Statistical Analysis of a Qualitative Evaluation on Feature Lines

Alexandra Baer, Kai Lawonn, Patrick Saalfeld, Bernhard Preim

Department for Simulation and Graphics, Otto-von-Guericke University Magdeburg
alexandra.baer@ovgu.de

Abstract. In this paper, we statistically analyze the results of a qualitative evaluation with 129 participants of the most commonly used feature line techniques on artificial and anatomical structures obtained from patient-specific data. For this analysis, we tested for significant differences to verify the results and the evaluation. We applied the Shapiro-Wilk test for normality, the Friedmann test to validate significant differences, and the Wilcoxon signed-rank test to compare paired samples. The results are used to give recommendations for which kind of surface which feature line technique is most appropriate.

1 Introduction

Illustrations in anatomical atlases are mostly depicted in a simplified representation to enhance essential anatomical structures. This means, only important information are illustrated to support the physician or the doctor-to-be to focus on a specific region. Traditional illustrations are hand-made which is therefore, not available on patient-specific data. Feature lines are a family of illustrative visualization techniques that can be applied on arbitrary surfaces and thus on patient-specific data. They are used to give a simplified representation of an object. For this simplification, the object is illustrated by using lines, which are placed at the most salient regions such that the object's shape can be perceived without additional shading. This abstraction of surfaces is important and has a high potential for depicting pathologies, anatomical and risk structures required for surgery planning [1] and for intraoperative visualizations [2]. Moreover, abstract illustrations of patient-specific data and therapy options support an individual patient documentation. The most commonly used six feature line techniques are described in Section 2. To assess the quality of such feature lines, evaluations are important. Previous evaluations only compare a subset of the feature line methods.

This work is an extension of [3] where we qualitatively compared all different feature line techniques on three anatomical and three artificial surfaces. This first evaluation was about a ranking of the different feature line techniques according to aesthetic and realistic depiction. The participants should order the methods and rank them from place 1 to 6, with 1 being the most realistic or aesthetic technique. The rank of the feature line techniques is determined by using the

Schulze method. This method is used to determine a final rank ordering, but no significance is tested. Therefore, a winner is determined, but it is not clear if this result is statistically relevant. The contribution of this paper is a statistical analysis of the feature line techniques for significant differences. For this analysis, we use different statistical tests and analyze the results on organic surfaces as they occur in biology and medicine. Afterwards, we give recommendations on what kind of surface is depicted well with which feature line technique. This guideline can be used to choose an appropriate feature line technique for an anatomical structure.

2 Materials and Methods

2.1 Feature Line Techniques

Interrante et al. [4] introduced *ridges and valley lines* (RV) to illustrate salient regions. This method is curvature-based and therefore view-independent. Thus, DeCarlo et al. [5] presented a view-dependent approach: *suggestive contours* (SC). SC are an extension of the conventional contour definition, but fail in depicting convex structures. Therefore, Judd et al. [6] combined the advantages of RV and SC and presented *apparent ridges* (AR). This method uses a view-dependent curvature term and applies the RV definition to illustrate the shape. Xi et al. [7] introduced *photic extremum lines* (PLs). This technique determines the maximum of the variance of illumination. The user can add additional spotlights to influence the result. Kolomenkin et al. [8] presented *demarcating curves* (DC), a view-independent approach. This method is best suited to enhance furrows. Zhang et al. [9] introduced *Laplacian lines* (LL). LL extend the Laplacian-of-Gaussian for images to 3D surfaces to illustrate the shape.

2.2 Evaluation and Quantitative Analysis

The evaluation was conducted with 129 participants. 68 men with an average age of 30.53 years and 61 woman with an average of 30.92 years participated

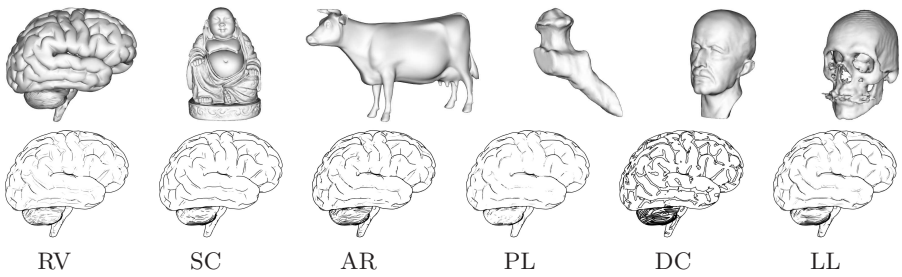

Fig. 1. In the first row, the different models for the evaluation are listed. In the second row, the different feature line techniques are illustrated with the brain model [3].

in this line drawing technique evaluation. The participants had a very broad spectrum of educational background. The tasks to assess the techniques' quality comprise an assessment of aesthetics, of realism and preference. For all tasks, the participants saw an original model, which was shaded using Gouraud shading. Furthermore, the feature line techniques were applied and the participants had to order them according to aesthetics and realism regarding the shaded model, see Fig. 1 for an overview of all models and techniques. The last task was about choosing a favorite technique. Six models were visualized with each of the six line drawing techniques. Three groups of 43 participants had to rate the six line drawing techniques. Each model was rated by 43 participants. The line drawing techniques are quantitatively analyzed for each model. Thus, we have 43 rank results for each model and technique. We analyzed the rating results of realism and of aesthetics in two steps; model-independent and model-dependent. Initially, the Shapiro-Wilk test was applied to analyze the normal distribution. Furthermore, we applied the Friedmann test and the non-parametric Wilcoxon signed-rank test for a pairwise comparison.

3 Results

Not normally distributed rank results were confirmed with a significant difference of $p \leq .05$ for realism and of $p \leq .01$ for aesthetics compared to a normal distribution. The non-parametric Friedmann test confirmed significant differences with $p \leq .01$ for both assessment tasks. The Wilcoxon sign-rank test confirmed significant difference for the first analysis step (model-independent) for realism with $p \leq .05$ between the techniques RV - SC, SC - LL, SC - PL and AR - LL and with $p \leq .01$ between RV - AR, RV - LL, AR - PL and PL - LL.

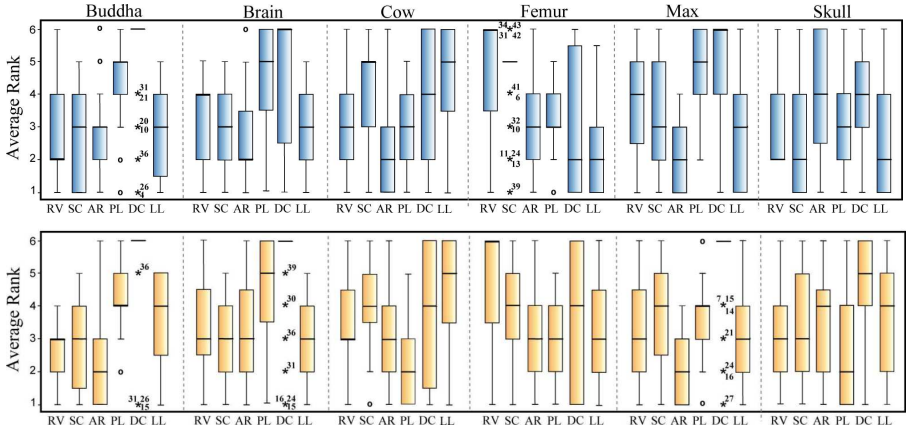

Fig. 2. The model-dependent results for realism in blue and for aesthetics in yellow visualized as boxplots. Detected outliers are displayed as an asterisk with the according sample number. Circles are potential outliers.

A significant difference for aesthetics with $p \leq .01$ was confirmed for AR and DC pairwise compared with any other technique. In detail, the best and worst ranked technique are significantly different compared to the other techniques. All model-dependent results for realism and for aesthetics are illustrated in Fig. 2. The boxplots show the results for each task and model and the median result is included. The Wilcoxon sign-rank test pairwise compared the techniques for each model and task. The results are shown in Fig. 3. Each table contains the results for realism (upper blue triangular matrix) and for aesthetics (lower yellow triangular matrix) for one model. The Wilcoxon sing-rank test confirmed significant differences for realism with $p \leq .01$ between each technique pair for the Buddha and the Max model, illustrated in Fig. 3 in the upper triangular matrix with green check marks. Additionally, significant differences with $p \leq .01$ and $p \leq .05$ were confirmed for the other models and tasks, as shown in Fig. 3. A detailed result discussion follows in Section 4.

4 Discussion

We divide this section into the discussion of the realism and aesthetics results. Finally, we analyze the results of the preferred techniques. In the following, a statistically significant difference is in short significant or a significant difference.

4.1 Realism Results

In most cases, starting from the resulting order of the feature line techniques, neighbored ranks are not significantly different. For example, regarding the result of the cow model the resulting order is: AR, PL, RV, DC, SC, LL. In this case, the statistical difference of the first place (AR) and the second place (PL) is not significant but AR compared to RV. Analyzing the first three ranks, AR, SC, and LL are dominant. In cases where AR is on the first rank (brain, Buddha, cow, Max), it is not significant with the second place only on the brain and cow model, on the Max model it is significant on the first place, and on the Buddha model it is even not significant with the fourth rank. Thus, the AR method is mostly ranked on the first place or the second place. The SC technique is placed once on the first place (skull), twice on the second place (brain, Buddha), and once on the third place (Max). On the skull model, SC has no significant difference with the second place, and on the Buddha model it has no significant difference with the first, third, and fourth place. Regarding the second and the third place of SC, it is not significant with the third and fourth place. The first rank of LL for the femur model is a significant result. Furthermore, LL is placed on rank 2 twice and once on rank 3. As mentioned, mostly there is no significant difference between neighbored ranks, but mostly between two ranks, i.e., rank 2 and 4. In summary, although the realism results are not uniquely defined, there is a tendency to the methods AR, SC, and LL. Counting the number of how often a technique was placed on a rank between 1-3, we have AR: 5, SC: 4, LL: 4, PL: 2, RV: 2, DC: 1.

4.2 Aesthetics Results

The aesthetics results are similar to the realism results regarding the significance. Mostly, neighbored ranks are not statistically significant. The best result was obtained by AR. AR reaches twice the first place (Buddha, Max) with significant difference to the second rank; it reaches twice the second rank (cow, femur), whereas it has no significant difference even to the fourth rank. The LL technique reached twice the first place (brain, femur). On the brain model it has no significant difference compared to the second place, and on the femur model it has no significance even on the third place. The PL technique reached also twice the first place (cow, skull). On the cow model it is significantly better compared to the second place, and on the skull model it has no statistical significant difference to the third rank. The results for this task contain several detected outliers for the DC technique, as shown in Fig. 2 in the lower row. The Buddha, brain and Max results revealed up to seven outliers. According to the median and other DC results, we assume that the participants misunderstood rank 1 and thus ranked the technique vice versa. For this task, the techniques AR, PL, and LL were placed best. Counting the number of how often a technique was placed on a rank between 1-3, we have AR: 5, RV: 4, SC: 3, LL: 3, PL: 3, DC: 0.

4.3 Preferred Techniques

The participants should also choose their favorite technique. Independent of the underlying model, we list how often which feature line technique was chosen: AR: 65, LL: 58, SC: 49, PL: 33, RV: 27, DC: 26. Here, we see a tendency to the preferred techniques AR, LL, and SC.

Buddha	RV	SC	AR	PL	DC	LL
RV		-.308	-.564	✓	✓	-.496
SC	.324		-.283	✓	✓	-.828
AR	✓	✓		✓	✓	-.496
PL	✓	✓	✓		✓	✓
DC	✓	✓	✓	✓		✓
LL	✓	✓	✓	✓	✓	

Brain	RV	SC	AR	PL	DC	LL
RV		-1.554	✓	✓	✓	-.686
SC	✓		-.471	✓	✓	-1.128
AR	.266	.192		✓	✓	✓
PL	✓	✓	✓		-.734	✓
DC	✓	✓	✓	✓		✓
LL	✓	.122	✓	✓	✓	

Cow	RV	SC	AR	PL	DC	LL
RV		✓	✓	.165	.070	✓
SC	✓		✓	✓	.388	✓
AR	.080	✓		.129	✓	✓
PL	✓	✓	✓		✓	✓
DC	.311	.264	.063	✓		✓
LL	✓	✓	✓	✓	✓	

Femur	RV	SC	AR	PL	DC	LL
RV		.083	✓	✓	✓	✓
SC	✓		✓	✓	✓	✓
AR	✓	✓		.430	.435	✓
PL	✓	✓	.350		.344	✓
DC	✓	.318	.060	.063		✓
LL	✓	✓	.372	.415	✓	

Max	RV	SC	AR	PL	DC	LL
RV		.200	✓	✓	✓	✓
SC	.099		✓	✓	✓	.075
AR	✓	✓		✓	✓	✓
PL	.134	.299	✓		.162	✓
DC	✓	✓	✓	✓		✓
LL	.325	✓	✓	✓	✓	

Skull	RV	SC	AR	PL	DC	LL
RV		✓	.096	✓	.155	✓
SC	.237		✓	✓	✓	.295
AR	.165	.378		✓	.422	✓
PL	.141	.061	✓		✓	.100
DC	✓	✓	✓	✓		✓
LL	✓	.059	.154	✓	✓	

Fig. 3. Each table contains the realism (upper blue triangular matrix) and the aesthetics (lower yellow triangular matrix) results of the Wilcoxon sign-rank for one model. A green check mark confirms a significant difference between the corresponding two techniques. If no significant difference was found, the *z-score* value is listed. If the *z-score* value is bigger than 1.96 (ignoring the minus sign), then the Wilcoxon sign-rank test confirms a significant difference with $p < .05$.

4.4 Summary

The statistical analysis showed that mostly neighbored ranks are not significantly different, but analyzing the techniques that differ from more than two ranks, the difference is mostly significant. Thus, the evaluation gives reliable results. Moreover, further evaluations can be conducted similarly, even if the results of the line techniques are visually hard to distinguish, the participants mostly agreed with the different tasks. We would strongly recommend to use one of the three different feature line techniques AR, LL, or SC for illustrating anatomical structures. In general, SC would be more appropriate as this technique uses second-order derivatives only. Compared to the other methods, which have third-order derivatives, SC is less susceptible to noisy surfaces. However, SC cannot depict convex regions and thus, surfaces with sharp edges. Sharp edges, however, rarely exist in anatomical structures. Nevertheless, in this case AR and LL are recommended. Here, the performance of AR is lower than the performance of LL. The LL method is not recommended for users without experience, as this technique needs user-defined values for calculating the Laplace operator on the surface. This can result in a trial-and-error loop where the user tests different parameters until a satisfying result is yielded. In overall, we recommend to use SC, as this is more robust against noise. For a detailed analysis of patient-specific data an extended evaluation is recommended.

References

1. Tietjen C, Isenberg T, Preim B. Combining Silhouettes, Surface, and Volume Rendering for Surgery Education and Planning. In: IEEE/Eurographics Symposium on Visualization; 2005. p. 303–310.
2. Ritter F, Hansen C, Dicken V, Konrad O, Preim B, Peitgen HO. Real-Time Illustration of Vascular Structures. IEEE Trans Vis Comput Graph. 2006; p. 877–884.
3. Lawonn K, Baer A, Saalfeld P, Preim B. Comparative Evaluation of Feature Line Techniques for Shape Depiction. In: Proc. of Vision, Modeling and Visualization; 2014. p. 31–38.
4. Interrante V, Fuchs H, Pizer S. Enhancing Transparent Skin Surfaces with Ridge and Valley Lines. In: Proc. of IEEE Visualization; 1995. p. 52–59.
5. DeCarlo D, Finkelstein A, Rusinkiewicz S, Santella A. Suggestive Contours for Conveying Shape. Proc of ACM SIGGRAPH. 2003; p. 848–855.
6. Judd T, Durand F, Adelson E. Apparent ridges for line drawing. In: Proc. of ACM SIGGRAPH; 2007. p. 19–26.
7. Xie X, He Y, Tian F, Seah HS, Gu X, Qin H. An Effective Illustrative Visualization Framework Based on Photic Extremum Lines (PELs). IEEE Trans Vis Comput Graph. 2007; p. 1328–1335.
8. Kolomenkin M, Shimshoni I, Tal A. Demarcating curves for shape illustration. In: Proc. of ACM SIGGRAPH Asia; 2008. p. 157:1–157:9.
9. Zhang L, He Y, Xia J, Xie X, Chen W. Real-Time Shape Illustration Using Laplacian Lines. IEEE Trans Vis Comput Graph. 2011; p. 993–1006.