Vision, Modeling, and Visualization (2019)
H.-J. Schulz, M. Teschner, and M. Wimmer (Eds.)

Learning a Perceptual Quality Metric for Correlation in Scatterplots

L. Wöhler[1], Y. Zou[1], M. Mühlhausen[1], G. Albuquerque[1,2], M. Magnor[1]

[1]Computer Graphics Lab, TU Braunschweig, Germany
[2] Software for Space Systems and Interactive Visualization, DLR Braunschweig, Germany

Abstract

Visual quality metrics describe the quality and efficiency of multidimensional data visualizations in order to guide data analysts during exploration tasks. Current metrics are usually based on empirical algorithms which do not accurately represent human perception and therefore often differ from the analysts' expectations. We propose a new perception-based quality metric using deep learning that rates the correlation of data dimensions visualized by scatterplots. First, we created a data set containing over $15,000$ *pairs of scatterplots with human annotations on the perceived correlation between the data dimensions. Afterwards, we trained two different Convolutional Neural Networks (CNN), one extracts features from scatterplot images and the other directly from data vectors. We evaluated both CNNs on our test set and compared them to previous visual quality metrics. The experiments show that our new metric is able to represent human perception more accurately than previous methods.*

CCS Concepts
• ***General and reference*** → ***Metrics;*** • ***Human-centered computing*** → ***Visual analytics;*** • ***Computing methodologies*** → *Perception; Neural networks;*

1. Introduction

With continuing advances of digital technologies, the amount of generated and collected data is rapidly increasing. In order to extract meaningful information from these collections, effective visual exploration techniques are required. Visual quality metrics are measures that were introduced to support the visual exploration of high dimensional data sets. They are used to automatically rate lower-dimensional mappings of data sets with regard to a previously defined task chosen by the analysts. Most current metrics are based on empirical algorithms and do not reliably represent human perception. As the interest of the analysts may differ from the mathematical definition of data properties, their expectations may not be matched by current metrics. Therefore, not all structures of interest for human observers may be found. To address this problem, we propose a new visual quality metric that evaluates data visualizations according to human perception. In this paper, we focus on the task of rating the correlation visualized by scatterplots using deep learning. Thereby, we are able to identify scatterplots that do not match the mathematical concept of correlation but nonetheless are of interest to analysts.

Previous work on Convolutional Neural Networks (CNNs) has shown their ability to match human perception when trained on appropriate data sets. This was demonstrated on various tasks like rating the similarity of natural images [PCMS18] or scatterplots [MTW*19]. Instead of computing a similarity score, we propose to use CNNs to learn a quality metric that rates the perceived correlation in scatterplots according to human perception. The graphical representation of the analyzed data property can be very diverse in scatterplots (e.g. linear vs. non-linear correlation), therefore rating the similarity alone is not sufficient. Despite this, humans can rate different correlation forms in scatterplots as similar, independently of their visual disparity. We perform an experiment using a pairwise comparison design to empirically gather a data set containing the perceived correlation in scatterplots. During the experiment, participants were instructed to select the scatterplot showing the strongest correlation between its dimensions.

Overall, our new data set consists of more than $15,000$ annotated pairs of scatterplots which we use to train two different CNN architectures. One uses images as input and is based on the popular ResNet [HZRS16] architecture which achieved outstanding performance for many image processing tasks. The other architecture is based on PointNet++ [QYSG17] and uses data vectors as input. By adapting PointNet++, our metric becomes independent of visual representations. The output of both networks is a score indicating the perceived correlation as captured by our experiment. The main contributions of our work are:

- A novel CNN-based visual quality metric to rate correlation in scatterplots according to human perception.
- A new data set consisting of over $15,000$ pairs of scatterplots annotated with human judgments on the perceived correlation.
- Several comparisons of our two CNN architectures accepting either images or data vectors.

© 2019 The Author(s)
Eurographics Proceedings © 2019 The Eurographics Association.

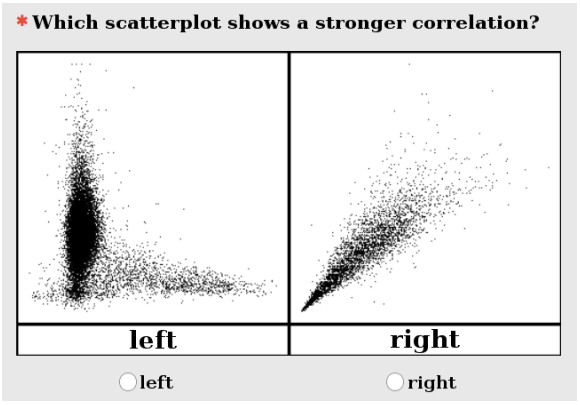


Figure 1: *Exemplar trial: During the experiment, participants were asked to choose which of the two presented scatterplots showed a stronger correlation.*

2. Related Work

In our work, we propose to learn a new visual quality metric respecting human perception using deep learning. Thus, we shortly outline perception related classical approaches and learning techniques that incorporate perception for related tasks.

Human Perception on Scatterplots

There has been extensive previous work on describing and detecting properties and patterns in visualizations. While many approaches were presented to automatically select or rank visualizations [WAG05, TAE*11, SSB*16], the majority of these works does not focus on modeling human perception. In order to understand and model human perception on scatterplots, different experiments have been conducted [RB10, PKF*16, Ren17, YHR*19]. One example of such an experiment is presented in the work of Rensink and Baldridge which uses classical psychological measures like the just-noticeable difference to investigate whether the perception of scatterplots can be modeled by mathematical concepts [RB10]. Alternatively, Pandey et al. performed an experiment where participants were asked to group scatterplots and to find names for the proposed groups [PKF*16]. The authors found that the participants' grouping criteria differed from previously defined categories [WAG05]. They suggested to use the new concepts to describe scatterplots with respect to human perception. Additional experiments showed that entropy is an important factor for participants [Ren17] and that only a small number of visual features are used when rating the correlation in scatterplots [YHR*19].
Building upon the experimental insights [RB10], a system to rank visualizations using Weber's law was introduced [HYFC14]. In contrast to manually designing a descriptive model, we train a CNN on human ratings which implicitly learns to rate the correlation of scatterplots according to human perception.

In order to investigate human perception on dimension reduction techniques, Sedlmair et al. performed an experiment with two visual analytics experts [STMT12]. During the study, the experts rated scatterplots based on their class separation property using a Likert scale from 1 to 5. Later, they successfully used their data set to evaluate existing quality metrics for class separation in hindsight of human perception [SA15]. To obtain reliable results using CNNs, the training data needs to be as complete and representative as possible. Thus, learning normally requires a large quantity of data. To obtain this amount of annotated scatterplots, we conduct an experiment via an online survey. Therefore, we use a pairwise comparison instead of a Likert scale as direct numerical assignments can be inconsistent between participants without clearly defining the meaning of each possible value beforehand [EF00].

Learning to Mimic Human Perception

In the field of visual analytics, different learning-based techniques were introduced to generate perceptually-motivated visualizations for human analysts [DBH14, DD19, HBL*19, HGH*19]. The possibility to automatically choose good design parameters for visualization (like color, shape, or size) was realized based on learned kernels [DBH14], while other works focused on predicting the design choices of visual analysts [HBL*19]. Instead of predicting visualization parameters, Dibia and Demiralp directly generated efficient visualizations from a data set using a recurrent neural network [DD19]. In order to enable easier comparisons, Hu et al. introduced a data set collection for visualization generation [HGH*19]. In contrast to these works, we do not aim to find perceptual-motivated visual representations but to rate specific data properties according to human perception.

Training neural networks to match human perception has been proposed for different tasks in the field of quality assessment for images like predicting visible distortions [BLBS17] or rating image similarity [ZIE*18]. Talebih and Milanfar trained a CNN for image enhancement by not only using ground truth information but also incorporating human ratings [TM18]. Therefore, human annotations on the aesthetics of images were included in their training data alongside undistorted ground truth information. Prashnani et al. introduced a network that learns image similarity through an experiment with pairwise comparisons [PCMS18]. While their approach aims to find the perceptual error with regard to a reference image, we want to obtain a score for a scatterplot without a reference. Moreover, these CNN-based methods only use input images, whereas we also designed a network architecture which operates directly on data vectors.

Similar to our work, Albuquerque et al. introduced a perception-based quality metric [AEM11]. Instead of deep learning, they used a machine learning approach based on non-metric multidimensional scaling [AWC*07]. While they can mimic human perception, their perceptual embedding has been trained only on a small set of correlation patterns. Furthermore, as their metric compares new scatterplot images to the learned examples using Principal Component Analysis, it can be only used to reliably rank scatterplots that are very similar to the learned samples. Recently, a deep learning approach has been proposed to rank scatterplots according to their similarity [MTW*19]. In their ScatterNet an experiment is conducted using triplets of scatterplot images where one scatterplot acts as a reference. The participant decides which of the other two plots is most similar to the reference. Afterwards, a CNN is trained to retrieve the most similar scatterplots based on a query image.

© 2019 The Author(s)
Eurographics Proceedings © 2019 The Eurographics Association.

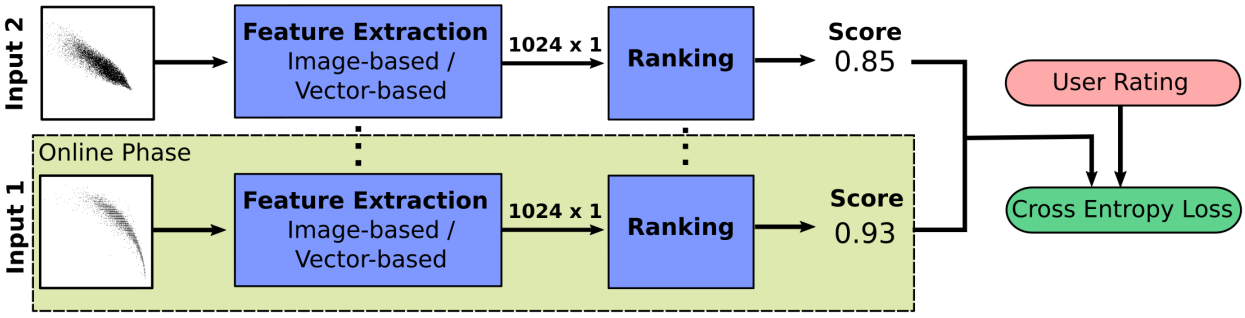


Figure 2: *An overview of our network architecture and training procedure. During online phase (indicated by the dashed box) our network calculates a score on a single input. Our architecture consists of two networks. First, we use a feature extraction network (either Image-based or Vector-based) to obtain a* 1024×1 *feature vector. We feed this feature vector into a ranking network which computes our final correlation score in* $[0, 1]$*. For training we compute the scores of annotated scatterplot pairs using two instances of our architecture with shared weights. For the loss computation, we apply a probabilistic-based cross entropy loss according to the user rating and computed scores.*

While the network can be used to find similar scatterplots, it cannot directly rate data properties for a defined user task. In contrast to ScatterNet, our network learns to rate scatterplots for a given task on a numerical scale. This way, our approach can match human ratings for diverse graphical representation like linear and non-linear correlation which cannot be retrieved based on similarity only. Additionally, as ScatterNet is based on input images, the quality suffers if the representation of scatterplots (e.g. the chosen point size) is changed. To avoid this problem we successfully train a network directly on the data vectors which achieves the same quality as our Image-based network variation.

3. Data Set

We create a data set with annotations collected through an experiment and use it to train a CNN to rate the correlation in scatterplots with respect to human perception. We use scatterplot pairs and ask the participants to decide which of the two shows a stronger correlation. The annotated scatterplot pairs enable our network to learn a meaningful scale for the rating implicitly. Thus, we avoid to manually normalize the ratings which is often seen for Likert scales, as the rating behaviour on a numerical scale can differ between participants. [EF00]. While previous work learns to rate similarity between two images [PCMS18, MTW*19], our approach allows to rate a specific perceptual characteristic from a single input. In theory, our metric can learn to rate any other data property by repeating the experiment for the desired property and retraining on the new data set.

To create the data set, we first choose 26 real data sets from the UCI machine learning repository [DG17] and 3 contained in the *ggplot2* R package [Wic16] which include different point distributions and forms of linear and non-linear correlation. We first split the data sets into a train and test set to make sure that data sets and scatterplots are mutually exclusive between both sets. We use 15 data sets during training: Abalone, AirQuality, Anuran Calls MFCCs, Avila, Cloud, Credit Card Clients, Energy, HTRU2, KEGG Metabolic, Relation Network, MAGIC Gamma Telescope, Wine Quality, Physicochemical Properties of Protein Tertiary Structure, Parkinsons and Telemonitoring. The test set uses 14 data sets: Baseball, Crwd Sourced, Diamonds, Electrical Grid Stability, Facebook, Google, iBeacon, Leaf, Occupancy, Online News Popularity, Page Blocks, Postures, Sensorless Drive Diagnosis and liver. We then render scatterplots for all possible dimension pairs for both sets at a size of 256x256 pixels. Afterwards, we manually choose subsets of 500 plots for training and 50 for testing which represent the different forms and patterns of correlation well. Additionally, we remove scatterplots containing categorical data or classes. We obtain enough data making it unnecessary to enlarge our data set with synthetic data. Finally, we randomly generate $15,000$ pairs for training and 200 pairs for testing and evaluations. Even though, the test set is considerably smaller than the training set, we used a similar amount of data sets and preserved diverse data properties.

Experiment

We created a web-based experiment and invited participants via email. Invitation emails were sent to different university mailing lists aimed at students of computer science and related fields. Anyone who obtained the link to the survey was able to participate.

As a first step, when participants accessed the experiment site, they were guided through a training section. This was designed to convey all necessary information for the experiment and familiarize the participants both with the task and the concept of correlation. In this training, the participants were presented with examples of linear and non-linear correlation as well as with examples where the data dimensions were not correlated at all. They were informed to rate negative and positive correlations equally, and to base their judgments on their impression rather than on mathematical definitions. After the participants read all the information, they were presented with five training trials that were not considered for the analysis. These exemplar trials showed clear cases of strong and weak correlation to test whether the participants understood the explained concepts. Thus, the participants could only continue with the real experiment by selecting the right answers.

The experiment consisted on a forced-choice task where the par-

© 2019 The Author(s)
Eurographics Proceedings © 2019 The Eurographics Association.

Accuracy	Pearson	RVM	Image-based	Vector-based
Train	-	-	**92.07%**	90.60%
Test	52.50%	69.00%	80.00%	**83.00%**

Table 1: *Accuracy values for the Pearson correlation, the RVM [AEM11] and both of our networks on our data set.*

ticipants needed to select the plot with the highest correlation based on the criteria previously explained to them. Participants were presented with a screen where two scatterplots were displayed side by side under the sentence "Which scatterplot shows a stronger correlation". They were able to introduce their answers by checking one of the two radio buttons available (see Fig. 1). For each trial of the experiment, a randomly selected scatterplot pair was presented and the participant needed to make their decision before the next pair was displayed.

The number of trials of the experiment was left to the decision of the participant and they were given the opportunity to conclude the experiment at their own convenience. The average number of trials fulfilled by a participant was 150. Overall, we collected at least three human annotations for each scatterplot pair. Based on the results gathered through this experiment, in order to build our data set, we labelled each pair based on the majority decision and stored the images as well as data vectors for all scatterplots.

4. Network

Our visual quality metric to rate the correlation in scatterplots according to human perception consists of two network parts: First, a *feature extraction network* computes a 1024×1 feature vector representing the scatterplot. Afterwards, a *ranking network* calculates a perceptual score based on the extracted features. An overview of the network architecture is given in Fig. 2.

For the *feature extraction network* we evaluate two different variations. As CNNs have shown impressive results for feature extraction from images for various tasks like classification and segmentation, we train an **Image-based feature extractor** for scatterplot images. This way the network is trained on the same images that were shown to the participants during the experiment and obtains no additional information. However, using scatterplot images as input also has a disadvantage: Scatterplots need to be rendered beforehand and use the same representation (e.g. point size) that was used for the training data. Otherwise, the rendering differences might reduce the performance of the metric [MTW*19]. To avoid this restrictions, we also train a second network using **Vector-based feature extraction**. This architecture directly processes the data vectors and is therefore independent of scatterplot images and their visual representation. The disadvantage is that the vector data may contain information the participants did not perceive, like data points that overlap and were not visible in the rendered scatterplot.

Image-based feature extraction. The Image-based architecture adapts ResNet [HZRS16] as it has shown outstanding performance on the task of image classification. The Residual Units of this architecture make deep network architectures without performance degradation possible. This enables networks to learn abstract and representative features resulting in good classification performance. We adopt the standard ResNet18 architecture with some modifications. As scatterplot images have less variations than natural images, we reduce the feature vector to a size of 1024×1 by changing the output size of the forth residual block to 256, removing the last fully-connected layer and setting the kernel size of average pooling to 2.

Vector-based feature extraction. The Vector-based architecture adapts PointNet++ [QSMG17] for feature extraction. PointNet++ has been designed to use point clouds and therefore is independent of the point order. Originally, the efficiency of PointNet++ has been demonstrated for 3D point cloud segmentation and classification. We adjust the architecture by setting the input dimension to 2 as we consider each scatterplots as a 2D point cloud. As proposed by the original PointNet++, we either up- or downsample our point cloud to create a vector with a fixed length. We use an input length of 4096 data points.

Ranking Network. After feature extraction, we want the ranking network to learn a numerical score, even though our data set only contains binary correlation information for each pair. Therefore, we took inspiration from RankNet which computes a ranking by sorting items using a probabilistic cross entropy loss [BSR*05]. We use a simple three-layer structure after feature extraction that outputs a one-dimensional value as score. Specifically, we use three fully connected layers with LeakyReLUs. The ranking is performed exactly in the same way for both Image-based and Vector-based feature extraction. As we want to normalize the score of our quality metric between 0 and 1 we conclude the ranking network with a sigmoid function. Limiting the score with an upper and lower bound makes the rating directly interpretable and more predictable for unseen scatterplots.

4.1. Loss Function

We train our networks end-to-end to predict a score for one input scatterplot representing the correlation's strength. For this purpose we use an adapted version of the probabilistic cross entropy loss proposed in RankNet [BSR*05]:

$$L = -\widehat{P}_{ij} \log P_{ij} - (1 - \widehat{P}_{ij}) \log(1 - P_{ij}), \tag{1}$$

where $\widehat{P}_{ij}$ is the ground truth label which was derived from the experiment by majority decision:

$$\widehat{P}_{ij} = \begin{cases} 1 & i \text{ shows stronger correlation} \\ 0 & \text{else} \end{cases} \tag{2}$$

Accordingly, P_{ij} is the probability that scatterplot i is perceived to be more correlated than scatterplot j based on the scores S_i and S_j estimated by the network. However, as we restrict our scores to be in the range of [0, 1], it puts a hard limit on the probability P_{ij} and therefore restricts the size of the gradient signal. To counteract this problem we adapted the probability function with a multiplier σ to enlarge the possible probability range:

$$P_{ij} = \frac{\exp(\sigma * (S_i - S_j))}{1 + \exp(\sigma * (S_i - S_j))} \tag{3}$$

© 2019 The Author(s)
Eurographics Proceedings © 2019 The Eurographics Association.

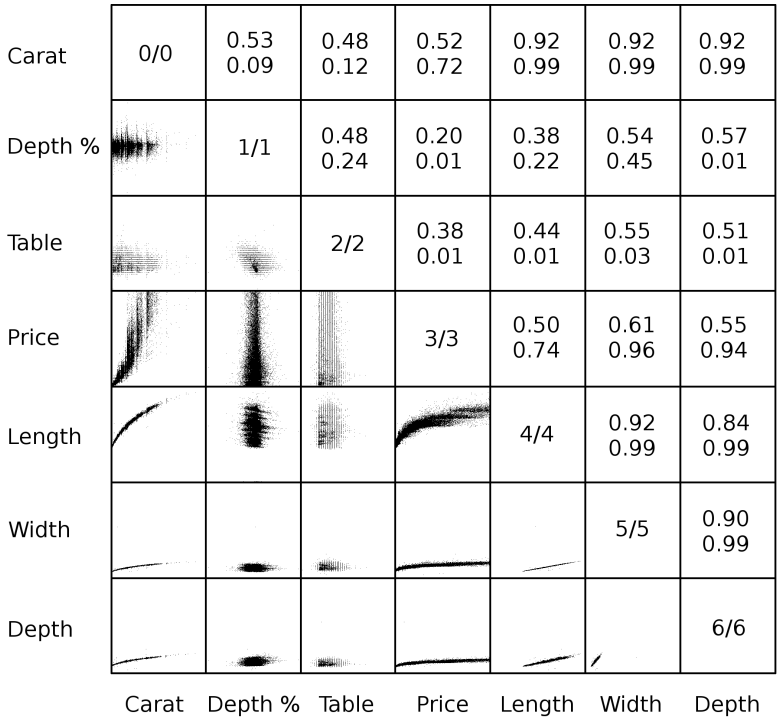


Figure 3: *Scatterplot matrices showing the scores computed by our Image-based metric (top) and Vector-based metric (bottom) for all pairs of dimensions of the diamond data set [Wic16].*

Training Details

During training the Image-based network, we use the random flips as data augmentation to prevent overfitting: We perform random flips in both horizontal and vertical directions, as we decided to treat positive and negative correlation equally. For image files, we avoid operations which would affect human perception. Especially cropping image regions may change the impression on the correlation of the scatterplots as data points can be cut off. As rotations might also require cropping of image edges, we exclude this operation as well. For the Vector-based architecture, we first perform data normalization to bring the point positions to a common value range. To increase the variability of the data vectors, we randomly shuffle the point order.

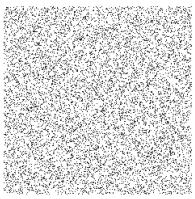
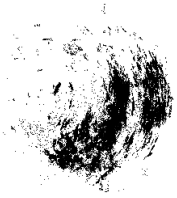
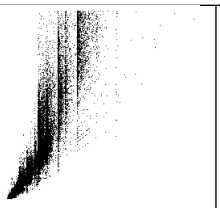
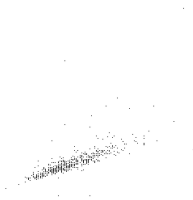
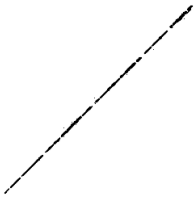
We use the following techniques and parameters: We set the multiplier $\sigma = 5$ from Eq.3 as this provided the best training results. We use SGD [GBC16] as a solver and set the momentum to 0.9. For both network architectures we use a batch size of 64 and a learning rate of 0.001. We decay the learning rate by multiplying it with 0.1 after every 10 epochs. While the loss basically converges after 15 epochs, we keep training for a total of 50 epochs. Training takes about 2.5 hours for the Image-based network and 9 hours for the Vector-based network on a *NVIDIA Titan Xp*. Our networks were implemented in *PyTorch* using *CUDA* support.

5. Evaluation

In this section, we evaluate how well our novel visual quality metrics match human perception and compare their results to two existing methods. We perform three different experiments. In the first experiment we compute the overall accuracy values on our data set for all metrics. In the second experiment, we show the differences between our novel metrics and present results for example scatterplots containing different correlation patterns. Lastly, we use a ranking example to further illustrate the behaviour of our metrics. We compare our metrics to the Pearson correlation which is based on the mathematical definition for linear correlation. As we do not distinguish between positive and negative correlation, we use absolute values for the Person correlation. Additionally, we compute the Rotating Variance Measure (RVM) [TAE*11] which was designed to rate both linear and non-linear correlation. This metric is based on the fact that correlated data is usually represented by fine structures in scatterplots. We chose the parameters of RVM as proposed by the original paper.

For our first test, we compare the accuracy of our metrics, Pearson and RVM on our data set. To check the accuracy of the different quality metrics, we compute scores for all scatterplots. Then, we check whether the order of the calculated scores matches the human annotation for each scatterplot pair. The resulting accuracy values are presented in Table 1. Both of our proposed network architectures achieve a high accuracy on the train set showing a successful training. As our metrics already saw all scatterplots in the

© 2019 The Author(s)
Eurographics Proceedings © 2019 The Eurographics Association.

					
Pearson	0.0024	0.5447	0.9216	0.7512	0.9999
RVM	0.0898	0.1478	0.3257	0.2445	1.0
Image-based	0.0012	0.2321	0.5215	0.6924	0.9999
Vector-based	0.0004	0.1183	0.7242	0.6462	0.9996

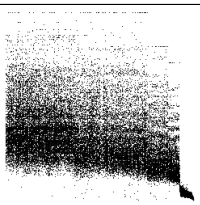
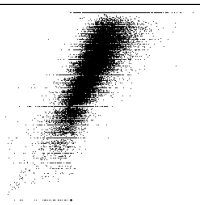
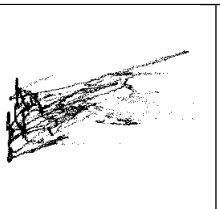
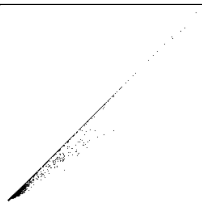
					
Pearson	0.4931	0.7278	0.478	0.7978	0.9949
RVM	0.3826	0.2658	0.2858	0.2114	0.2801
Image-based	0.0044	0.2359	0.4592	0.7203	0.9871
Vector-based	0.0019	0.4565	0.601	0.7648	0.9984

Table 2: *Two scatterplot sets ordered by human participants. From left to right, the scatterplots appear increasingly correlated. Our Image-based feature extraction performed particularly well in this experiment.*

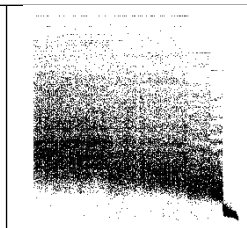
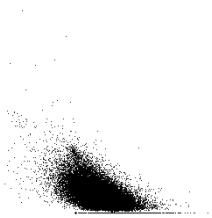
train set, we perform all following experiments only on data from the test set to ensure a fair comparison. Our metrics significantly outperform the Pearson Correlation baseline as well as the RVM achieving an accuracy of 80.00% and 83.00% for the unseen scatterplots in the test set. The accuracy is close to chance-level for the Pearson Correlation at 52.50%. The reason for this is most likely the inclusion of non-linear correlation which is not anticipated by the Pearson correlation. In contrast, the RVM achieves a reasonable accuracy of 69.00%. We present two examples of scatterplot pairs with scores for both of our metrics as well as the Pearson Correlation and RVM in Table 3. Even though the metrics do not match human perception, their values for both scatterplots in the lower example are very similar. In contrast to the compared metrics, both of our networks match the human judgment for these examples. Overall, these results indicate that both of our Learning-based techniques are in better alignment with human annotations.

Next, we investigate the differences between our Image-based and Vector-based architectures. Fig. 3 shows a comparison of scores on the diamond data set [Wic16]. The diamond data set consists of the following dimensions: Carat, Depth Percentage, Table, Price, Length, Width and Depth. We visualize the scores in a scatterplot matrix which enables us to detect interesting dependencies between data dimensions. As a diamonds carat is defined by its weight, the carat value is correlated to the length, width and depth of diamonds. Additionally, our metrics show that the length, width and depth of diamonds correlate to each other. This indicates that diamonds are usually cut while preserving an equal ratio between its sizes. While both of the metrics produce reasonable scores, the Image-based metric outputs scores distributed over the full value range while the Vector-based architecture scores are closer to the upper and lower bound.

To gain a better understanding of our metrics, we performed an analysis of their scores in our test set. The average scores are 0.5095 for the Image-based architecture and 0.4412 for the Vector-based architecture with a standard deviation (STD) of 0.3148 and 0.3415. The STD of the Vector-based metric is 8.48% larger, indicating that the scores of this metric are generally more extreme than the ones from the Image-based metric. Additionally, the average difference between scores is 0.1847. The differences between both metrics likely result from a different amount of information in the input data of both architectures. Especially, if several data points are in close proximity, they cover each other and are therefore not visible for the Image-based network. In contrast, the Vector-based network receives information based on the overall point distribution. Table 4 shows an example where the score of the Vector-based metric does not match the human rating from our experiment and also strongly differs from the Image-based variation. The right scatterplot was judged to have a stronger correlation, however, the Vector-based metric differs from this judgment. As can be seen in the lower left corner of the right image, there is a cluster of outliers. While our participants as well as the Image-based network only register a small amount of outliers, the actual number is considerably higher. As the actual number of outliers is not registered by participants and the Image-based network, they are still able to detect the strong correlation along the center diagonal. In contrast, this Vector-based and classical methods are unable to represent the interesting data distribution in their score.

In terms of runtime, the Image-based metric rates an input im-

© 2019 The Author(s)
Eurographics Proceedings © 2019 The Eurographics Association.

		
Pearson	**0.4931**	0.472
RVM	**0.3826**	0.2918
Image-based	0.0044	**0.4353**
Vector-based	0.0019	**0.1473**

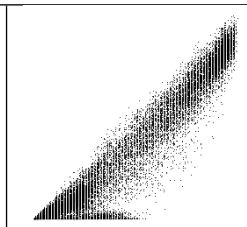
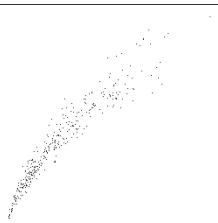
		
Pearson	**0.958**	0.9548
RVM	**0.2967**	0.2658
Image-based	0.4919	**0.7224**
Vector-based	0.8479	**0.9708**

Table 3: *Examples of scatterplot pairs where the right one was perceived to show stronger correlation. In contrast to our metrics, Pearson correlation and the RVM do not match the perception of the participants.*

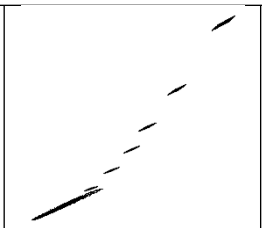
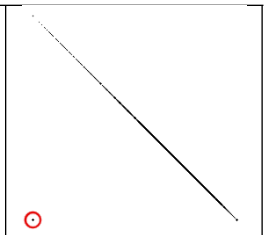
		
Pearson	**0.9863**	0.531
RVM	**0.7687**	0.5951
Image-based	0.9937	**0.9999**
Vector-based	**0.9982**	0.4725

Table 4: *For this scatterplot pair, the human participants perceived the right one to display a stronger correlation, which is not represented by our Vector-based quality metric. The reason is the outlier cluster in the left bottom of the right plot (marked with a red circle). The score for this scatterplot is also not in union with human perception for both classical metrics.*

age in 7.85ms whereas the Vector-based approach requires roughly 282ms. Overall, our Image-based and Vector-based architectures achieve similar results which shows that CNNs can learn to match human judgment based not only on images but also on data vectors. This is a valuable insight, as it enables training and testing without the need to render scatterplots, making it independent of the scatterplot's representation parameters like point size and other rendering effects.

We further illustrate the behaviour of our visual quality metrics using a sorting example. We prepared two subsets of five scatterplots and asked participants to sort them from weak to strong correlation. We recruited five participants who are familiar with visual analytics and asked them to sort the plots independently. We then computed scores for all scatterplots using our networks as well as the classical metrics and compared their results to the order proposed by the human participants. The results for this experiment are given in Table 2. The image order is based on the majority decision of the participants. For the upper row, all participants agreed on the displayed order, while one participant switched the second and third scatterplot for the lower row. From this experiment, we see that the classical metrics do not always align well to human ordering. Overall, our Image-based quality metric performed better than the Vector-based metric in this test. One example of this can be seen for the second and third image in the upper row where the order was only preserved by our Image-based metric. In contrast, the lower row was correctly sorted by both of our metrics. This example suggest that our metrics could also be used to compute rankings based on human perception.

Overall, our experiments indicate that our network architectures are able to represent human perception better than previous visual quality metrics when rating the correlation in scatterplots.

6. Conclusion

In this paper, we presented a novel visual quality metric based on human perception to rate the correlation in scatterplots. Our metric utilizes CNNs which are trained on a novel data set consisting of more than 15,000 scatterplot pairs with human annotations. To create the data set we conducted an experiment and obtained human judgments on which of two scatterplots were perceived as having stronger dimensional correlation. We trained and evaluated two CNNs: one based on images and the other based on data-vectors. Our results indicate that both versions improve the performance to predict human judgment compared to traditional correlation metrics.

There are some promising lines of work that we will like to address in the future, like adjusting the Vector-based network to enable the processing of variable input length. Another interesting extension is investigating how well the presented networks can be adapted for different criteria like class separation or other visualization techniques like parallel coordinates. Finally, it might be worthwhile to align both metrics' score more closely by matching the information contained in the image and vector files.

7. Acknowledgements

The authors gratefully acknowledge funding by the L3S Research Center, Hanover, Germany. We thank our South African colleagues for the inspiring discussions during the Workshop on Advances in Knowledge Engineering, Reasoning and Sensemaking (WAKERS 2019) at Stellenbosch.

© 2019 The Author(s)
Eurographics Proceedings © 2019 The Eurographics Association.

References

[AEM11] ALBUQUERQUE G., EISEMANN M., MAGNOR M.: Perception-based visual quality measures. In *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)* (Oct 2011), pp. 13–20. doi:10.1109/VAST.2011.6102437. 2, 4

[AWC*07] AGARWAL S., WILLS J., CAYTON L., LANCKRIET G., KRIEGMAN D., BELONGIE S.: Generalized non-metric multidimensional scaling. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics* (San Juan, Puerto Rico, 21–24 Mar 2007), Meila M., Shen X., (Eds.), vol. 2 of *Proceedings of Machine Learning Research*, PMLR, pp. 11–18. 2

[BLBS17] BERARDINO A., LAPARRA V., BALLÉ J., SIMONCELLI E.: Eigen-distortions of hierarchical representations. In *Advances in Neural Information Processing Systems 30*, Guyon I., Luxburg U. V., Bengio S., Wallach H., Fergus R., Vishwanathan S., Garnett R., (Eds.). Curran Associates, Inc., 2017, pp. 3530–3539. 2

[BSR*05] BURGES C., SHAKED T., RENSHAW E., LAZIER A., DEEDS M., HAMILTON N., HULLENDER G. N.: Learning to rank using gradient descent. In *Proceedings of the 22Nd International Conference on Machine Learning* (New York, NY, USA, 2005), ICML '05, ACM, pp. 89–96. doi:10.1145/1102351.1102363. 4

[DBH14] DEMIRALP Ç., BERNSTEIN M. S., HEER J.: Learning perceptual kernels for visualization design. *IEEE Transactions on Visualization and Computer Graphics 20*, 12 (Dec 2014), 1933–1942. doi:10.1109/TVCG.2014.2346978. 2

[DD19] DIBIA V., DEMIRALP Ç.: Data2vis: Automatic generation of data visualizations using sequence to sequence recurrent neural networks. *IEEE Computer Graphics and Applications* (2019), 1–1. doi:10.1109/MCG.2019.2924636. 2

[DG17] DUA D., GRAFF C.: UCI machine learning repository, 2017. URL: http://archive.ics.uci.edu/ml. 3

[EF00] ELLERMEIER W., FAULHAMMER G.: Empirical evaluation of axioms fundamental to stevens's ratio-scaling approach: I. loudness production. *Perception & Psychophysics 62*, 8 (Dec 2000), 1505–1511. doi:10.3758/BF03212151. 2, 3

[GBC16] GOODFELLOW I., BENGIO Y., COURVILLE A.: *Deep learning*. MIT press, 2016. 5

[HBL*19] HU K., BAKKER M. A., LI S., KRASKA T., HIDALGO C.: Vizml: A machine learning approach to visualization recommendation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2019), CHI '19, ACM, pp. 128:1–128:12. doi:10.1145/3290605.3300358. 2

[HGH*19] HU K., GAIKWAD S. N. S., HULSEBOS M., BAKKER M. A., ZGRAGGEN E., HIDALGO C., KRASKA T., LI G., SATYANARAYAN A., DEMIRALP C.: Viznet: Towards a large-scale visualization learning and benchmarking repository. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2019), CHI '19, ACM, pp. 662:1–662:12. doi:10.1145/3290605.3300892. 2

[HYFC14] HARRISON L., YANG F., FRANCONERI S., CHANG R.: Ranking visualizations of correlation using weber's law. *IEEE Transactions on Visualization and Computer Graphics 20*, 12 (Dec 2014), 1943–1952. doi:10.1109/TVCG.2014.2346979. 2

[HZRS16] HE K., ZHANG X., REN S., SUN J.: Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2016), pp. 770–778. doi:10.1109/CVPR.2016.90. 1, 4

[MTW*19] MA Y., TUNG A. K. H., WANG W., GAO X., PAN Z., CHEN W.: Scatternet: A deep subjective similarity model for visual analysis of scatterplots. *IEEE Transactions on Visualization and Computer Graphics* (2019), 1–1. doi:10.1109/TVCG.2018.2875702. 1, 2, 3, 4

[PCMS18] PRASHNANI E., CAI H., MOSTOFI Y., SEN P.: Pieapp: Perceptual image-error assessment through pairwise preference. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (June 2018), pp. 1808–1817. doi:10.1109/CVPR.2018.00194. 1, 2, 3

[PKF*16] PANDEY A. V., KRAUSE J., FELIX C., BOY J., BERTINI E.: Towards understanding human similarity perception in the analysis of large sets of scatter plots. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2016), CHI '16, ACM, pp. 3659–3669. doi:10.1145/2858036.2858155. 2

[QSMG17] QI C. R., SU H., MO K., GUIBAS L. J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (July 2017), pp. 77–85. doi:10.1109/CVPR.2017.16. 4

[QYSG17] QI C. R., YI L., SU H., GUIBAS L. J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (USA, 2017), NIPS'17, Curran Associates Inc., pp. 5105–5114. 1

[RB10] RENSINK R. A., BALDRIDGE G.: The perception of correlation in scatterplots. In *Proceedings of the 12th Eurographics / IEEE - VGTC Conference on Visualization* (Chichester, UK, 2010), EuroVis'10, The Eurographs Association & John Wiley & Sons, Ltd., pp. 1203–1210. doi:10.1111/j.1467-8659.2009.01694.x. 2

[Ren17] RENSINK R. A.: The nature of correlation perception in scatterplots. *Psychonomic Bulletin & Review 24*, 3 (Jun 2017), 776–797. doi:10.3758/s13423-016-1174-7. 2

[SA15] SEDLMAIR M., AUPETIT M.: Data-driven evaluation of visual quality measures. *Computer Graphics Forum 34*, 3 (June 2015), 201–210. doi:10.1111/cgf.12632. 2

[SSB*16] SHAO L., SCHLEICHER T., BEHRISCH M., SCHRECK T., SIPIRAN I., KEIM D. A.: Guiding the exploration of scatter plot data using motif-based interest measures. *Journal of Visual Languages & Computing 36*, C (Oct. 2016), 1–12. doi:10.1016/j.jvlc.2016.07.003. 2

[STMT12] SEDLMAIR M., TATU A., MUNZNER T., TORY M.: A taxonomy of visual cluster separation factors. *Computer Graphics Forum 31*, 3pt4 (June 2012), 1335–1344. doi:10.1111/j.1467-8659.2012.03125.x. 2

[TAE*11] TATU A., ALBUQUERQUE G., EISEMANN M., BAK P., THEISEL H., MAGNOR M., KEIM D.: Automated analytical methods to support visual exploration of high-dimensional data. *IEEE Transactions on Visualization and Computer Graphics 17*, 5 (May 2011), 584–597. doi:10.1109/TVCG.2010.242. 2, 5

[TM18] TALEBI H., MILANFAR P.: Learned perceptual image enhancement. In *2018 IEEE International Conference on Computational Photography (ICCP)* (May 2018), pp. 1–13. doi:10.1109/ICCPHOT.2018.8368474. 2

[WAG05] WILKINSON L., ANAND A., GROSSMAN R.: Graph-theoretic scagnostics. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.* (Oct 2005), pp. 157–164. doi:10.1109/INFVIS.2005.1532142. 2

[Wic16] WICKHAM H.: *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. URL: https://ggplot2.tidyverse.org. 3, 5, 6

[YHR*19] YANG F., HARRISON L. T., RENSINK R. A., FRANCONERI S. L., CHANG R.: Correlation judgment and visualization features: A comparative study. *IEEE Transactions on Visualization and Computer Graphics 25*, 3 (March 2019), 1474–1488. doi:10.1109/TVCG.2018.2810918. 2

[ZIE*18] ZHANG R., ISOLA P., EFROS A. A., SHECHTMAN E., WANG O.: The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 586–595. 2

© 2019 The Author(s)
Eurographics Proceedings © 2019 The Eurographics Association.