Scaling Up Parallel Coordinate Plot with Color-coded Stacked Histograms

Jinwook Bok*
Seoul National University

Bohyoung Kim†
Hankook University of Foreign Studies

Jinwook Seo‡
Seoul National University

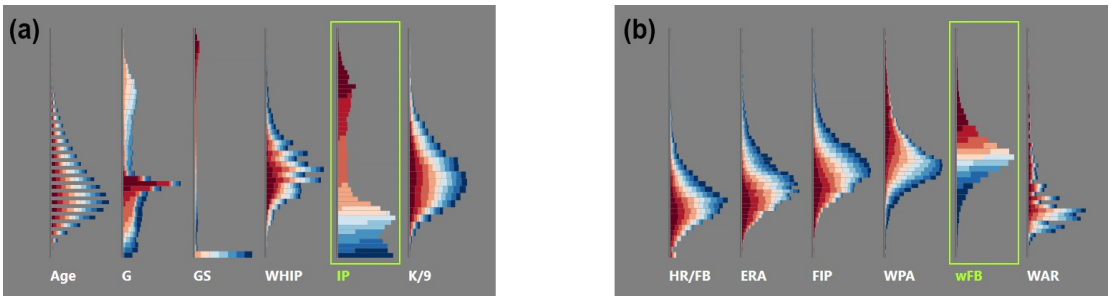

Figure 1: Histogram-Scented PCP (HSPCP) utilized on the baseball pitchers' data gathered from FanGraphs [4]. The attribute selected for color-coding is bordered by a green rectangle ((a): IP; (b): wFB).

Abstract

The original visual encoding of parallel coordinates plot (PCP) backfires as limitations when the number of items and/or attributes increases. Polylines for individual items clutter with each other and the linear ordering of vertical PCP axes makes it difficult to interpret relationship between physically distant attributes. In this paper, we introduce a novel technique that overcomes the innate limitations of PCP by attaching stacked-bar histograms with discrete color schemes to PCP. The color-coded histograms enable users to grasp an overview of the whole data without cluttering or scalability issues. Each rectangle in the histograms is color-coded according to the ranking of data by a user-selected attribute. The color-coding scheme allows users to perceptually examine relationships between attributes, even between the ones displayed far apart, without repositioning or reordering axes. We adopt the Visual Information Seeking Mantra so that the polylines of the original PCP can be used to show details of a small number of selected items when the cluttering problem subsides.

Index Terms: H.5.m [Information interfaces and presentation (e.g., HCI)]: Miscellaneous

1 Introduction

Parallel coordinates plot (PCP) [1] is a visualization technique that arranges multiple attributes parallel to each other on a 2D plane. Clusters of data items and linear relationships between adjacent attributes, including correlation, can be perceived by the line-crossing patterns of the lines on PCP. This pattern recognition becomes harder when lines overlap severely with each other as data becomes bigger in terms of the number of items and attributes. Furthermore, relationship between attributes is difficult, if not impossible, to infer from visual patterns in PCP when the axes are not adjacent.

* bok@hcil.snu.ac.kr
† bkim@hufs.ac.kr
‡ jseo@snu.ac.kr

In this paper, we introduce Histogram-Scented PCP (HSPCP), a novel visualization technique that deals with the innate limitations of PCP while preserving its perceptual advantages and characteristics. Following the Visual Information-Seeking Mantra [2], HSPCP augments the original PCP with color-coded stacked histograms that provide a scalable overview of the whole data without clutter. On each axis of PCP, we attach a stacked-bar histogram where each stacked bar is color-coded the ranking of a user-selected attribute. This histogram provides an overview of each attribute by showing the distribution of data items on the attribute. Visual comparison of the color distributions on histograms for multiple attributes reveals relationships between the attributes without suffering from cluttering or overlapping of lines in PCP. Especially, relationships between distant attributes that are hard, if not impossible, to grasp on the original PCP can be readily perceived on HSPCP through the visual comparison of color distributions for the attributes. The polylines are used on the later stages of the Visual Information Seeking process, when the cluttering problem is not severe after filtering or selection.

2 Constructing HSPCP

To construct the color-coded histograms of HSPCP, we first split data into equal-sized groups according to a user-selected attribute. Instead of using the original data value of the attribute to derive groups, we use ranking as the criterion to create groups. Grouping by ranking is more robust to outliers or skewed distributions that could skew the distribution of colors which will be applied to each group in the next step. Unequal-sized groups could occur because we make sure that elements with the same value are placed on the same group when splitting groups. This prevents the elements with the same value on the selected attribute from inconsistently assigned to a different group.

Second, we apply a unique color to each group of the split data using a discrete diverging color scheme. The discrete color scheme is designed to distinguish between different groups, and to help users perceive the difference (or similarity) between them. In this paper, we use a ten-level diverging red-blue color scheme acquired from ColorBrewer2 [3] (low to high rankings from blue to red). We adopt a diverging color scheme to emphasize low and high ranked ones with more saturated (thus more salient) colors because they are usually more valuable in data analysis.

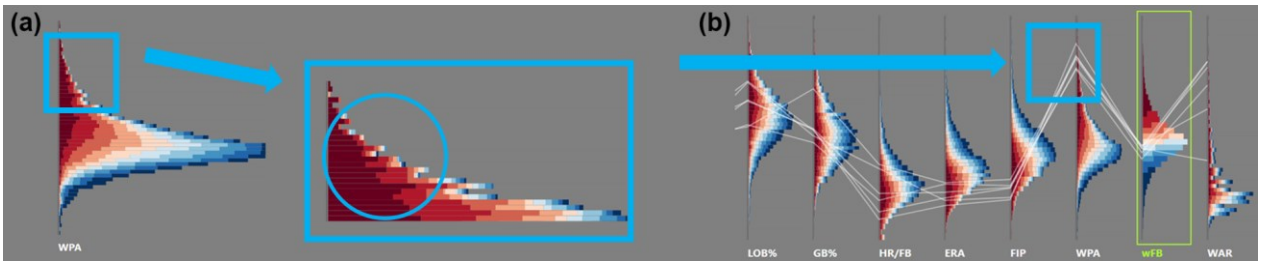

Figure 2: Example of the Visual Information Seeking Mantra [2] applied. (a) Some items of interest (blue item in the light blue box, implying outliers) are spotted from the overview (histogram). (b) By displaying the items as polylines, detail of the selected items can be spotted.

Finally, with the pre-processed data (grouped by the ranking and then color-mapped), we render a horizontal histogram that represents the distribution of data items on each attribute on the corresponding axis of PCP. The histogram is drawn as a stacked bar chart with the colors previously applied. When stacking the groups, the order of color-coded elements must agree with the order of the colors in the color scheme so that elements in the same color are grouped together, which helps users effectively perceive information from the distribution of colors. Then, the bar of the histogram looks like a stack of color-coded blocks. Such stacking of elements makes the layout similar to the Attribute Explorer [5], where the histogram consists of stacks of 'lightbulbs' which represent individual data items. In HSPCP, a single stacked element represents a group of data, and its length is proportional to the number of items belonging to the group in the corresponding attribute.

3 Visual Information Seeking with HSPCP

HSPCP utilizes color-coding derived from a single selected attribute. Figure 1 shows two different color-codings applied to the same data. In Figure 1(a) and 1(b), the attributes, IP and wFB are selected as the criterion for color-coding, respectively. On HSPCP, changing the attribute for color-coding is a pivotal step for revealing features in the data. Visually comparing distribution of colors on histograms can reveal relationships between the selected attribute and all the other attributes. Users can steer their data exploration by changing the attribute for color-coding to seek relationships between attributes from different aspects. This methodology can be effectively utilized in situations where only a part of the attributes on the dataset are familiar. Users can start their exploration by first setting a familiar attribute as the criterion for color-coding, then expand their knowledge about the unknown attributes by analyzing them from the perspective of the known.

Correlation between the attribute for color-coding and an attribute of interest can be readily estimated just by visually comparing the color distributions of the two histograms. In Figure 1(b) the wFB is selected as the criterion for color-coding with the color changing from blue (for low ranked values) to red (for high ranked values) from bottom to top, values low to high. The histogram for the WPA (4th from the left) attribute shows a color distribution similar to the one for the wFB attribute, implying a positive correlation between the two attributes. But the histogram for the FIP (3rd from the left) attribute shows an inverted color distribution compared with the selected attribute, implying a negative correlation between the two. To measure how this approach performs compared to the original, we conducted a controlled user study with 20 subjects to compare the original PCP and HSPCP in terms of correlation coefficient estimation. Results showed that the approach of inferring the correlation coefficient from the colored distribution outperformed the polylines, especially on positive correlation conditions.

The process of finding correlation between attributes can be done with multiple attributes regardless of the distance between them. In Figure 1(b) we can also visually grasp that wFB has a negative correlation with HR/FB (1st from left) and ERA (2nd from left). Unlike polylines that are only effective in revealing relationships between adjacent attributes, the indirect connection provided by colors in HSPCP is not cluttered and much less influenced by the distance between attributes. With such adoption of the color encoding, the axes ordering becomes much less important as the relationship between attributes can be easily perceived through indirect connection by matching color, even when they are distant from each other.

Users also can recognize clusters of items that follow a certain trend, or outliers that do not follow the trend from the histograms. Figure 1(a) shows a part of the baseball pitchers' dataset where the histograms are color-coded by the IP attribute. From the color distribution, it can be easily found that items with high IP are gathered in a tight region of the G attribute. Meanwhile, on Figure 2(a) some items in the light blue box have salient colors that are different from the overall distribution of its surroundings. These outliers have lower rankings of the selected attribute compared to most of the nearby items, which can be readily spotted by the distinct colors.

Following the Visual Information Seeking Mantra [2], HSPCP utilizes the polylines in harmony with the color-coded histograms. The colored histograms effectively display the overview of data and relationship between attributes; however, it is not efficient in showing the more accurate value of each data item. Meanwhile, polylines excel in showing the exact value of each data items but suffer from cluttering when there are many of them. Thus, we combine these two components to complimentarily support each other. At the beginning, color-coded histograms show the overview of the data. This overview is scalable as histograms don't suffer from overlapping issues as the polylines. After zooming and filtering out redundant data items in the histograms, the polylines show details of a small group of items selected from the histograms, enabling users to take advantage of the original PCP design. In this way, the cluttering problem is diminished because the polylines are only shown for a selected small portion of the data. Figure 2 shows an example of a scenario where this approach is in action. Upon finding items of interest, the items can be displayed as polylines to show the detail of the selected items with much less cluttering.

References

[1] A. Inselberg and B. Dimsdale, "Parallel coordinates: a tool for visualizing multi-dimensional geometry," *Proc. IEEE Conf. on Visualization*, pp. 361-378, 1990.

[2] B. Shneiderman, "The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations," *Proc. IEEE Symp. on Visual Languages*, pp. 336-343, 1996.

[3] ColorBrewer, http://colorbrewer2.org, June 2018.

[4] FanGraphs Baseball, https://www.fangraphs.com/, June 2018

[5] L. Tweedie, B. Spence, D. Williams, and R. Bhogal, "The Attribute Explorer", *Proc. ACM Conf. Human Factors in Computing Systems*, pp. 435-436, 1994.