Erschienen in: EuroVis 2014 : the Eurographics Conference on Visualization ; 9-13 June 2014, Swansea, Wales, UK ; EuroVis Short Papers / Margit Pohl ... (Hrsg.). - Eurographics Association, 2014. - S. 91-95. - ISBN 978-3-905674-69-9
https://dx.doi.org/10.2312/eurovisshort.20141163

Revisiting Perceptually Optimized Color Mapping for High-Dimensional Data Analysis

Sebastian Mittelstädt[1], Jürgen Bernard[2], Tobias Schreck[1], Martin Steiger[2], Jörn Kohlhammer[2] and Daniel A. Keim[1]

[1]University of Konstanz, Germany
[2]Fraunhofer Institute for Computer Graphics Research IGD, Germany

Abstract

Color is one of the most effective visual variables since it can be combined with other mappings and encode information without using any additional space on the display. An important example where expressing additional visual dimensions is direly needed is the analysis of high-dimensional data. The property of perceptual linearity is desirable in this application, because the user intuitively perceives clusters and relationships among multidimensional data points. Many approaches use two dimensional colormaps in their analysis, which are typically created by interpolating in RGB, HSV or CIELAB color spaces. These approaches share the problem that the resulting colors are either saturated and discriminative but not perceptual linear or vice versa. A solution that combines both advantages has been previously introduced by Kaski et al.; yet, this method is to date underutilized in Information Visualization according to our literature analysis. The method maps high-dimensional data points into the CIELAB color space by maintaining the relative perceived distances of data points and color discrimination. In this paper, we generalize and extend the method of Kaski et al. to provide perceptual uniform color mapping for visual analysis of high dimensional data. Further, we evaluate the method and provide guidelines for different analysis tasks.

1. Introduction

Ware and Beatty [WB88] performed an experiment, in which five dimensional data was mapped to two spatial and three color dimensions. The results indicated that each additional color dimension is as useful as an additional spatial dimension for cluster identification. Other guidelines [Bre96, War12] suggest mapping two dimensions to hue and saturation (or lightness). This results in few distinguishable colors, which is in most cases enough to visualize effective overviews but lacks in precision [War12]. In high dimensional data analysis the focus is typically on exploring the relations of data items. Perceptual similarity is already modeled in color spaces such as CIELAB. If the distances in the data space are mapped to perceptual distances in the color space, the analysts will perceive the relations of data items by interpreting the perceptual similarity of their colors. In this case, the color mapping is not bound to a fixed number of dimensions and is able to encode high-dimensional data relations. Unfortunately, only a subspace of CIELAB can be visualized on current displays. This subspace (or *bounds*) is of non-rectangular shape that makes interpolation and other arithmetics for color mapping very complex (see Figure 1C). Rectangular parts of this subspace as defined by a maximum surrounded box provide perceptual linear mappings but result in fewer discriminable colors (see Figure 1B). Other techniques use two-dimensional color maps that are often created by interpolation between four corner colors. This results in highly discriminable colors

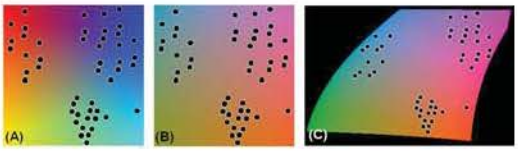

Figure 1: ***Two dimensional colormaps.*** *The position of black dots represent the color of multidimensional data points. (A) 2D colormap in RGB: colors are saturated, however, not perceptually uniform; (B) Rectangular sub plane of CIELAB: perceptually uniform, but less saturated colors; (C) Kaski et al.: saturated and perceptually uniform colors.*

Figure 2: ***Wine data set [BL13]. 13 attributes describe three classes of wines. The data is projected with MDS to four dimensions and visualized in the scatter plots (x-, y-axis and two dimensions are mapped to color). (A) 2D RGB color map: classes are not separated and colors reveal (wrong) large distance between data points; (B) CIELAB sub plane: distances are preserved but classes are not separated; (C) Our Method: Three classes are separated and local distances of class elements are preserved.***

but these color maps are not perceptually linear. The user may group data points of the same cluster differently (in Figure 1A clusters span over two or more color hues). Kaski et al. previously introduced a method [KVK00] that projects high-dimensional data with a self-organizing map to two dimensions and fits the data into the bounds of CIELAB (see Figure 1C). The color assignment supports the user in recognizing clusters and preserves the relationships of clusters while maximizing the exploitation of the color space.

In this paper, we revisit the method of Kaski et al. and adjust it to the needs of visual analysis. Our method provides improved color mapping for high dimensional data points, which can be used in any visual design since color is an additional design variable that is most effective in combination with other visual variables such as position. A result of our method is illustrated in Figure 2. We claim the following contributions: 1) **generalization** of the method with further projection methods, and extension to 3D target color spaces; 2) **efficient heuristics** for practical use; 3) **cost functions** to further support analysis tasks; 4) **evaluation** of different configurations and methods in a user study.

2. Related Work

General guidelines on selecting color maps can be found in [War88, BRT95, RTB96, Rhe00, SSSM11, War12]. For more than one dimension, color seems to be problematic. If mapped to the receptor level (e.g., RGB or LMS) we perceive the mixture and can infer similarity [WB88] but cannot separate the input from each dimension. Bivariate color schemes that meet several perceptual issues are discussed in [Bre96, HB03]. These schemes, however, do only support a limited number of color levels. An extension to the approach is introduced in [GGMZ05, GCML06]. The method uses interaction and bell shaped rasters in the CIELAB space to produce diverging colors. There is evidence that two-dimensional color maps are unintelligible for encoding certain dimensions [WF80]. However, under a different perspective of visualizing the similarity of data points or clusters these color maps have shown their usefulness in many papers. For example, in [Him00, BvLBS11] high dimensional data is projected to a lower (two) dimensional space and then scaled to fit a two dimensional color map. Most methods interpolate in RGB or CIELAB between fixed color anchors in the corners. Some methods also use uniform planes of CIELAB [WD08].

3. Color Mapping for High-Dimensional Data Analysis

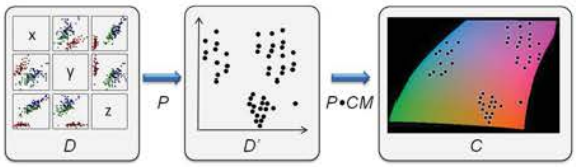

Figure 3: ***Schematic Approach. High dimensional data D is projected with P to low dimensional space D', which is transformed with P • CM to fit into color space C.***

For color mapping of high-dimensional data, we see different requirements for the visual analysis tasks as described in [TFS08] on the task model in [AA06]: Group 1: identification and comparison of data points and clusters; Group 2: lookup of clusters and classes. Group 1 requires an accurate match of all distances in the data space and perceptual distances in the mapped colors. Group 2 requires perceptual separation of classes and known clusters. Assigning clusters to distinct categorical colors works well for group 2. However, with this approach properties of the clusters are lost (e.g., the correlation of dimensions or relations of cluster elements). In the ideal case, data relations are preserved in the coloring. This requires a model of perceptual similarity that is implemented in CIELAB. The exploitation of the whole color space supports color discrimination and thus, lookup of clusters. To guarantee full exploitation, the method must adapt the data to the non-linear shape of CIELAB. The intuition behind the method of Kaski et al. is that high-dimensional data is projected into the low-dimensional color space and then *fitted* to the bounds of the color space (see Figure 3). The fitting is an optimization algorithm that minimizes target cost functions. In the following, we generalize the method and provide cost functions that meet the requirements of different analysis tasks.

3.1. Cost Functions & Perceptual Metrics

Definitions: D is the set of all model vectors $m_i \in \mathbb{R}^m$ describing all data elements i. C is the set of all colors $c_i \in \mathbb{R}^n$ in the target color space. $P : \mathbb{R}^m \mapsto \mathbb{R}^n$ is the projection of the high-dimensional model vectors in the lower dimensional target space. D' being the set of model vectors $m'_i \in \mathbb{R}^n$ (note that $D' \neq C$). $CM : \mathbb{R}^n \mapsto C$ is the color assignment of m'_i

to color c_i. G is the set of clusters in D with $g(i)$ being the cluster of data element i. H_x being the convex hull and $V(H_x)$ the volume of a set or cluster x (e.g., $V(H_D)$ represents the volume of the convex hull of D in $\mathbb{R}^m$).

Preservation of data relations. A quality measure of a projection from high- to low-dimensional space is the preservation of all relative distances that can, for example, be measured by the Sammon's stress measure (1). However, the preservation of all pairwise distances is typically impossible. Therefore, Kaski et al. preserve the relative distances within a cluster to increase the accuracy of the projection locally (2).

$$f_1 = \sum_{i \in D} \sum_{j \neq i} \frac{\left(d\left(m_i, m_j\right) - d\left(m_i', m_j'\right)\right)^2}{d\left(m_i, m_j\right)} \tag{1}$$

$$f_2 = \sum_{i \in D} \sum_{j \in g(i)} \left(d\left(m_i, m_j\right) - d\left(m_i', m_j'\right)\right)^2 \tag{2}$$

Color space exploitation. Another important property of good color mappings is that the mapping exploits as much of the color space in order to provide distinguishable colors. Kaski et al. rigidly scale the vectors m_i' with a parameter k. It is increased to let D' occupy more of the available color space C. The original method estimates the distance of m_i' to its perceptually closest color c_i that can be displayed on the output device. This does not measure the exploitation of the color space. It measures the distortion of CIELAB colors that lay beyond the color space bounds (3). The exploitation of the color space can be measured by the overlap of the color space in $\mathbb{R}^n$ and the projected data $D' \in \mathbb{R}^n$. This can be approximated by computing the volume of the intersection of the convex hulls of $H_{D'}$ and H_C (4).

$$f_3 = \sum_{i \in D} d\left(m_i', c_i\right) \tag{3}$$

$$f_4 \approx 1 / V\left(H_C \cap H_{D'}\right) \tag{4}$$

Preservation of clusters. Preserving the local distances within a known cluster and ignoring the interrelations of clusters makes the color mapping very flexible. The data can adapt to the non-linear shape of the color space, which separates clusters well. However, if the task requires also to perceive interrelations of clusters, this method will produce misleading results. Kaski et al. introduced a heuristic that measures the "orderliness" of clusters based on a SOM grid. We propose a different function that preserves the relative distances of cluster centroids $\bar{m}_r$ with $r \in G$ (5), because the heuristic cannot be applied in high-dimensional spaces.

$$f_5 = \sum_{\bar{m}_r, r \in G} \sum_{\bar{m}_s \neq \bar{m}_r} \left(d\left(\bar{m}_r, \bar{m}_s\right) - d\left(\bar{m}_r', \bar{m}_s'\right)\right)^2 \tag{5}$$

Further, the original method does not measure how well clusters are separated or do overlap. This can be approximated with the inverse centroid distance (6) and the intersection of convex hulls (7). Another issue in visualizing clusters with color is that we will overestimate the number of clusters or see noise if there are only few present [WB88]. Our cognitions tries to differentiate between groups and objects based on their color (hue). If a cluster spans over the whole color space, it is likely that it is perceived as multiple clusters. A cost function (8) measures the pairwise color distances of cluster elements that are higher than a threshold t (we found that for $t = 30$ in CIELAB clusters are correctly perceived).

$$f_6 \approx 1 / \sum_{\bar{m}_r, r \in G} \sum_{\bar{m}_s \neq \bar{m}_r} d\left(\bar{m}_r', \bar{m}_s'\right) \tag{6}$$

$$f_7 \approx \sum_{r \in G} \sum_{s \neq r} V\left(H_r \cap H_s\right) \tag{7}$$

$$f_8 = \sum_{r \in G} \sum_{i, j \in r} \sum_{j \neq i} max\left(d\left(m_i', m_j'\right) - t, 0\right) \tag{8}$$

Combination of cost functions. The optimization goal is to minimize the multi-objective cost functions. We scalarize and sum the functions (9). Note, that this may be different with other optimization methods. Scalar α_i is used to make the cost functions comparable. This parameter can be estimated, for example, by evaluating a "bad" random solution and normalize all cost functions. λ_i steers the influence of the cost function i on the mapping and configures the method for different analysis tasks. Details can be found in Section 4.

$$f = \sum_{i=1}^{8} \lambda_i \cdot \alpha_i \cdot f_i \tag{9}$$

3.2. Optimization Algorithms & Heuristics

The optimization goal can be reached by minimizing the sum of cost functions by a variety of optimization algorithms. Kaski et al. use a stochastic gradient method. We found that *particle swarming* [KE*95] provided good results. However, we consider the choice of the optimization algorithm as interchangeable part of our method. The optimization goal $min(f)$ has several issues: 1) f is not continuous so that f' can only be approximated; 2) in high dimensional spaces f_1 and f_2 suffer under the curse of high dimensionality. Sophisticated projections P exist that effectively map $\mathbb{R}^m$ to $\mathbb{R}^3$. We, in practice, use a standard projection technique P such as MDS. The fitting to CIELAB is then applied in a post-processing step (see Figure 3). Global and/or local distances can be preserved by P. Therefore, a heuristic can use translation (in three dimensions), scaling and rotation (about three axis; centers as fix points) on the projected data D' or on clusters in D' to minimize the cost functions. This has the advantage that the parameter vector in the optimization is of low dimensionality. This results in seven dimensions for the whole data D' if all pairwise distances shall be preserved or seven dimensions per cluster if the task is focused on the lookup of clusters.

4. Evaluation

Goal and Task. We evaluated our method empirically with an experiment introduced by Ware and Beatty [WB88]. The goal was to measure the accuracy of users identifying the number of clusters in a visualization. The participants were shown a multi-dimensional data set in a scatter plot (as in Figure 2). Two spatial dimensions were encoded by x- and y-axis and two or three dimensions were encoded by color (note,

Task / Costfunction	Elementary		Synoptic		
	Compare data points global	Compare data points local (cluster)	Identify clusters	Lookup clusters	Compare clusters
f1	✔	✘	✔	✘	✘
f2	✘	✔	✘	(✔)	(✔)
f5	✘	(✔)	✘	(✔)	✔
f6	✘	(✔)	✘	✔	✔
f7	✘	(✔)	✘	✔	✔
f8	✘	(✔)	✘	✔	✔

✔ Activated ✘ Deactivated (✔) Optional

Table 1: *Combinations of cost functions for analysis tasks.*

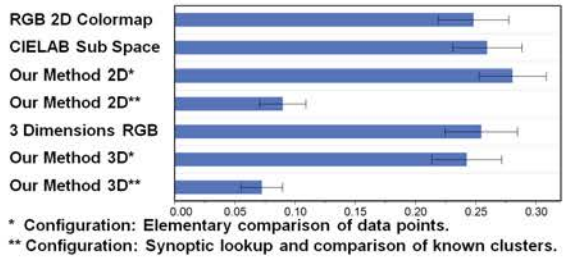

Figure 4: ***Evaluation Results.*** *Averaged normalized error* $(|\frac{\#userEstimate}{\#clusters} - 1|)$ *and standard deviation.*

that we reduced the number of dimensions in order to be comparable with related methods). The participants were asked to estimate the number of clusters in each scatter plot. Note, that counting the number of clusters is not trivial and involves elementary and synoptic tasks (see Table 1). The participant has to compare the spatial and color distribution of the data points, which is the elementary task of comparing data points globally. The participant has to group the data points and further has to differentiate between spatial distribution and color since clusters may overlap spatially or in the color space. With this, the participant identifies clusters (synoptic task) and is able to count the number of clusters in the plot.

Experiment Factors. We evaluated seven color mappings with three state-of-the-art techniques and our method. Our method can be configured in multiple ways, however, we selected two versions. One was configured for the elementary comparison task and the other was configured to preserve known clusters (lookup and comparison task, see Table 1). For four dimensional data we used our method with a fixed lightness of $L = 60$ (2D version) and state-of-the-art methods that were two dimensional color maps in RGB and CIELAB (see Figure 1). For five dimensional data we used Ware's and Beatty's method to map three dimensions directly to red, green and blue [WB88] and our method that exploits the full CIELAB space (3D version). The color mapping of Kaski et. al. requires the SOM projection. We excluded the uncertainties of projections. Thus, our method was comparable to the state-of-the-art but not to the method of Kaski et al.

Experimental Design. We conducted a user study with 8 visualization and data analysis experts. The study was within-subject designed. Each participant performed 18 tasks with each color mapping. The order of color mappings was randomized. The data was created according to [WB88], with the number of clusters (1 to 6 clusters), number of cluster elements (min: 30, max: 80), cluster positions and cluster shapes being randomized in each trial.

Results and Discussion. The summary of results is illustrated in Figure 4. With our method preserving clusters (2D and 3D version) users were significantly more accurate than with all other mappings on estimating the correct number of clusters (paired U-Test: $p<0.001$). This method supports the synoptic lookup and comparison task of clusters and still preserves the local data distances. The configuration implies that clusters are known a priori, which is typically not the case in the cluster identification task. However, this shows the advantage of concerning separation of known clusters in the color mapping. Our method for cluster identification provides correct perceptually mappings. The 3D version performed well, however, not significantly better than the state-of-the-art methods. The effect of perceiving more clusters if few are present [WB88] seems to compensate the benefits of perceptual linearity. Especially, since our method tries to exploit the whole color space and preserves all pairwise distances. We presented cost functions that are designed to support two opposing groups of analysis tasks. We argue that these functions are a sound basis for the analysis in realistic scenarios. However, wee see further research to support different analysis tasks and to improve visual cluster identification. It will be interesting to find trade offs in real applications. Further, we see future work to estimate the benefit of preserving global cluster relations and local cluster element relations in comparison to categorical color mapping.

Implications. Our guidelines are summarized and illustrated in Table 1. Note, that $f3$ and $f4$ are independent of the task and should always be activated. If the task is to visually identify high dimensional clusters, standard two dimensional color maps will perform as well as our technique. However, if the task also implies the comparison of data items, our technique ($f1$) will provide perceptual correct mappings. When clusters are known a priori and should be perceptually preserved, our method ($f2$, $f5$-$f8$) should be used since it preserves local distances and supports lookup of clusters.

5. Conclusions

In this paper, we present an extension to the method of Kaski et al. [KVK00] to project high dimensional data to perceptual linear color spaces. Our method preserves the relationships of data items and supports the user in recognizing clusters while maximizing the exploitation of the CIELAB color space. We provide guidelines on how to configure our method for different analysis tasks and evaluated different versions of our method empirically. The results show that our method outperforms other methods in the lookup task of clusters but also highlighted that further research is required to improve cluster identification with color.

References

[AA06] ANDRIENKO N., ANDRIENKO G.: *Exploratory analysis of spatial and temporal data.* Springer, 2006. 2

[BL13] BACHE K., LICHMAN M.: UCI machine learning repository, 2013. URL: http://archive.ics.uci.edu/ml. 2

[Bre96] BREWER C. A.: Guidelines for selecting colors for diverging schemes on maps. *The Cartographic Journal 33*, 2 (1996), 79–86. 1, 2

[BRT95] BERGMAN L., ROGOWITZ B., TREINISH L.: A rule-based tool for assisting colormap selection. In *Proceedings of the 6th conference on Visualization* (1995), p. 118. 2

[BvLBS11] BREMM S., VON LANDESBERGER T., BERNARD J., SCHRECK T.: Assisted descriptor selection based on visual comparative data analysis. In *Computer Graphics Forum* (2011), vol. 30, Wiley Online Library, pp. 891–900. 2

[GCML06] GUO D., CHEN J., MACEACHREN A. M., LIAO K.: A visualization system for space-time and multivariate patterns (vis-stamp). *IEEE Transactions on Visualization and Computer Graphics 12*, 6 (2006), 1461–1474. 2

[GGMZ05] GUO D., GAHEGAN M., MACEACHREN A. M., ZHOU B.: Multivariate analysis and geovisualization with an integrated geographic knowledge discovery approach. *Cartography and Geographic Information Science 32*, 2 (2005), 113–132. 2

[HB03] HARROWER M., BREWER C.: Colorbrewer.org: an online tool for selecting colour schemes for maps. *The Cartographic Journal 40*, 1 (2003). 2

[Him00] HIMBERG J.: A SOM based cluster visualization and its application for false coloring. In *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks* (2000), vol. 3, IEEE, pp. 587–592. 2

[KE*95] KENNEDY J., EBERHART R., ET AL.: Particle swarm optimization. In *Proceedings of the IEEE international conference on neural networks* (1995), vol. 4, pp. 1942–1948. 3

[KVK00] KASKI S., VENNA J., KOHONEN T.: Coloring that reveals cluster structures in multivariate data. *Australian Journal of Intelligent Information Processing Systems 6*, 2 (2000), 82–88. 2, 4

[Rhe00] RHEINGANS P.: Task-based color scale design. In *28th AIPR Workshop: 3D Visualization for Data Exploration and Decision Making* (2000), International Society for Optics and Photonics. 2

[RTB96] ROGOWITZ B., TREINISH L., BRYSON S.: How not to lie with visualization. *Computers in Physics 10*, 3 (1996). 2

[SSSM11] SILVA S., SOUSA SANTOS B., MADEIRA J.: Using color in visualization: A survey. *Computers & Graphics 35*, 2 (2011), 320–333. 2

[TFS08] TOMINSKI C., FUCHS G., SCHUMANN H.: Task-driven color coding. In *Proceedings of the 12th International Conference on Information Visualisation.* (2008), IEEE, pp. 373–380. 2

[War88] WARE C.: Color sequences for univariate maps: Theory, experiments and principles. *IEEE Computer Graphics and Applications 8*, 5 (1988). 2

[War12] WARE C.: *Information visualization: perception for design.* Elsevier, 2012. 1, 2

[WB88] WARE C., BEATTY J. C.: Using color dimensions to display data dimensions. *Human Factors: The Journal of the Human Factors and Ergonomics Society 30*, 2 (1988), 127–142. 1, 2, 3, 4

[WD08] WOOD J., DYKES J.: Spatially ordered treemaps. *IEEE Transactions on Visualization and Computer Graphics 14*, 6 (2008), 1348–1355. 2

[WF80] WAINER H., FRANCOLINI C. M.: An empirical inquiry concerning human understanding of two-variable color maps. *The American Statistician 34*, 2 (1980), 81–93. 2