitrogen dioxide (NO₂)
- Nitrogen oxides (NOₓ)
- Sulphur dioxide (SO₂)
- Ozone (O₃)
- Carbon monoxide (CO)
- Solar Radiation
- Air Pollution Index (API)
- Contributed Pollutant to API
Visualization Tasks

• Finding correlations between different attributes
  ▫ E.g. correlations between air pollution index (API) and pollutants for pinpointing air pollution sources

• Comparing data from different stations
  ▫ Examine similarity or difference at different locations
  ▫ Geographic information can affect the weather behavior

• Detecting the trend for Hong Kong’s weather and air quality
  ▫ Predict the future tendency based on the pattern we observe today
Contents

• Introduction
• Related Work
• System Overview
• Visualization Techniques
  ▫ Polar System
    • Circular Pixel Bars
    • Time-Series Polar System
  ▫ Parallel Coordinates
  ▫ Weighted Complete Graph
• Experimental Results
• Conclusion and Future Work
Our Approach

• Integrate well established visualization techniques into a comprehensive system

• Develop novel techniques specifically designed for weather data
  ▫ Polar system with embedded circular pixel bar charts
    • Detects correlations between wind direction, wind speed and other scalar attributes
  ▫ Parallel coordinates with vector and time axes
  ▫ Weighted complete graph
    • Shows the overall correlation of all data dimensions
    • Determines the order of axes in parallel coordinates
Polar System

- One of the most common representations for vectors
- Low learning curve for domain scientists
  - Heavily applied in the environmental area
- Wind speed and direction frequently used as key

Distance from the center ➔ Wind Speed
Angle from the north ➔ Wind Direction
Pixel Color ➔ Scalar Attribute
Area-Preserving Mapping

- Common practice in the environmental field to generate more reliable display
- Area-preserving mapping on distance from the center
- Points located closer to the center not overcompressed
- Simplest: take the square root
Circular Pixel Bars

- X-position
- Y-position
- Pixel color

- Extended from Pixel Bar
- Users select a sector to plot the circular pixel bar on the data items falling inside the sector region, i.e. lying in a certain range of wind direction and speed
- Complement circular pixel bar blended underneath
Circular vs. Regular Pixel Bars

- Circular plots arranged intuitively on wind direction and speed
- Although accuracy of data analysis may be diminished due to the circular shape
  - Overall patterns preserved in the sector for rapid comparison
  - Numerical analysis on supplement rectangular pixel bars
Polar System with Time Domain

- X-position ➔ Month
- Y-position ➔ SO₂
- Color ➔ Temperature

- X-position ➔ Day
- Y-position ➔ SO₂
- Color ➔ Temperature

- X-position ➔ Month
- Y-position ➔ Day
- Color ➔ Temperature
Contents

• Introduction
• Related Work
• System Overview
• Visualization Techniques
  ▫ Polar System
  ▫ Parallel Coordinates
  ▫ Weighted Complete Graph
    • Definition and Distance Metrics
    • Encoding Scheme
    • Axis Order Selection for Parallel Coordinates
• Experimental Results
• Conclusion and Future Work
Parallel Coordinates

- Well-established visualization tool for multivariate data
- Each parallel vertical axis represents an attribute
- Data item plotted by a polygonal line intersecting each axis at respective attribute data value
S-Shape Axis for Vector

- Traditional straight-line axis not good for encoding vectors and directions
- S-shape axis introduced
  - More natural to represent wind direction
  - Stands out among all axes, attracting user’s attention
Parallel Coordinates with Scatterplot

Enhanced Parallel Coordinates with S shape axis to encode wind direction and scatterplot to reveal bivariate relationship between neighbor axes.
Weighted Complete Graph

- For exploring overall relationship among all data dimensions
- Each node represents one data dimension
- Distance between nodes encodes correlation between adjacent nodes
  - Use LinLog energy model with Barnes-Hut algorithm
  - Strongly correlated nodes located closer to each other
Definition & Distance Metrics

- **Weighted**: each edge associated with weight
  - Strength of correlations between two nodes
- **Complete**: graph complete, each pair of nodes connected by an edge
  - Correlations between any two attributes are of interest
- Standard correlation coefficient used for computing correlations:

\[
C_s(X_i, Y_j) = \frac{\| (X_i - \bar{X}_i)(X_j - \bar{X}_j)^T \|}{((X_i - \bar{X}_i)(X_i - \bar{X}_i)^T)^{\frac{1}{2}}((X_j - \bar{X}_j)(X_j - \bar{X}_j)^T)^{\frac{1}{2}}}
\]
Encodings Scheme

- Weight of edge encodes correlation between adjacent nodes
  - Edges eliminated by setting thresholds to avoid visual clutters
  - Reinforces users’ interpretation and perception
    - E.g. pattern, width, color of edges

- Size of node encodes accumulated correlation coefficients with other attributes
  - A bigger node likely to have strong relationship with other nodes

Color (brightness) encodes correlation measures - Sharp red color represent high correlation.
Axis Order Selection for Parallel Coord.

- Different orders of axes in parallel coordinates could reveal different patterns
  - Order of axes critically important
  - Axes of attributes with potential correlations should be placed closer for better results
- How to determine optimal axis order from the weighted complete graph
  - **Manually**: user decide the order manually
  - **Automatically**: find the shortest path in the graph to maximize possible correlations
Axis Order Selection - Example

- Data with only 13 dimensions, manual selection feasible
- Users manually select the order of nodes in the weighted graph
- Corresponding parallel coordinates generated with color encoding API
  - Attributes on the left strongly correlated, yielding clear clusters
Contents

- Introduction
- Related Work
- System Overview
- Visualization Techniques
- Experimental Results
  - Correlation Detection
  - Similarities and Difference
  - Time-Series Trend
- Conclusion and Future Work
Correlation Detection 1 - Polar

- Finding the correlation between Air Pollution Index (API) and Respirable Suspended Particulates (RSP) with solar radiation, SO$_2$ and O$_3$
- RSP correlated with SO$_2$ and O$_3$, not solar radiation
- High API value (red pixels) not found when SO$_2$ is high, revealing SO$_2$ contributed little to API
- API strongly correlated with O$_3$ which is known to experts
- Suspicious clusters are shown in [SO$_2$] and [O$_3$] - a blue cluster behind a green one, immediately holding domain experts’ attention
Correlation Detection 2 - Parallel

• Color denotes API value
• Gradual color change perceived at RSP and O₃ as expected, indicating they are positively correlated with API
• High API reading does not necessarily attribute to a large amount of SO₂, as shown by group of red lines
• Solar radiation and temperate not related to API suggested by messy lines
• NO₂ and CO / NO and NOₓ display partial relationships worth investigating
• Correlations between multiple dimensions can be explored more easily from parallel coordinates than polar system
Similarities and Differences

• The Hong Kong society mostly weighs external pollution factors more
  ▫ Air pollutants blown in from factories on the Pearl River Delta located at the northwest of Hong Kong

• Local pollution often ignored
  ▫ Monopolistic power plants
  ▫ Excessive number of vehicles and vessels
Similarities and Differences 1

- 9 stations of 3 years data
- Color represents amount of SO$_2$
- Large SO$_2$ amount with strong northwest wind in most stations (blown from external source)
- Station Kwai Chung has the highest SO$_2$ value with southwest wind of all wind speed (internal)
  - Energy sector and vehicular exhaust as major emission sources of SO$_2$
  - Due to cargo ships at Kwai Tsing Container Terminals
Similarities and Differences 2

- Sector with high API $\text{SO}_2$ value selected
- Kwai Chung data generally shows a higher API value for higher recorded $\text{SO}_2$ values than Tung Chung station
  - Recall: $\text{SO}_2$ is not the main pollutant contributing to API
  - Local pollution resulted from heavy $\text{SO}_2$ emission by vessels is dominating in the Kwai Chung region
Similarities and Differences 3

- Tung Chung
  - API strongly related to the wind direction suggested by clusters of **red** and **blue** lines (north / northwest winds) at API axis
- Kwai Chung
  - Noticeable **yellowish** lines (southwest winds) marks highest API
  - Some **cyan** (east winds) lines gives high O$_3$ value
Time-Series Trend

• Weather varies with time in seasonal basis; useful for short-term forecasting
• Trends observed over time when the global climate is changing in the long-run
Time-Series Trend 1: Three Years

- Typical subtropical region with distinguishable seasons
  - Direction of winds opposes each other
- Higher API (color) in winter than in summer
- No obvious growing trend for API value
Time-Series Trend 2: Kwai Chung

- Prominent red pixels are mainly seen in year 2004 plot only
  - Local pollution from SO$_2$ emission was significant
- Slight improvement observed in the following years: lower API
  - Local pollution has become less dominating
Time-Series Trend 3: Time of Day

- Mongkok: \((x, y, \text{color}) \rightarrow (\text{day}, \text{hour}, \text{API})\)
- Year 2005 generally has less severe air pollution
- High API (red pixels) tended to appear in the afternoon and is mostly found in year 2006
- Lowest API is found around April to June
Time-Series Trend 4: Parallel Coord

• Apply polar system to select data of interest first to reduce clustering in parallel coordinates

• Weighted complete graph for axis ordering
  ▫ Dash density encodes correlation: solid line most correlated
  ▫ Oxygenic attributes more correlated
Time-Series Trend 4: Parallel Coord

- Time axis added; color also encodes time in year
- 2006 plot
  - Lines elegantly clustered together for most dimensions
  - Temperature varies dramatically
- 2004 plot
  - Unusual yellow lines (near the end of year) seen at high RSP and NO\textsubscript{2} values, resulting in the largest API in this set of data
Time-Series Trend 4: Parallel Coord

- Other dimensions reveal a rather constant pattern in all 3 years.
- Decreasing trend of $O_3$ observed in this sector, i.e. when strong winds are blowing from the north.
Contents

• Introduction
• Related Work
• System Overview
• Visualization Techniques
• Experimental Results
• Conclusion and Future Work
Conclusion

• Proposed a comprehensive system for weather data visualization

• Integrated:
  ▫ Polar system
  ▫ Parallel coordinates

• Developed:
  ▫ Circular pixel bars embedded in polar system
  ▫ Enhanced parallel coordinates with vector and time axes
  ▫ Weighted complete graph for parallel axes ordering

• Analyzed the air pollution problem in Hong Kong
  ▫ Known findings revealed effectively
  ▫ Unknown patterns detected by domain scientists
Future Work

• Incorporate new datasets into the existing system for further exploration
  ▫ Visibility, PM2.5, etc

• Allow data transformation
  ▫ Very often experts are only interested in the oxide content of the pollutant
  ▫ May compute the sum of oxygenate substances with different weight to seek any revealing patterns
Acknowledgments

• Collaborate work with:
  ▫ Professor Huamin QU
  ▫ Mr. Anbang XU
  ▫ Mr. Peter Kai-Lun CHUNG

• Institute for Environment, HKUST
  ▫ Professor Alexis LAU
  ▫ Dr. Zibin YUAN
The End

Thank You
Q & A

Polar system with embedded circular pixel bars

Weighted complete graph

Enhanced parallel coordinates with S-shape vector axis