Visual Quality Metrics

Enrico Bertini
DIS
University of Rome
“La Sapienza"
Via Salaria 113 - 00198
Rome, Italy
bertini@dis.uniroma1.it

Giuseppe Santucci
DIS
University of Rome
“La Sapienza"
Via Salaria 113 - 00198
Rome, Italy
santucci@dis.uniroma1.it

ABSTRACT

The definition and usage of quality metrics for Information Visualization techniques is still an immature field. Several proposals are available but a common view and understanding of this issue is still missing. This paper attempts a first step toward a visual quality metrics systematization, providing a general classification of both metrics and usage purposes. Moreover, the paper explores a quite neglected class of visual quality metrics, namely Feature Preservation Metrics, that allow for evaluating and improving in a novel way the effectiveness of basic Infovis techniques.

1. INTRODUCTION

In the last years several visual metrics have been proposed. First attempts come from Tufte that in [8] proposes a set of measures to estimate the "graphical integrity" of static (i.e., paper based) representations. Measures like the *lie factor*, that is the ratio between the size of an effect, as shown graphically, to its size in the data, or *data density* that takes into account the size of the graphic in relation to the number of displayed data, are examples of his attempt to systematically provide indications about the quality of the displayed image. Brath, in [7], starting from Tufte's proposal, defines new metrics for static digital 3D images. He proposes metrics such as *data density* (number of data points/number of pixels) that recall Tufte's approach. He provides metrics aiming at measuring the visual image complexity like the *occlusion percentage*, or the *number of identifiable points*, that is the number of visible data points whose position is identifiable in relation to every other visible data point. More recently, Peng et al. in [5] propose metrics able to measure clutter in parallel coordinates and scatter plot matrices, while Bertini et al. in [2] propose metrics for measuring several kinds of scatter plot degradation rising from data points collisions.

The way and the purpose of defining such metrics are quite different and, in order to provide a visual metrics systematization, we propose a classification based on three main classes, namely *Size Metrics*, *Visual Effectiveness Metrics*, and *Features Preservation Metrics*.

- Size Metrics are the basis of any other computation: number of data items, density, screen occupation percentage are metrics belonging to this class. Size Metrics can benefit from various perceptual studies [9], indicating the limits of human perception and thus providing some useful threshold values. They are quite intuitive and we do not discuss them in the rest of the paper.
- Visual Effectiveness Metrics are intended for measuring the image degradation, taking into account some disturbing factors. Measures of collisions, occlusion, outliers are example of this kind of metrics. Most of the available metrics belong to this class.
- Features Preservation Metrics are intended for measuring how correctly an image is representing some data characteristic. The Tufte lie factor or the comparison between the number of identifiable points with the actual data items number are examples of such kind of metrics. Very few proposals are available in this class and it is one of the goals of this paper to analyze these metrics in detail. It is worth noting that Features Preservation Metrics are grounded on Size and Visual Effectiveness Metrics.

In the following we will use the term *Visual Quality Metrics* for both Visual Effectiveness and Feature Preservation metrics.

Another, orthogonal, issue is metrics usage. Metrics should not be defined in splendid isolation but used for practical purposes: together with the metric classification we provide an overview of the typical visual quality metrics usage.

Summarizing, the paper contribution is twofold:

- it presents a visual metrics classification, focusing on visual quality metrics: Visual Effectiveness and Feature Preservation metrics. The methodological steps involved in the definition process are described, together with the typical visual quality metrics usage;
- it explores in detail, through a complete example, the quite neglected class of Features Preservation Metrics, providing a novel way of evaluating a visual technique effectiveness.

BELIV 2006 Venice, Italy

The paper is structured as follows: Section 2 describes the Visual Effectiveness Metrics, Section 3 deepens the notion of Features Preservation Metrics, showing a running example, Section 4 analyzes how quality metrics can be used in practice, and, Section 5 points out open issues.

2. VISUAL EFFECTIVENESS METRICS

These metrics measure the quality of a visualization in terms of one or more aesthetic criteria. A classical example is the principle of minimizing the number of edge crossings in a graph, with the basic assumption that crossings hinder the perception of paths in the graph. This class of metrics is characterized by the idea that the quality of a visualization depends on the presence of a disturbing factor and its level of influence. There are numerous examples of such kind of metrics besides the example of line crossings. Brath in [7] defines the *occlusion percentage* as: "number of data points completely obscured divided by the number of data points". Peng et al. in [6] propose a *clutter measure* for parallel coordinates: "*Due to the fact that outliers often obscure structure and thus confuse the user, clutter in parallel coordinates can be defined as the proportion of outliers against the total number of data points*". Similarly, Keim in [4], proposes optimal arrangements of data item measuring the similarity between data dimensions and devising algorithms that try to position similar dimensions as close as possible one to another. Bederson et al. measure the quality of a treemap in terms of how square-like its rectangles are: "we define the average aspect ratio of a treemap layout as the unweighted arithmetic average of the aspect ratios of all leaf-node rectangles, thus the lowest average aspect ratio would be 1.0 which would mean that all the rectangles were perfect squares" [1].

It is interesting to note that these metrics vary in terms of their level of generality, e.g., occlusion percentage can be applied to several visualizations while aspect-ratio, as defined in [1], is specific to treemaps. In order to apply the same principle to other visualizations, one needs to provide a broader definition and its area of intervention. We will discuss this aspect in later sections as it is common to other classes of metrics.

The steps needed to define Visual Effectiveness Metrics are:

1. *Come up with a visual optimization principle*: as a first step one needs to "detect" a quality principle that permits to state that a configuration is "better" than another (e.g., treemaps with square-like rectangles are more visually appealing than those with thin rectangles). A typical approach is to consider a (somewhat) self-evident principle and use it as reference to judge the quality of a visualization.

2. *Define a measure*: after a quality criterium has been drawn, it is necessary to find a way to measure it in practice for any visualization instance. There may be cases in which this is straightforward, as for example when calculating the number of occluded objects, while there may be others in which the connection between the criteria and its measure is not obvious.

The typical method, however, often misses some relevant parts. First, there should be an initial assessment to understand the extent to which the proposed metric adheres to the aspect one wants to measure. In principle, the same aspect (or similar aspects) may be captured by different metrics with different expressive power (e.g., clutter in parallel coordinates can be measured in terms of number of outliers or number of line crossings); selecting a good one should be part of the design of a novel metric. Then, the impact of the metric on real cases should be taken with more care. Quality principles should always be grounded on user studies, assuring a close matching between the principle and the real performance of users. This last point is particularly relevant since too often the proposed metrics lack of empirical grounding.

3. FEATURES PRESERVATION METRICS

In spite of the fact that Tufte introduced in 1986 a nice example of these metrics, the lie factor, this class is quite unexplored and few proposals are available. Here we deepen the Feature Preservation Metrics definition providing several methodological steps that, in our opinion, are needed to define such kind of metrics. In order to make these steps more clear we use a running example, dealing with density differences in the context of 2D scatter plot; the example is not a toy one and it is intended to provide all the details needed to make clear our methodology.

Feature Preservation Metrics are comparative metrics: a measure is performed on the data set, capturing a data characteristic; a second measure is performed on the visualization, measuring the same characteristic. The two figures are compared, calculating how the characteristic is correctly represented in the visualization. As a very simple example, assume that we are interested in measuring how correctly a 2D scatter plot represents a data set cardinality. We can count the data elements number n, the turned on pixels p (because of collisions very likely $p < n$) and define the metric $K = \frac{p}{n}$. The greater the K value (ranging between 0 and 1) the better the image represents the data set cardinality. This example let us point out the following considerations:

- the metric is associated with a feature that we want to preserve in the actual visualization; to this aim, we need to devise the interesting feature *and* formally characterize it. Note that this is, in general, not a trivial task;

- we need two measures, one in the data space and one in the visualization space; moreover we need a comparison mechanism. Again, these activities may be not trivial ones;

- the metrics depends on the *actual* data set and on the *actual* image parameters: e.g., changing the image size results in a different metric value: e.g., the K metrics improves increasing the image size.

Starting from these considerations we are ready to outline the methodological steps involved in the metric definition process.

1. *Choose the target feature*: analyzing the clues the Infovis technique presents to the end users gives us the list of features we can use in the process of metric definition. As an example, a 2D scatter plot allows the

end user to discover, among the others, trends, correlations, clusters, and density differences. Among them we choose the last one as the feature we want to analyze.

2. *Formally define the feature in (1)the data space and (2)in the visualization space*: Once the feature has been selected, two formal definitions are needed, one in the data space and one in the visualization space. As stated before, this is not a trivial task and strongly depends on the chosen visualization technique and feature. Again, we illustrate this step using the 2D scatter plot example. In order to formally define density differences we assume the image is displayed on a rectangular area and that small squares of area A divide the space in *sample areas* (SA) where density is measured. As an example, if we are plotting data points on a monitor of 1280×1024 pixels (visualization space) using real values from a X-Y semi-plane of size of 13"x10.5" (data space) and we choose SA of side $l = 0,08$ inch, the area is covered by 160×128 squared sample areas whose dimension in pixels is 8×8 (note that all these metrics are Dimension Metrics).

   For each $SA_{i,j}$ we calculate two different densities: *data density* and *represented density*.

   Data density is defined as $D_{i,j} = n_{i,j}$ where $n_{i,j}$ is the number of data points that fall into sample area $A_{i,j}$. Measuring data density differences is the formal definition of the feature in the data space.

   Represented density is defined as $RD_{i,j} = p_{i,j}$ where $p_{i,j}$ is the number of distinct active pixels in $SA_{i,j}$. The number of different values that a represented density can assume depends on the size of sample areas. If we adopt sample areas of 8×8 pixels the number of different not null represented densities is 64. Measuring represented density differences is the formal definition of the feature in the visualization space.

   Data densities are measured in a continuous space, while represented densities in a discrete one; because of collisions the number of active pixels on a sample area $SA_{i,j}$ will likely be less than the plotted points so $RD_{i,j} \leq D_{i,j}$.

3. *Validate the visual definition against user perception*: while the feature definition in the data space is an abstract one, the visual counterpart is a physical one and can be influenced by user perception. In this phase we have to relate the numerical values devised in step 2 against users' perceptions. In the 2D scatter plot example we are analyzing the density differences presented to the user through the represented density. Using pure numerical differences between sample areas to decide whether a data density difference is well represented or not by the corresponding represented densities is not correct. As an example, a sample area containing 55 active pixels is considered denser than another one containing 54 active pixels while both of them look the same to the end user. In order to overcome this limitation we have to perform a user study to understand how users perceive differences in represented densities. We conducted such a study, asking the users to recognize few more dense areas on a uniform background (basis), repeating the test for different bases and different density differences. In the end we came up with a function $minimum\delta(RD_{i,j})$ returning the minimum increment a sample area must show to be perceived as denser than $SA_{i,j}$ by 70% of users. Using this function we can claim that a difference in data density between two sample areas (feature in the data space) is well represented iff the corresponding represented densities (feature in the visualization space) differ at least of $minimum\delta$.

4. *Define the metric*: once the metric building blocks have been evaluated against the user perception it is possible to compare the figures coming from the data space against the visual ones. In the example of the K metric defined in the beginning of this section a simple division was enough; considering the running 2D scatter plot example we need a more complex calculation and we used an algorithm. The idea is to compare all the possible sample area pairs, checking whether the density differences in the data (data densities) are correctly represented by the corresponding visual density (represented densities). Moreover, in order to take into account the relevance of a comparison between two sample areas, we weight each pair using the number of points falling in the two sample areas. More precisely we defined the PDiff (Perceptual Diff) function as follows:

$$PDiff(x,y) = \begin{cases} 1 & \text{if } x \geq y + y \times minimum\delta(y) \\ -1 & \text{if } y \geq x + x \times minimum\delta(x) \\ 0 & \text{otherwise} \end{cases}$$

   and we defined the $match(i,j,k,l)$ function that returns true iff $PDiff(D_{i,j}, D_{k,l}) = PDiff(RD_{i,j}, RD_{k,l})$. The *match* function captures the spirit of the Features Preservation Metrics: it compares a feature (density differences) in the data space against the same feature in the visualization space. The *match* function is used to compute the WPLDDr (Weighted *Perceptually* Lost Data Densities ratio) metric:

```
function WPLDDr(){
 Let couples=0; \* weighted SA couples
 Let sum=0;     \* weighted non matching SA couples
 foreach distinct pair(SA[i][j], SA[k][l]){
   couples = couples + pt(SA[i][j]) + pt(SA[k][l]);
   if ( NOT match(i, j, k, l) )
      sum = sum + pt(SA[i][j])+ pt(SA[k][l]);}
 return (sum / couples);}
```

   where $pt(SA_{i,j})$ returns the number of data points falling in a SA.

   The variable *sum* contains the number of weighted non matching couples encountered during the iterations; dividing it by the weighted total number of possible distinct couples we obtain the weighted percentage of non matching sample areas ranging between 0 and 1 (the lower the better).

Roughly speaking we can say that the WPLDDr metric measures the percentage of data whose density differences are hidden from the users.

As the example clarifies, the overall process requires a great effort but the advantages are quite evident: Feature Preservation Metrics allows for a quick, sound, and *objective* evaluation of an image quality. Concrete examples of their usage will be provided in Section 4

4. METRIC BASED EVALUATION

This section deals with the practical visual quality metrics usage, describing three main paths that can be followed to evaluate/ameliorate an image.

- *Threshold comparison.* The simplest way of using a visual quality metrics is to compare it against a threshold value. This allows for evaluating the quality of a certain image and can be used as a trigger to start some ameliorating techniques or to discard the image. Moreover, since most quality metrics improve while increasing the image size, given a data set and a set of threshold values, it is possible to compute the minimum display area size preserving all the threshold values, allowing the system to optimize the screen usage. This is particularly useful when the screen hosts multiple visualizations.

- *Comparison and evaluation of algorithms and visual techniques.* Using a visual quality metric it is possible to evaluate the effectiveness of an ameliorating algorithm analyzing the metric improvement (if any). Moreover, the same approach allows for comparing two or more algorithms that share the same goal. A more complex usage foresees the possibility of using a Feature Preservation Metric to compare the effectiveness of two different visual techniques in representing the same data feature. As an example, assume that we computed the metric Kpc for parallel coordinates in a similar way we defined the metric K (in the begin of Section 3) for scatter plots: in principle we can visualize the same data set on both a scatter plot and parallel coordinates (with defined size and axes ordering) and to decide, comparing K and Kpc, which technique best represents the data set cardinality. Note that this relies on the assumption that (1) a formal definition of distinguishable polyline has been provided and (2) suitable user studies have validated the two metrics.

- *Algorithm driving.* In some cases, visual quality metrics are used to drive an ameliorating algorithm. As an example, Peng et al. in [5] propose a metric able to measure clutter in parallel coordinates: such a metric is used by an axes reordering algorithm to find the optimum configuration. In a similar way, Bertini et al. in [2] propose a composition of two or more quality metrics that constitutes the objective function a sampling algorithm optimizes.

5. CONCLUSION AND OPEN ISSUES

In this section we want to outline some open issues that raised during visual quality metrics discussion.

- *Finding appropriate threshold values* - when a metric is compared against a quality threshold, it is necessary to find meaningful threshold values. This usually requires to run some user studies that permit to state that a visualization is considered "unacceptable" when some metric values are over a predefined limit. The problem is that these studies may be not trivial since they often have to cope with unexplored aspects of human perception.

- *Designing composite objective functions* - in some cases, as when metrics are used to guide optimization algorithms, there is the need to use compound metrics to take into account an overall quality factor. However, the composition of metrics is not straightforward. For example, assuming that it is reasonable to make a linear combination of the chosen metrics, there is the basic problem of deciding what weights to assign to the single involved metrics.

- *Describing visual features* - when dealing with feature preservation metrics, it is necessary to: 1)find features that capture relevant aspects of a visualization; 2)find a way to describe them in formal terms. Both activities are not trivial. Given a specific visualization, it is not clear what are the kind of visual patterns the user looks for when exploring the data. In addition, even when a feature is clearly detected, its expression in formal terms may be cumbersome. As an example, consider the definition of visual cluster in parallel coordinates. While it is evident that an n-dimensional cluster is a relevant visual aspect the user may want to detect, it is not clear how to describe it in precise terms.

- *Level of generality* - metrics can be expressed at various levels of generality. A large proportion of metrics are developed for the specific purpose of a single visualization (typically with the intent of guiding some optimization algorithms), without any attempt or opportunity to employ them beyond their scope. More generic metrics can be useful to compare different visualization techniques on common problems and can be useful to guide the design of new visualizations. To date, we do not know of any attempts to understand the scope of existing metrics or to classify them. For example the "occlusion percentage" is a generic metric that cannot be applied to visualizations where items cannot overlap (a quite large class of visualizations!). We believe such a classification is feasible. One approach could be to characterize the design space (e.g., as in [3]) in classes of visualizations and provide a set metrics that make sense for entire classes of visualizations.

In conclusion, visual quality metrics are a necessary step for information visualization but in literature very few attempts to analyze them in general terms are available. In this paper we made an attempt in that direction and we described a novel class of metrics we called "feature preservation". Our goal is to elicit some fresh ideas on these quite neglected topics, suggesting potential issues to look for in order to progress in this direction.

6. REFERENCES

[1] B. B. Bederson, B. Shneiderman, and M. Wattenberg. Ordered and quantum treemaps: Making effective use of 2d space to display hierarchies. *ACM Trans. Graph.*, 21(4):833–854, 2002.

[2] E. Bertini and G. Santucci. Give chance a chance: modeling density to enhance scatter plot quality through random data sampling. *Information visualization*, to appear, 2006.

[3] S. K. Card and J. Mackinlay. The structure of the information visualization design space. In *INFOVIS '97: Proceedings of the 1997 IEEE Symposium on Information Visualization (InfoVis '97)*, page 92, Washington, DC, USA, 1997. IEEE Computer Society.

[4] D. A. Keim. Designing pixel-oriented visualization techniques: Theory and applications. *IEEE Transactions on Visualization and Computer Graphics*, 6(1):59–78, 2000.

[5] W. Peng, M. O. Ward, and E. A. Rundensteiner. Clutter reduction in multi-dimensional data visualization using dimension reordering. In *Proc. of IEEE Symposium on Information Visualization*, pages 89–96. IEEE Computer Society, 2004.

[6] W. Peng, M. O. Ward, and E. A. Rundensteiner. Clutter reduction in multi-dimensional data visualization using dimension reordering. In *INFOVIS '04: Proceedings of the IEEE Symposium on Information Visualization (INFOVIS'04)*, pages 89–96, Washington, DC, USA, 2004. IEEE Computer Society.

[7] B. Richard. Concept demonstration: Metrics for effective information visualization. In *Proc. of IEEE Symposium on Information Visualization*, pages 108–111. IEEE Service Center, Phoenix, AZ, 1997.

[8] E. R. Tufte. *The visual display of quantitative information.* Graphics Press, 1986.

[9] C. Ware. *Information visualization: perception for design.* Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2000.