DOI: 10.1111/cgf.12877
Eurographics Conference on Visualization (EuroVis) 2016
K.-L. Ma, G. Santucci, and J. van Wijk
(Guest Editors)

Volume 35 (2016), Number 3

Enhancing Scatterplots with Multi-Dimensional Focal Blur

J. Staib, S. Grottel and S. Gumhold†

TU Dresden, Germany

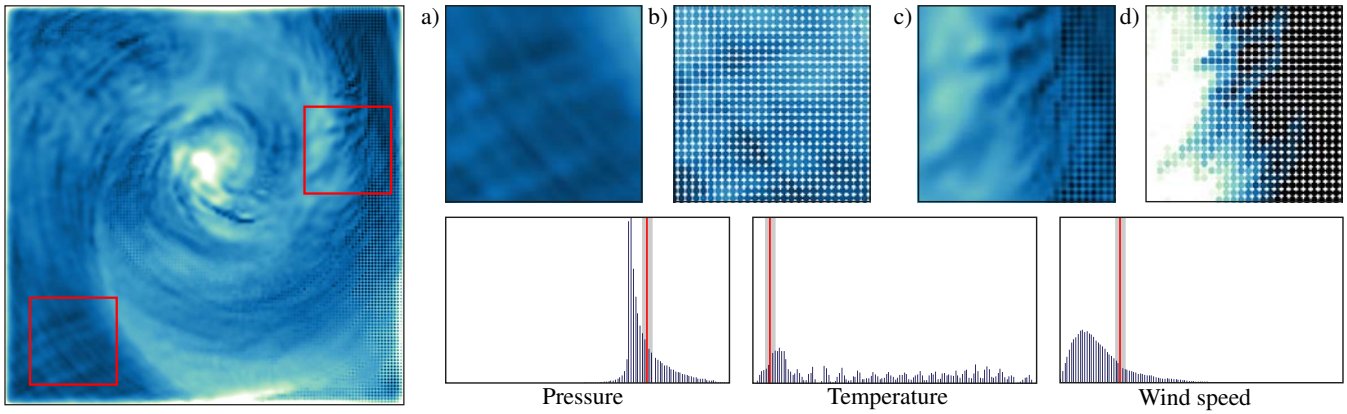


Figure 1: *Hurricane Isabel data set shown in different plots. Left: scatterplot of* x×y *using our blur method. Top row: zoomed-in views of the two marked regions. a) and c) use our method. b) and d) use transparent splats for comparison. Bottom row: histograms of three relevant dimensions and the selected focus coordinates marked with red vertical lines. (cf. Sec. 5.2)*

Abstract

Scatterplots directly depict two dimensions of multi-dimensional data points, discarding all other information. To visualize all data, these plots are extended to scatterplot matrices, which distribute the information of each data point over many plots. Problems arising from the resulting visual complexity are nowadays alleviated by concepts like filtering and focus and context. We present a method based on depth of field that contains both aspects and injects information from all dimensions into each scatterplot. Our approach is a natural generalization of the commonly known focus effects from optics. It is based on a multi-dimensional focus selection body. Points outside of this body are defocused depending on their distance. Our method allows for a continuous transition from data points in focus, over regions of blurry points providing contextual information, to visually filtered data. Our algorithm supports different focus selection bodies, blur kernels, and point shapes. We present an optimized GPU-based implementation for interactive exploration and show the usefulness of our approach on several data sets.

Categories and Subject Descriptors (according to ACM CCS): I.3.3 [Computer Graphics]: Picture/Image Generation—Point Rendering I.3.8 [Computer Graphics]: Applications—Visualization

1. Introduction

Scatterplots are among the fundamental techniques for data visualization. A single plot directly depicts data points in two dimensions and discards all other information. In order to visualize multi-dimensional data, scatterplots are extended to scatterplot matrices (SPLOM). Since a single data point is distributed over many plots, the visual complexity impedes a human's analysis. Further concepts are required to alleviate this issue. These include interactive methods, like selecting, linking and brushing, or filtering. An important concept is focus and context, allowing a user to focus on a relevant subset of the data but retaining context of surrounding elements. This context can be depicted less accurate, as it will only be used to help understanding the data in focus.

Our main contribution is the transfer of the concept of *depth of field* to the multi-dimensional case for scatterplots. The approach is based on a *focus selection body* and provides a natural understanding of focus and context using points in focus and blurry context.

† {joachim.staib,sebastian.grottel,stefan.gumhold}@tu-dresden.de

Our method allows to *depict multi-dimensional structures*, *assess qualitative properties of defocused data* and to *provide context with only minor distraction* in a continuous way.

In optics, depth of field can be described by the distance or region at which objects appear sharp. In the multi-dimensional case, we understand that region as a focus selection body. This is either a point, a parametric shape or a clustering-based selection of data points. We denote the distance from a point to that focus selection as *out-of-focus distance*. Similar to the distance in the 3D case being perpendicular to an image plane, we define the out-of-focus distance in a subspace of the data domain not containing the dimensions directly depicted in the individual scatterplots. This way, information from all other dimensions is injected into the visualization, revealing multi-dimensional structures and their characteristics.

We use blur as visual attribute to explicitly encode out-of-focus distance, similar to a shallow depth of field. It is present in almost any optical system with lenses, including, of course, the human eye. In fact, most parts of our vision are blurred, and the brain appropriately interprets this content, even before awareness. The effect is considered preattentive recognition and allows for coarse qualitative cues as has been shown in user studies [KMH02]. This makes depth of field a perfect candidate for focus and context applications. In classical scatterplots, individual data points only affect limited regions in the plot domain, i.e. usually single image points or image points inside small shapes, like discs, squares or crosses. Using blur allows for *extending the influence of the data points forming context* in the plot domain and are thus visible behind and between the sharp points. Using an extensive amount of blur visually removes data points from a visualization, implicitly implementing filtering. In this way, our method provides a consistent and continuous transition from data points in focus, over regions of blurry points providing qualitative contextual information, to filtered data points.

The remainder of this work is structured as follows. In Sec. 2, we review existing works on extensions of scatterplots with a focus on a large number of multi-dimensional point data and uses for depth of field in visualization. In Sec. 3, we lay out the mathematical fundamentals of distance calculation for various selection bodies, filtering kernels and glyphs that we use. We briefly describe our implementation in Sec. 4. In Sec. 5, we evaluate our approach exemplary concerning effectiveness and usefulness on several data sets and discuss the benefits and drawbacks in Sec. 6. We conclude our work in Sec. 7, followed by ideas for future work.

2. Related Work

Already in the late eighties Tufte [Tuf86] noted that scatterplots had been the most frequently used visual tool to analyze the relationship between variables. Their first use has been historically analyzed in [FD05]. Being actively researched for decades, a variety of sophisticated concepts have been developed to handle large data sets and multi-dimensional cases. The most important concepts rely on abstraction or distortion.

2.1. Frequency Based Techniques

In order to overcome overplotting issues and facilitate data understanding, more abstract visualizations of data density are common. This metaphor is denoted as frequency based, since plots do not show individual items but the frequency of their occurrence per plot unit [LH11]. Being one of the most basic approaches, Wilkinson proposes to use transparency and additive blending [Wil05]. Eilers and Goeman rasterize points into discrete bins and blur the result [EG04].

Other density based approaches rely more directly on the concept of Kernel Density Estimation (KDE) [Sil86]. For each position on the plot area, a density is estimated based on a distribution kernel, e.g. a Gaussian. Lampe and Hauser propose a fast renderer for streaming using a point kernel and a line kernel [LH11]. Zinsmaier et al. use a KDE approach to aggregate nodes for a GPU based visualization of large graphs [ZBDS12]. Jourdan et al. blur a rasterization of the projected points using a Gaussian kernel and perform a segmentation of homogeneous areas to identify data sub sets of similar properties [JPKM07]. Mayorga and Gleicher extract contours of high point densities and fill the enclosed area with a constant color, depending on the point group [MG13]. In order to identify outliers, points outside these groups are sampled and rendered individually. In our method, we also employ kernels for blurring that can be interpreted as a density distribution. Most of the mentioned methods generate the density distribution in screen space by blending kernel splats. We take a similar approach, except that splats are parametrized via their distance to a multi-dimensional selection body.

Bachthaler and Weiskopf treat the discrete data points as samples from a continuous function. They visualize a reconstruction of this function using respective interpolation schemes [BW08]. They also optimized their approach for fast rendering [BW09].

2.2. Distortion Based Techniques

To make data points distinguishable, several works propose to distort the projected points. Keim et al. present a density-equalizing distortion [KHD*09]. Points are moved to the next free space in the vicinity of its original position. Janetzko extends the work to account for the local distribution of the data [JHM*13], also providing a comprehensive review on point distortion techniques. Instead of only selecting two dimensions for a scatterplot, multi-dimensional projection techniques reduce the dimensionality using more advanced schemes. The idea is to conserve characteristics of multi-dimensional data either globally or locally [JPC*11]. The tool presented in [JNB*12] allows to steer a local projection and selection of points by back projecting low-dimensional points into the original high-dimensional space.

2.3. Multi-dimensional Visualization

To visualize more than two dimensions, SPLOMs are commonly used. The emerging problem of high visual complexity is typically addressed by selection and filtering. Brushing and linking is one common approach [Kei02]. The technique was already proposed in the late eighties by Becker and Cleveland. The main idea is to

select an area of interest in one plot [BC87]. All points are simultaneously highlighted in linked plots. Turkay et al. propose to visualize and brush statistical aggregates on selected dimensions as well [TFH11]. Eisemann et al. present a visualization of hierarchy of scatter plots [EAM14]. Inside one panel, a subset of data points is shown for other axes as nested plot. Van Long describes an interaction scheme based on hierarchical clustering [Lon13]. The user interactively selects data subsets from a tree-like structure, which are then shown in a scatterplot matrix. Piringer et al. present linked visualizations of 2D and 3D scatterplots [PKH04]. They vary point size and color in 3D plots to enhance depth perception.

For filtering of multi-dimensional data, Furnas and Buja describe and analyze the concept of *Prosection*, which is a combination of projection and section [FB94]. The authors propose to project points on two dimensions and only consider data in an interval of a third dimension. Spanning the full range with this interval creates traditional scatter plots. Tweedie and Spence generalize this technique to Prosection Matrices for multidimensional data [Spe14]. In each dimension an interval can be adjusted. Points are filtered in all plots. In the limit for intervals of infinitesimal width, the result is identical to a section of a 2D plane. This is similar to the concept of HyperSlices, presented by van Wijk and van Liere, which was designed for visual analysis of continuous functions [vWvL93]. The authors propose a visualization of the cross sections and an interaction scheme based on a set of orthogonal 2D planes that can be freely rotated. All planes intersect in a user definable point.

Similarly we can define focus intervals for each dimension. In our work we extend the concept of a focus point to a multi-dimensional body.

2.4. Depth of field in Visualization

Depth of field using blur has been an active subject of research. Human perception of blur is considered preattentive [KMH02]. This means, that the brain processes this information very fast before awareness [HE12]. User studies have shown that depth of field can help depth perception, enhances basic ordering [MS02] and even allows for the estimation of absolute sizes and distances [HCOB10]. Schrammel et al. found that blurred and sharp objects can be discriminated reliably [SGT*03]. Kosara et al. [KMH01] most notably generalized this concept as Semantic Depth of Field where the strength of blurriness is controlled by a relevance function. In various user studies, they concluded that blur is not suitable to directly encode information, but can be combined with other visual variables without imposing significant stress on the interpretation [KMH02]. In the same user studies, the authors furthermore found that users can discriminate between two to four object groups.

The technique was applied to a number of applications, e.g. GIS systems and text highlighting. Kosara et al. also experimented with scatter plots [KMH02]. They propose to map one dimension of the data to the amount of blur. We extend the case for scatter plots and use it as a means to represent the out-of-focus distance.

3. Multi-Dimensional Focal Model

We consider multi-dimensional data sets as an $n \times m$ dimensional matrix $\boldsymbol{P}$. Each row $\underline{\boldsymbol{p}}_i$ is one of n data points living in a m-dimensional feature space $\mathbb{F}^m$. Data points are indexed with i and dimensions with j. Each dimension $\mathbb{F}_j$ can have a different scale of measurement. For the sake of simplicity, we limit the following descriptions to metric scales and nominal features, being the two extreme cases.

We denote the blurring kernel by $\Theta(r_\Theta)$. The amount of blur is controlled by an effective blur kernel radius $r_\Theta(\underline{\boldsymbol{p}}_i)$, individually for each data point i. This blurriness of a data point is related to the circle of confusion in optics. We derive $r_\Theta(\underline{\boldsymbol{p}}_i)$ from a multi-dimensional out-of-focus distance:

$$r_\Theta(\underline{\boldsymbol{p}}_i) = \sigma \cdot d^F_{j_1,j_2}\left(\underline{\boldsymbol{p}}_i\right) , \tag{1}$$

with F being the focus selection body. The parameter σ is a global factor to intuitively scale the overall blur effect.

This distance is defined for a scatterplot over the dimensions j_1 and j_2. Thus, these two dimensions are omitted in the distance calculation. An interpretation from information theory strengthens this approach, as information from the dimensions j_1 and j_2 is already visualized, and would be represented twice.

3.1. Focus Selection Bodies

Focus selection bodies are a definition of what is in focus. $d^F_{j_1,j_2}(\underline{\boldsymbol{p}}_i)$ is the distance of $\underline{\boldsymbol{p}}_i$ to this selection, i.e. it is zero for selected points and increases with growing distance. The actual definition of d^F depends on the scale of measurement of the different dimensions. For metric scales, this distance is defined as absolute difference between values. For nominal scales, the distance is based on class equality.

We distinguish between the geometric selection bodies, hybercube and hypersphere, and cluster-based selections. Geometric bodies select a subvolume of $\mathbb{F}^m$. The cluster-based selection body is used to select nearly connected structures (cf. Fig. 2), like the hidden word in the well-known "pollen" data set.

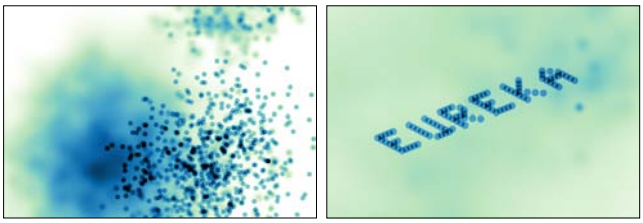
Focus Hypercube

The focus hypercube is the simplest selection body. In each dimension an interval (of size $\pm s_j$) from the respective coordinate of the focus point $\underline{\boldsymbol{f}}$ is defined to be in focus. For metric scales the distance in $\mathbb{F}_j$ reads:

$$d^{HC}_j\left(\underline{p}_{i,j}\right) = \left(|\underline{p}_{i,j} - \underline{f}_j| - s_j\right)\Big|_+ . \tag{2}$$

Negative values are clamped to zero, as indicated by $|_+$. For nominal scales, $\underline{f}_j$ is the set of classes of points in focus and the distance reads:

$$d^{HC}_j\left(\underline{p}_{i,j}\right) = \begin{cases} 0 & \text{if } \underline{p}_{i,j} \in \underline{f}_j \\ 1 & \text{else .} \end{cases} \tag{3}$$



hypersphere selection cluster-based selection

Figure 2: *Applied selection bodies for synthetic data (left) and the "pollen" dataset (right).*

These individual 1D distances are then combined to a distance in all $m-2$ dimensions:

$$d^{HC}_{j_1,j_2}\left(\underline{\boldsymbol{p}}_i\right) = \sqrt{\sum_{j\neq j_1,j_2}^{m}\left(d^{HC}_j\left(\underline{p}_{i,j}\right)\cdot\lambda_j\right)^2}\,. \tag{4}$$

Distances d^{HC}_j need to be comparable to each other for different j. This can be achieved by normalization. To compensate for scaling problems and to give additional modeling freedom, we introduce weights λ_j. These adjust the influence of each dimension in the summary distance. However, as these weights are only used to model exceptions from the normalization rule, $\lambda_j = 1$ is the common case. One interesting use case is $\lambda_j = 0$, which removes dimensions from the distance altogether.

Focus Hypersphere

We also provide a selection based on a hypersphere around the focus point $\underline{\boldsymbol{f}}$, similar to the hypercube. The distance reads:

$$d^{HS}_{j_1,j_2}\left(\underline{\boldsymbol{p}}_i\right) = \left(\sqrt{\sum_{j\neq j_1,j_2}^{m}\left(d^{HS}_j\left(\underline{p}_{i,j}\right)\cdot\lambda_j\right)^2} - s\right)\Bigg|_+ \tag{5}$$

with

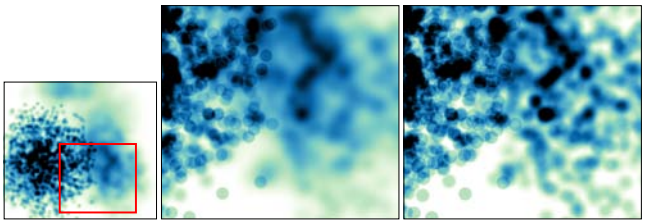
$$d^{HS}_j\left(\underline{p}_{i,j}\right) = \left(\underline{p}_{i,j} - \underline{f}_j\right) \tag{6}$$

being the distance in dimension j for ratio scales and s being the selection radius of the hypersphere. λ_j can be used to form a hyperellipsoid. Usually $\lambda_j = 1$ is sufficient and the global s should be used for intuitive interaction. For nominal scales Eq. (3) can be directly used in Eq. (5) as $d^{HS}_j = d^{HC}_j$.

Focus Cluster

Both hypercube and hypersphere might be too inflexible for some data, due to their simple shapes. We want to be able to define *an object of the scene* to be in focus. For this object in focus we define the cluster $\boldsymbol{C}$ as follows. All pairs of points with a distance smaller than a maximal neighborhood distance s are considered connected. Cluster $\boldsymbol{C}$ is the connected component starting at the focus point $\underline{\boldsymbol{f}}$. The distance between two points $\underline{\boldsymbol{p}}_{i1}$ and $\underline{\boldsymbol{p}}_{i2}$ is computed similarly to Eq. (5) by:

$$d^{Cp}\left(\underline{\boldsymbol{p}}_{i1},\underline{\boldsymbol{p}}_{i2}\right) = \sqrt{\sum^{m}\left(d^{Cp}_j\left(\underline{p}_{i1,j},\underline{p}_{i2,j}\right)\cdot\lambda_j\right)^2}\,. \tag{7}$$



a) b) c)

Figure 3: *Blurring with different kernels. a) whole data set; b) zoom-in with Gaussian blur; c) zoom-in with disc blur*

Note, that in Eq. (7) all m dimensions are included, since this focus selection body should be independent from the final scatterplot.

The out-of-focus distance d^C is the distance between the query point $\underline{\boldsymbol{p}}_i$ and closest point $\underline{\boldsymbol{c}}_{k(i)}$ in $\boldsymbol{C}$:

$$\underline{\boldsymbol{c}}_{k(i)} = \underset{\underline{\boldsymbol{c}}_k\in\boldsymbol{C}}{\arg\min}\; d^{Cp}\left(\underline{\boldsymbol{p}}_i,\underline{\boldsymbol{c}}_k\right)\,, \tag{8}$$

and reads

$$d^{C}_{j_1,j_2}\left(\underline{\boldsymbol{p}}_i\right) = \left(\sqrt{\sum_{j\neq j_1,j_2}^{m}\left(d^{C}_j\left(\underline{p}_{i,j}\right)\cdot\lambda_j\right)^2} - s\right)\Bigg|_+ \tag{9}$$

with

$$d^{C}_j\left(\underline{p}_{i,j}\right) = \left(\underline{p}_{i,j} - \underline{c}_{k(i),j}\right)\,. \tag{10}$$

For nominal scales Eq. (3) is adjusted accordingly.

3.2. Glyph Blur Kernels

Every data point is visualized using one of several glyphs G_r, where r is a user definable size. For every position $\underline{\boldsymbol{g}}$ in the plot, the contributing v from each data point $\underline{\boldsymbol{p}}_i$ is obtained by convolution of the point's glyph with a blur kernel $\Theta\left(r_\Theta\left(\underline{\boldsymbol{p}}_i\right),\underline{\boldsymbol{g}}\right)$ in X and Y direction as:

$$v = \int \Theta\left(r_\Theta(\underline{\boldsymbol{p}}_i),\underline{\boldsymbol{g}}+\underline{\boldsymbol{x}}\right)\cdot G_r\left(\underline{\boldsymbol{p}}_i-\underline{\boldsymbol{x}}\right)\mathrm{d}\underline{\boldsymbol{x}}\,. \tag{11}$$

Our blurring kernels represent distributions, similar to density estimation kernels. Any kernel with an integral of 1 fulfills the requirements of being *mass preserving*, i.e. not changing the influence of the data points on the plot, but only distributing their influence over larger areas. Typically a Gaussian kernel is used to smoothly fade out the glyphs. They are identical to bivariate normal distributions with the mean at one of the data points $\underline{\boldsymbol{p}}_i$ and the standard deviation $r_\Theta(\underline{\boldsymbol{p}}_i)$. Alternatively, we propose a disc-shaped kernel for sharper boundaries of the defocused points. This can reveal additional information on the context's structure, but reemphasizes defocused regions. This kernel contains all points that do not exceed a distance of $r_\Theta(\underline{\boldsymbol{p}}_i)$ from its center. Fig. 3 shows an example of both kernels.

By default, the glyphs are discs. For nominal dimensions our methods allows for different glyph shapes to be used. But, with increasing blur the shapes become increasingly indistinguishable. Alternatively, using additional visual attributes, e.g. color, should be considered (cf Sec. 5.1).

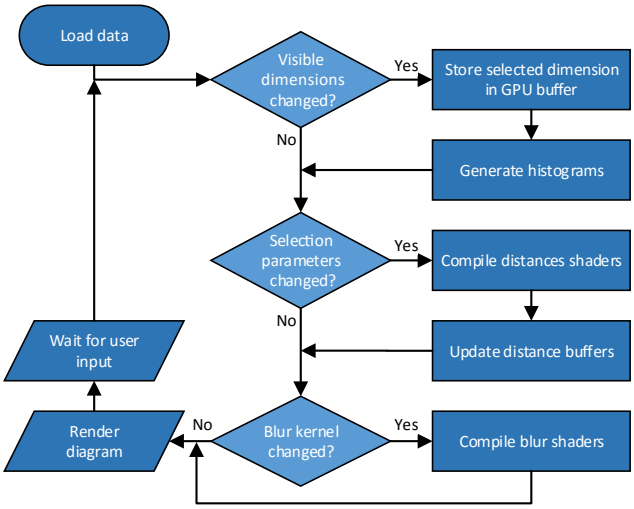


Figure 4: *Outline of the control flow during rendering*

4. Algorithmic Overview

We generate the rendering for each panel by splatting point sprites per projected data point, which evaluate the glyph shape convolved with the blur kernel.

4.1. Implementation

We implemented our visualization for the most part on the GPU using OpenGL to allow for interactive, real-time visual analysis and exploration of the data. The basic control flow during rendering is outlined in Fig. 4. The complete data is initially transferred to graphics memory as array of points. Also m atomic integer buffers for histograms are generated. They are directly created in OpenGL compute shaders. For multi-dimensional distances, $m \cdot (m-1)/2$ floating point buffers of size n are set up. These buffers are updated only if required, that is if parameters or the type of the focus selection body change.

During rendering, we generate a quadratic primitive covering the blurred glyph for each data point in each plot. Our implementation supports different glyphs as well as different blur kernels. For the general case we precompute the blurred glyph images and use a 3D lookup texture during rendering. Since disc glyphs are suitable in most cases, we provide optimized implementations for them. In combination with Gaussian blur we use a 2D lookup texture. Each row represents a rotational symmetric cross section of the blurred disc. In combination with a disc blur kernel the convolution result is calculated analytically in the shader. For each point on the splat, the intersection area of the kernel shape, centered on that point and the disc at the splat's origin is calculated. All point's blurred splats are blended additively into a high precision buffer, followed by tone mapping. Finally, the histograms and user controls are rendered.

4.2. User Interaction

Our prototype presents the data set as SPLOM, more precisely the upper triangle matrix of pairwise combinations. The user can interactively add or remove dimensions from the visualization. Histograms are included in the diagonal elements of the SPLOM, and serve as interaction elements for defining the multi-dimensional focal point for the focus selection bodies. The value in each dimension can be adjusted by dragging a vertical line in the respective histogram. Influence ranges, e.g. for the hypercube, are shown as gray bars centered at that position and can be interactively adjusted with a slider. Every change of parameters triggers an immediate update of the visualization.

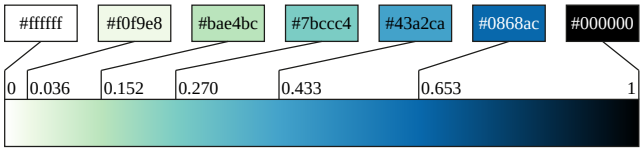


Figure 5: *The color map used in our visualizations (cf. ColorBrewer 5-class GnBu)*

5. Results

We first compare blurring to color and scale, and demonstrate the effectiveness and benefits of our method. Then, we exemplarily present two case studies showing the usefulness. Finally, we provide run-time performance figures and conclude, that our implementation allows for interactive exploration even with larger data sets.

We use the color map shown in Fig. 5 for tone mapping for all visualizations in this paper, if not stated otherwise. Based on a sequential multi-hue color map from ColorBrewer, designed for ordinal data, we adjust the interpolation positions of the key colors to obtain a linear luminance slope, suitable for visualization of metric scale data.

5.1. Comparison to Color and Scale

In our method, we map a multi-dimensional distance to the amount of *blur*. For comparison we map this distance to *color*, and to *color and scale* of opaque disc glyphs. Since opaque rendering is sensitive to the drawing order, we draw the glyphs based on the distance, far to near. This improves the visualization massively as shown in Fig. 7. Note, however, that the distance itself is unsigned. Ordering the points can thus lead to a wrong interpretation: points with similar colors could be assumed to be close to each other, but in fact lie on opposed sides with the same distance to the selection body. For comparison we prepared two experiments and two data sets: a *separation task* and a *shape perception task*. For each experiment we set the parameters to be as good as possible for the respective visualization, including color tables for the comparison methods.

Separation Task

For this task we use a three-dimensional synthetic data set, consisting of four point groups drawn from normal distributions (cf. Fig. 6). We further added a noise layer based on perlin noise. This noise exhibits macro structures and thus simulates structured, but uninteresting data. All five structures can be separated in only

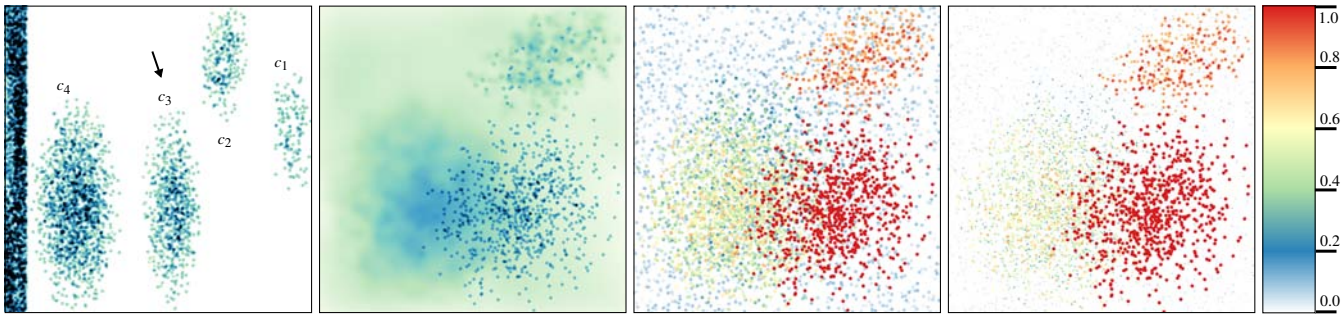


Figure 6: *Separation task; a) the data set in $x \times z$; for $x \times y$: b) results of blur, c) results of color and depth sorting, d) results with added varying scale, e) used color table (cf. ColorBrewer diverging 5-class spectral)*

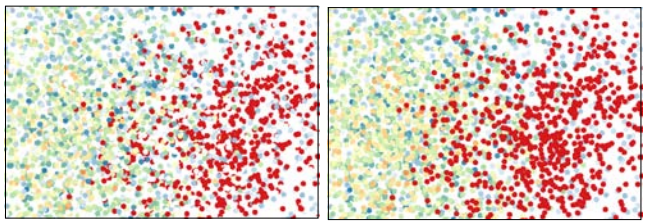


Figure 7: *Effects of ordered rendering for opaque glyphs; the color map shown in Fig. 6 is used.*

one dimension. The task is to emphasize group c_3 of the point groups visually. We select c_3 using a spherical selection body. The Figs. 6 b) to d) show the results using blur, colors and colors with scale. Fig. 6 e) shows the color table used in c) and d). It contains well-distinguishable colors to leverage perception of the distance. All methods are able to separate the selected cluster.

Color allows for identifying all four groups, but shows clutter from the background noise. Additionally using *scale* decreases clutter, at the cost of less distinguishability between the unselected groups. With our method, the selected group is salient, while the rest of the data is visible as context and does not visually distract. The contextual point groups are still qualitatively assessable, e.g. in terms of extents or thickness. The latter is visible because of additive blending and is not available using color and scale. For example, the blurred cluster c_4 (lower left area) appears thicker in Fig. 6 b) than the selected cluster. Thus its extent in the non-visible dimensions can be assumed to be larger.

Shape Perception Task

A crucial aspect of data analysis is understanding shape in more than two dimensions. For this experiment, we use a three-dimensional synthetic data set containing uniformly distributed point samples inside a cylindrical shape (cf. Fig. 8 a)). The domain is, again, polluted with point samples based on perlin noise. The goal of the task is to grasp the shape of the cylinder from only one plot. For the comparison visualizations we use the color table depicted in Fig. 8 e).

The Figs. 8 b) to d) show the results. As can be seen, using *color* leads to high frequencies in the rendering. While it is possible to see the inner element of higher density, the projected shape is barely visible. Scale does not help for this task in this setting, due to the continuous nature of the surrounding noise. Through blur, an effect similar to a projected volumetric reconstruction of the cylinder is achieved which visually separates the cylinder from the perlin noise. The orientation from bottom left to top right is clearly visible. The smooth changes in blur suggest the continuous smooth shape of the cylinder as well.

5.2. Case Studies

Case Study 1: Hurricane Isabel

We use a down-sampled version of the hurricane Isabel data set provided for the IEEE SciVis Contest 2004. The original data is a multi-attribute volume with a resolution of $500 \times 500 \times 100$ data points and a total of 13 variables describing wind speed, pressure, temperature and meteorological values, like cloud moisture mixing ratio. We reduced the spatial resolution to $1/4$ in each dimension and we stored the position as additional attribute at each data point. We finally added the magnitude of the wind speed vector, resulting in a total of 17 variables.

Fig. 1 shows the resulting scatterplot of the spatial domain $x \times y$. Note that the visible structure of the hurricane in the plot emerges through blur only, as the data points themselves are positioned on a regular grid. Our method increases the area of contribution for blurred points. Points close to the selected focus body, are thus visible in the space between the data points. The stronger the blur values, i.e. the larger d^F, the less impact data points have on individual image points. Dark areas, e.g. the upper right corner of the plot, only contain data points with very small distances. Bright areas, e.g. at the bottom, contain data points with large distances. Points in focus are directly visible by their sharp disk glyphs, as can be seen on the right side and in c) of Fig. 1. This not only clearly shows the vortex structure, but also shows *hatched* structures like in the upper left and lower left areas of the plot. These hint to wind directions changing with the height coordinate z within the vicinity of the selected focus point values.

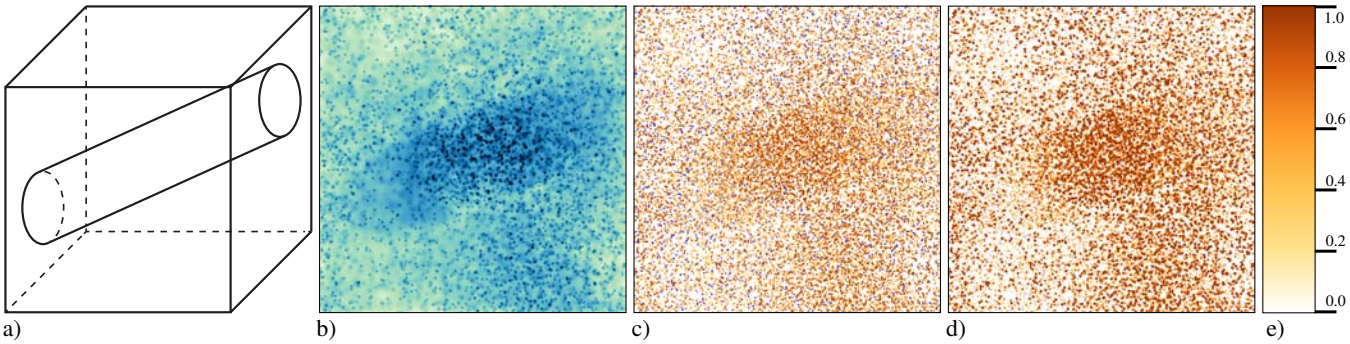


Figure 8: *Shape perception task. a) sketch of the data set, b) results of blur, c) results of color and depth sorting, d) results with added varying scale, e) used color table (cf. ColorBrewer sequential 5-class YlOrBr)*

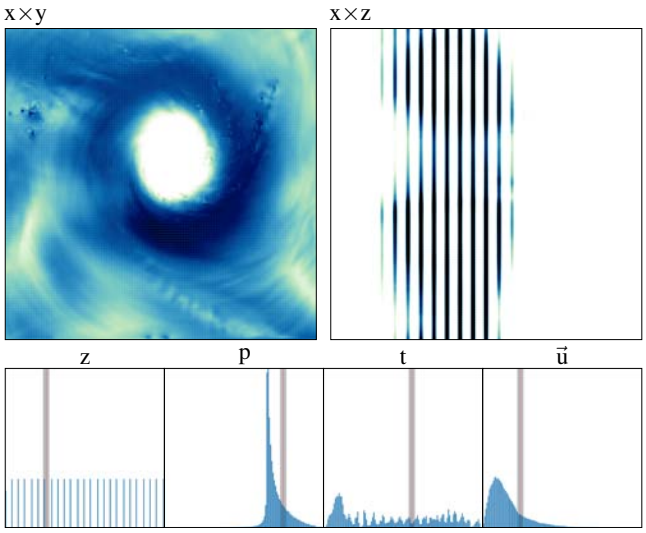


Figure 9: *Scatterplots of data set "Hurricane Isabel": Top row: two scatterplots of the spatial dimensions. The focus selection is on the hurricane's structure for increased pressure, mid-high temperatures and increased wind speed at lower height. Bottom row: histograms of the involved dimensions showing the focal point.*

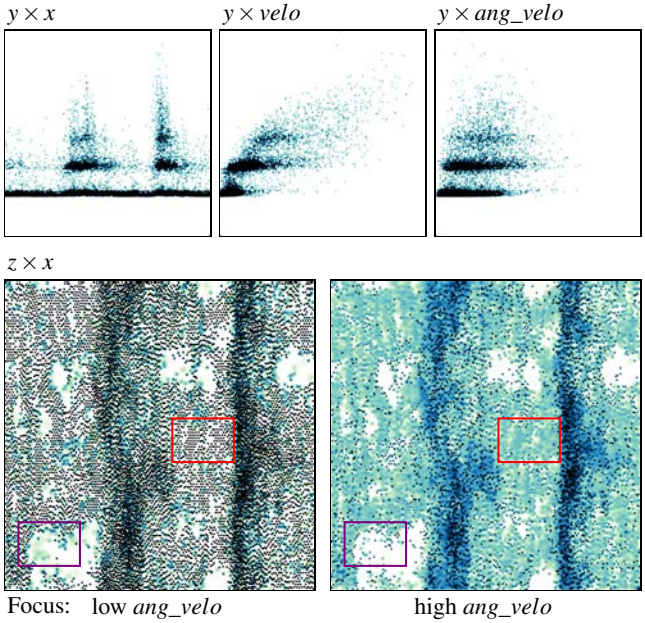


Figure 10: *Scatterplots of data set "Particles in Fluid simulation"; Upper row: classical scatterplots for dimensions* y *by either* x, velo *(linear velocity), and* ang_velo *(angular velocity). Correlation between angular velocity and height is not clear. Lower row: in scatterplots* x×z *particles with low angular velocity appear in the bulk material (e.g. red marked areas), while particles with high values appear at borders of holes (e.g. purple marked areas).*

The histograms in the lower row show the data of three relevant dimensions used in the distance computation: pressure, temperature and wind speed. The red and gray markings are $\underline{\boldsymbol{f}}$ and s of the hypercube focus selection.

In Fig. 9 the focal point is changed by first increasing the temperature. The x×z plot on the right side hints that data points with lower z values have smaller distances. Data points with too small or too large z values have large distances, are thus heavily blurred, and do not visually contribute to the plot. Choosing a correspondingly low z value for the focus point (cf. lower left histogram) reveals the structure of the hurricane again. Note, that the cross hatched pattern in the lower left area of the plot almost vanished compared to Fig. 1 a). Thus, the flow seems to be much more stable at this selected focus point and its vicinity.

Case Study 2: Particles in Fluid Simulation

This data set originates from a simulation of particle transport in a fluid (i.e. sand or small stones in a river). The particles store a total of 21 attributes, including position, linear speed, rotational speed and further simulation related values, like collision forces. Analyzing these values is not trivial but highly relevant for the application domain researchers. Our collaboration partners focus on the linear velocities (*velo*) and angualr velocities (*ang_velo*), as these values describe the dynamic and give insight into the interaction between particles and the surrounding fluid.

Table 1: *Data sets used for evaluation*

Name	# Dim.	# Points	Description
D_{pol}	5	3,848	"pollen" data set
D_{flw}	21	13,500	Solid particles in fluid simulation
D_{syn}	3	20,246	Synthetic (perlin noise with cylinder)
D_{hur}	17	315,500	Hurricane Isabel (NCAR)

Figure 10 shows selected plots and results found with our method. The direction of flow, and thus direction of primary particle motion is x, to the right in the plots. y is the elevation over the river's ground, upwards in the plots. Initially, the particles are all located at the bottom. The fluid's motion accelerates the particles and lifts some up. Because of the finite simulation extent distinct waves of particles emerge, which are clearly visible, e.g. in plot $y \times x$. As particles at higher elevation can move more freely with the fluid, not hindered by surrounding particles, there is a clear positive correlation between *velo* and y. Surprisingly, this is not true for *ang_velo*.

To study this aspect, the lower plots in Fig. 10 show the spatial dimensions $z \times x$. Glyphs are blurred based on different values for *ang_velo*. Particles in bulk (e.g. red marked area) are in focus for low values, as expected, as their packing hinders free motion. Particles with high *ang_velo* seem to be scattered everywhere, especially independent of elevation. This is understandable for higher elevations, where particles can move freely. At lower elevation, in-focus glyphs are located at borders of larger holes (cf. purple marked area) in the bulk regions. This indicates that rotation shown by high *ang_velo* values are induced by being pushed by other particles. The close packing in bulk regions inhibits rolling and, instead, particles slide as groups. Our methods can focus on particles from either of these two fundamentally different states of particle motion. It also allows for smooth transition in between these states and always retains context of the remaining, blurred particles.

5.3. Performance Evaluation

Our test system was equipped with an Intel i7-2770K @ 3.50GHz, with 16GB RAM and an Nvidia GeForce GTX 680 GPU with 2GB graphics memory. The rendering window was set to a resolution of 1920×1080 pixels. All data sets were zoomed in all setups to be completely visible, filling up most of the screen.

We use four data sets, as shown in Tab. 1. Data set D_{pol} is the "pollen" data set used in Fig. 2. Data set D_{flw} and D_{hur} are used in the case studies in Sec. 5.2. D_{syn} is the data set that is used for the shape perception task in Sec. 5.1. All parameters that influence the rendering speed, like blurriness and disk radius, were set to be the same for each data set.

Tab. 2 shows the overall rendering performance of our implementation. The values in fps are given for single scatterplots showing all data points for the data sets. Compared to classical scatterplots, our method (second column) is slower, due to more complex shader computations and texture look-ups, but still maintains interactive frame rates. The rendering load per point (third column) is similar for the larger data sets D_{syn} and D_{hur}. The higher values for D_{pol} result from the fixed overhead of the graphics pipeline, which outweighs the comparable small amount of points.

Table 2: *Overall rendering speeds for classical scatterplots (*classic SP*) and our method (*our SP*).* ns/pt. *is the average time required per data point for one plot in our method.*

Data set	classical SP	our SP	ns/pt.
D_{pol}	2411 fps	1106 fps	0.234 ns
D_{flw}	1966 fps	639.8 fps	0.116 ns
D_{syn}	1549 fps	553.2,fps	0.089 ns
D_{hur}	451.7 fps	33.78 fps	0.094 ns

Table 3: *Rendering performace scalability based on number of visible dimensions, i.e. number of visible plots (data set D_{hur} with 315,500 points)*

# dim	2	3	4	5	6	7	8
# plots	1	3	6	10	15	21	28
fps	33.78	34.42	19.59	18.49	13.23	16.62	12.51
ms/plot	29.60	9.684	8.508	5.408	5.039	2.865	2.855

Our implementation is designed to render a scatterplot matrix composed of many plots. Tab. 3 shows the performance scaling of our approach. Naturally, the number of plots increases quadratic with the number of visible dimensions. But since the window size and the zoom level remain constant, the sizes of the individual plots within the matrix decrease. The lower row shows the average milliseconds required to render one plot. The rendering time for the 1-plot case is highest, since it covers nearly the whole screen. The performance is mainly influenced by texture lookups and the blending operations. With decreasing plot sizes, the blending becomes the defining performance factor.

6. Discussion

Presenting multi-dimensional data with many data points in scatterplots can result in several problems, depending on the size and complexity of the data. Individual plots can become cluttered and hard to read without further aids such as focus and context or selection and filtering. Selecting points through a multi-dimensional body and incorporating blur to defocus unselected points directly combines selection, filtering, and focus and context in one consistent way. This allows for high numbers of points to be shown in plots, where only important points are focused and others provide context.

Depth of field is an easy to understand concept. We believe that providing context by blurring is superior to other methods such as using transparency (cf. Fig. 1), color or scale. Blending of colored glyphs, especially for multi-dimensional data points, has several problems considering the interpretability of the final color, the influence of one data point or the blending order.

The comparative studies show that blur visually emphasizes important structures while still allowing to assess qualitative properties of unfocused data. In case study 1, the $x \times y$ plot delivered hardly any information but only a regular grid in traditional scatterplots. Using our method, emerging visual features help steer the exploration. Also, points that would be normally completely overplotted become visible as the plotting area gets larger with increasing distance. This is done without the need of point distortion or complex multi-dimensional projection schemes. As shown in case study 2, gradually deemphasizing unselected data greatly helps in

finding characteristics of data from only a single plot. Of course, for more in-depth analysis all plots of a SPLOM are still required, in which case the classical scatterplots also contain all information.

Various selection strategies are suitable for different data configurations. The convex hyper bodies are best suited for exploration tasks, and are easily parameterizable by the user. Our clustering strategy helps when groups of points are separable based on distance or density.

Our implementation utilizes programmable GPUs and is thus fast enough to allow for real time exploration. In cases of many points covering the same pixel, fill rate problems occur and rendering speed can drop considerably. In practice, however, such a situation is not very common. Still, the performance depends on a trade-off between the number of points and the visible dimensions.

Our method shares all benefits and problems with SPLOMs, e.g. the quadratic growth of necessary panels with increasing dimensionality. However, our method can be combined with methods addressing these issues, like dimension reduction or rearrangement and rescaling of panels.

For distance calculation we use the L_2 norm. Per plot, this corresponds to the length of the vector composed of all but two dimensions. For nominal data, we decide its influence based on whether the point's class is part of the selected set. Even if all dimensions represent metric values, the issue of individual scaling and comparing dimensions in the single metric remains. Variables with much higher scale than others have a higher impact on the distance calculation. Renormalization of the value ranges of the different dimensions would solve this issue, but might not be plausible in terms of the meaning for the data domain. Individual scaling factors and even non-linear mappings might be required in future.

7. Conclusion and Future Work

We presented an extension of scatterplots to emphasize a local multi-dimensional area of interest, defined by a selection body. Based on a distance measure, projected data points are blurred to provide context and to convey information from other dimensions. Our model builds upon the intuitive effect of shallow depth-of-field, that has been shown to be beneficial for focus and context applications. We presented several focus selection bodies and blur kernels. We showed the usefulness of our method in comparative studies and two case studies. The studies indicate, that more correlations and trends can be inferred faster from individual plots than would be possible with classical scatterplots.

For future work, we plan to elaborate and refine work flows that exploit all advantages of this method. This includes improvements in the user interface, especially integration of methods to place high dimensional points. Furthermore we want to study the relations of different blur kernels to the bokeh, known from depth of field in cameras, and how it can be used to further improve separability of nominal data. To reduce performance drops related to fill rate problems, we want to investigate possibilities of multi-resolution approaches, without sacrificing accuracy.

8. Acknowledgements

The authors wish to thank Prof. Jochen Fröhlich (TU Dresden) for the fluid simulation data set used in case study 2. We also thank Ludwig Schmutzler for contributing to the supplemental video. This work was partially funded by BMBF Project No. 01IS14014 (ScaDS).

References

[BC87] BECKER R. A., CLEVELAND W. S.: Brushing Scatterplots. *Technometrics 29*, 2 (May 1987), 127–142. doi:10.2307/1269768. 3

[BW08] BACHTHALER S., WEISKOPF D.: Continuous scatterplots. *IEEE Transactions on Visualization and Computer Graphics 14*, 6 (Nov. 2008), 1428–1435. doi:10.1109/TVCG.2008.119. 2

[BW09] BACHTHALER S., WEISKOPF D.: Efficient and adaptive rendering of 2-d continuous scatterplots. In *Proceedings of the 11th Eurographics / IEEE - VGTC Conference on Visualization* (Aire-la-Ville, Switzerland, Switzerland, 2009), EuroVis'09, Eurographics Association, pp. 743–750. doi:10.1111/j.1467-8659.2009.01478.x. 2

[EAM14] EISEMANN M., ALBUQUERQUE G., MAGNOR M.: A nested hierarchy of localized scatterplots. In *Graphics, Patterns and Images (SIBGRAPI), 2014 27th SIBGRAPI Conference on* (Aug 2014), pp. 80–86. doi:10.1109/SIBGRAPI.2014.14. 3

[EG04] EILERS P. H. C., GOEMAN J. J.: Enhancing scatterplots with smoothed densities. *Bioinformatics 20*, 5 (2004), 623–628. doi:10.1093/bioinformatics/btg454. 2

[FB94] FURNAS G. W., BUJA A.: Prosection views: Dimensional inference through sections and projections. *Journal of Computational and Graphical Statistics 3* (1994), 323–385. doi:10.1080/10618600.1994.10474649. 3

[FD05] FRIENDLY M., DENIS D.: The early origins and development of the scatterplot. *Journal of the History of the Behavioral Sciences 41*, 2 (2005), 103–130. 2

[HCOB10] HELD R. T., COOPER E. A., O'BRIEN J. F., BANKS M. S.: Using Blur to Affect Perceived Distance and Size. *ACM Trans. Graph. 29*, 2 (Apr. 2010), 19:1–19:16. doi:10.1145/1731047.1731057. 3

[HE12] HEALEY C., ENNS J.: Attention and visual memory in visualization and computer graphics. *IEEE Transactions on Visualization and Computer Graphics 18*, 7 (July 2012), 1170–1188. doi:10.1109/TVCG.2011.127. 3

[JHM*13] JANETZKO H., HAO M., MITTELSTADT S., DAYAL U., KEIM D.: Enhancing scatter plots using ellipsoid pixel placement and shading. In *System Sciences (HICSS), 2013 46th Hawaii International Conference on* (Jan 2013), pp. 1522–1531. doi:10.1109/HICSS.2013.197. 2

[JNB*12] JOIA P., NONATO L. G., BRAZIL E. V., SOUSA M. C., DANIELS J., DOS SANTOS AMORIM E. P.: ilamp: Exploring high-dimensional spacing through backward multidimensional projection. In *Proceedings of the 2012 IEEE Conference on Visual Analytics Science and Technology (VAST)* (Washington, DC, USA, 2012), VAST '12, IEEE Computer Society, pp. 53–62. doi:10.1109/VAST.2012.6400489. 2

[JPC*11] JOIA P., PAULOVICH F., COIMBRA D., CUMINATO J., NONATO L.: Local Affine Multidimensional Projection. *IEEE Transactions on Visualization and Computer Graphics 17*, 12 (Dec. 2011), 2563–2571. doi:10.1109/TVCG.2011.220. 2

[JPKM07] JOURDAN F., PARIS A., KOENIG P.-Y., MELANÇON G.: Pixelization paradigm: First visual information expert workshop. Lévy P. P., Grand B., Poulet F., Soto M., Darago L., Toubiana L., Vibert J.-F., (Eds.), Springer, pp. 202–215. doi:10.1007/978-3-540-71027-1_18. 2

[Kei02] KEIM D. A.: Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics 8*, 1 (Jan. 2002), 1–8. doi:10.1109/2945.981847. 2

[KHD*09] KEIM D. A., HAO M. C., DAYAL U., JANETZKO H., BAK P.: Generalized Scatter Plots. *Information Visualization Journal (IVS)* (2009). 2009/12/24/online. doi:10.1057/ivs.2009.34. 2

[KMH01] KOSARA R., MIKSCH S., HAUSER H.: Semantic depth of field. In *Proceedings of the IEEE Symposium on Information Visualization 2001 (INFOVIS'01)* (Washington, DC, USA, 2001), INFOVIS '01, IEEE Computer Society, pp. 97–. 3

[KMH02] KOSARA R., MIKSCH S., HAUSER H.: Focus+context taken literally. *IEEE Computer Graphics and Applications 22*, 1 (Jan. 2002), 22–29. doi:10.1109/38.974515. 2, 3

[LH11] LAMPE O., HAUSER H.: Interactive visualization of streaming data with Kernel Density Estimation. In *Visualization Symposium (PacificVis), 2011 IEEE Pacific* (Mar. 2011), pp. 171–178. doi:10.1109/PACIFICVIS.2011.5742387. 2

[Lon13] LONG T. V.: isplom: Interactive with scatterplot matrix for exploring multidimensional data. In *Knowledge and Systems Engineering - Proceedings of the Fifth International Conference, KSE 2013, Volume 1, Hanoi, Vietnam, 17-19 October, 2013* (2013), pp. 175–186. doi:10.1007/978-3-319-02741-8_16. 3

[MG13] MAYORGA A., GLEICHER M.: Splatterplots: Overcoming Overdraw in Scatter Plots. *IEEE Transactions on Visualization and Computer Graphics 19*, 9 (Sept. 2013), 1526–1538. doi:10.1109/TVCG.2013.65. 2

[MS02] MATHER G., SMITH D. R. R.: Blur discrimination and its relation to blur-mediated depth perception. *Perception 31*, 10 (2002), 1211–1219. 3

[PKH04] PIRINGER H., KOSARA R., HAUSER H.: Interactive focus+context visualization with linked 2d/3d scatterplots. In *Second International Conference on Coordinated and Multiple Views in Exploratory Visualization, 2004. Proceedings* (July 2004), pp. 49–60. doi:10.1109/CMV.2004.1319526. 3

[SGT*03] SCHRAMMEL J., GILLER V., TSCHELIGI M., KOSARA R., MIKSCH S., HAUSER H.: Experimental evaluation of semantic depth of field, a preattentive method for focus+context visualization. In *Proceedings of the Ninth IFIP TC13 International Conference on Human-Computer Interaction* (2003), vol. Interact 03. 3

[Sil86] SILVERMAN B. W.: *Density Estimation for Statistics and Data Analysis*. CRC Press, apr 1986. 2

[Spe14] SPENCE R.: *Information Visualization: An Introduction*. Springer International Publishing, 2014. 3

[TFH11] TURKAY C., FILZMOSER P., HAUSER H.: Brushing Dimensions - A Dual Visual Analysis Model for High-Dimensional Data. *IEEE Transactions on Visualization and Computer Graphics 17*, 12 (Dec. 2011), 2591–2599. doi:10.1109/TVCG.2011.178. 3

[Tuf86] TUFTE E. R.: *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT, USA, 1986. 2

[vWvL93] VAN WIJK J. J., VAN LIERE R.: Hyperslice: Visualization of scalar functions of many variables. In *Proceedings of the 4th Conference on Visualization '93* (Washington, DC, USA, 1993), VIS '93, IEEE Computer Society, pp. 119–125. 3

[Wil05] WILKINSON L.: *The Grammar of Graphics (Statistics and Computing)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005. 2

[ZBDS12] ZINSMAIER M., BRANDES U., DEUSSEN O., STROBELT H.: Interactive Level-of-Detail Rendering of Large Graphs. *Visualization and Computer Graphics, IEEE Transactions on 18*, 12 (2012), 2486–2495. doi:10.1109/TVCG.2012.238. 2