Quantitative Analysis of Saliency Models

F. P. Tasse[1] J. Kosinka[1,2] N. A. Dodgson[1,3]
[1]University of Cambridge [2]University of Groningen [3]Victoria University of Wellington

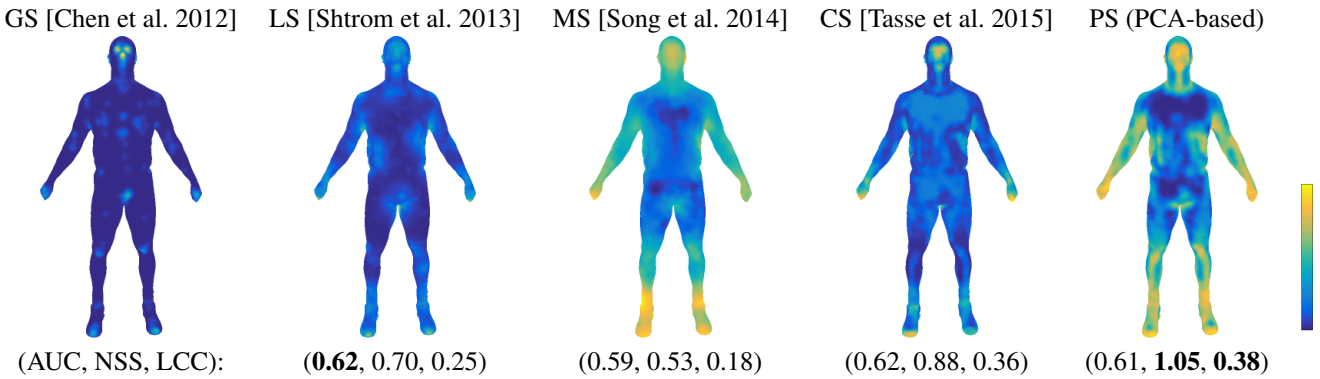

Figure 1: *Examples of saliency maps generated by selected saliency models. The scores underneath each map are evaluation metrics that indicate how well the map compares to ground-truth (GS). The metric for the top-performing saliency map is indicated in* ***bold****.*

Abstract

Previous saliency detection research required the reader to evaluate performance qualitatively, based on renderings of saliency maps on a few shapes. This qualitative approach meant it was unclear which saliency models were better, or how well they compared to human perception. This paper provides a quantitative evaluation framework that addresses this issue. In the first quantitative analysis of 3D computational saliency models, we evaluate four computational saliency models and two baseline models against ground-truth saliency collected in previous work.

Keywords: shape saliency, evaluation, fast point feature histogram

Concepts: •**Computing methodologies** → **Shape analysis;**

1 Introduction

We introduce three metrics for quantitative analysis and comparison of 3D saliency methods.

Despite recent interest in saliency detections of 3D surfaces [Liu et al. 2016], there are no evaluation metrics available, and no previous quantitative evaluation of 3D saliency models, making it impossible to objectively compare one method against another. Previous work evaluates results qualitatively, by 3D renderings of saliency maps. This approach does not objectively determine whether one saliency model is better than another. With the absence of an evaluation benchmark, new techniques cannot be compared to previous methods on a common dataset. To the best of our knowledge, no previous work performs a quantitative evaluation of computational models to ground truth. We propose such a quantitative evaluation, based on surface feature points acquired from users by Chen et al. [2012]. Inspired by Judd et al.'s benchmark of computational models to predict human eye fixations in images [Judd et al. 2012], we evaluate 4 saliency models, and 2 baseline models against ground truth 3D saliency on watertight meshes.

Benchmarks for 3D keypoint detection have been proposed [Dutagaci et al. 2012], as well as quantitative analysis that explain how mesh properties such as curvature are correlated with ground-truth saliency using information theory statistics [Chen et al. 2012]. However, none of these works compare saliency performance using raw saliency maps. For instance, metrics for keypoint detection will need to use thresholding parameters if applied to saliency maps. In contrast, there are well-established 2D image saliency benchmarks [Judd et al. 2012; Borji et al. 2013], which provide a framework for quantitative evaluation of different 2D saliency algorithms. Thus, we use available ground-truth saliency data [Chen et al. 2012] and present evaluation metrics, inspired by the 2D case, to compare saliency models. We select this particular dataset, which is derived from salient points selected by humans, because it is often used for qualitative evaluation [Song et al. 2014; Tasse et al. 2015].

Our contributions are three metrics for estimating the performance of saliency models, evaluated on six saliency models' ability to predict ground truth saliency, on 400 meshes from the SHREC'07 Watertight models track [Giorgi et al. 2007] (SHREC07).

2 3D Saliency methods

Early saliency models compute a multi-scale representation of a mesh and observe how a local vertex property such as curvature, surface variation or normal displacement changes at different scales [Pauly et al. 2003; Lee et al. 2005; Castellani et al. 2008]. These methods are tightly linked to local properties which are not robust against noise and topological changes. Song et al. [2014] address these issues by using spectral properties.

Other saliency models achieve robustness and speed by segmenting a mesh into patches represented by descriptors, followed by a rank-

SA '16 Technical Briefs, December 05 - 08, 2016, Macao
ISBN: 978-1-4503-4541-5/16/12
DOI: http://dx.doi.org/10.1145/3005358.3005380

ing process that specifies patch distinctiveness [Gal and Cohen-Or 2006]. Recent saliency models described in the above section, such as Shtrom et al. [2013] and Tasse et al. [2015], focus on point sets. They fill a gap in the literature that has become important owing to the proliferation of low-range scanning devices that produce point sets that may not always be suitable for full reconstruction.

This paper shows how saliency metrics could be used for evaluation, by focusing on three recent saliency models briefly mentioned above, and selected because they are the state-of-the-art:

Saliency of large point sets ($\mathbf{LS}$) Shtrom et al. [2013] propose the first method that supports saliency detection on large points sets. Saliency is a combination of point distinctiveness at two scales with point association, a function that assigns higher saliency to regions near foci of attention. Distinctiveness is computed by comparing local neighbourhoods described by the Fast Point Feature Histograms (FPFH) [Rusu et al. 2009]. FPFH consist of 33-D histograms of angles between oriented points in a local region.

Mesh saliency via spectral processing ($\mathbf{MS}$) Song et al. [2014] propose a spectral-based approach, described as more robust compared to previous saliency methods that analyse changes in local vertex properties. The new approach uses spectral properties of a mesh at multiple scales using the n lowest frequencies of its log-Laplacian spectrum L. The log-Laplacian spectrum amplifies variation in the low-frequency parts of the Laplacian spectrum and detects the most 'fundamental' saliencies.

Cluster-based point set saliency ($\mathbf{CS}$) Tasse et al. [2015] propose a cluster-based saliency model presented as being able to detect fine-scale saliency with better time complexity. They segment point sets into K clusters, and compute cluster saliency as a sum of cluster distinctiveness and spatial distribution. The point-level saliency is obtained by smoothing cluster-level saliency. Cluster distinctiveness is based on the mean FPFH of points belonging to that cluster, using a method similar to Shtrom et al.'s [2013].

Given that two of the methods above combine FPFH with some uniqueness heuristics, we introduce a simple model based on FPFH to investigate how it compares.

PCA-based saliency ($\mathbf{PS}$) Saliency is computed as the absolute value of the FPFH descriptors projected onto the largest principal axis after mean centering.

3 Methodology

3.1 Datasets

Ground-truth saliency is obtained from Chen et al.'s user study [2012], which comprises salient points, referred to by the authors as Schelling points. Users were asked to select points that were likely to be selected by other users. Each of the 400 meshes in the SHREC07 dataset was annotated by at least 22 participants. Chen et al. also compute a scalar field over a mesh by smoothing, with a Gaussian filter, the frequency with which each vertex was selected by all participants. We use this scalar field as ground-truth saliency.

3.2 Evaluation metrics

Saliency evaluation benchmarks in 2D images are well-established. We adapt evaluation scores used in these benchmarks to 3D saliency. Each saliency model is compared against ground-truth ($\mathbf{GS}$) using the following 3 metrics:

Area under the ROC curve (AUC): The Receiver Operating Characteristic (ROC) curve is obtained by thresholding the saliency map into a binary mask that separates positive samples (salient points) from negative samples (non-salient points) and, for different threshold values, plotting the true positive rate against the false positive rate. This metric is commonly used to compare saliency models in the $2D$ case [Judd et al. 2012; Le Meur and Baccino 2013]. The ideal saliency model has AUC of 1.0. AUC disregards regions with no saliency, and focuses on the ordering of the saliency values. Other, more selective, metrics are needed to support evaluation.

Normalized scanpath saliency (NSS): Also widely used in comparing 2D saliency maps to human eye fixations [Le Meur and Baccino 2013], NSS measures saliency values at fixation points along each user's eye scanpath. In our 3D case, we consider points selected by users as fixation points. For each participant, a NSS score is computed by a weighted sum of the computational saliency at points selected by the participant. This is in contrast with ground-truth saliency which is computed by smoothing the frequency with which points were selected. The final NSS metric is the average over all participants. The higher this metric, the closer the evaluated computational model is to ground-truth since interest points selected by users should have large saliency values.

Linear correlation coefficient (LCC): This coefficient measures the strength of the linear relationship between two variables [Borji et al. 2013]. The coefficient ranges between -1 and 1, with values closer to 0 implying a weak relationship. We use its absolute value as the metric score. With X the ground-truth and Y the saliency map under consideration, the correlation coefficient is

$$LCC(X,Y) = \frac{|\mathrm{cov}(X,Y)|}{\sigma_X \sigma_Y},$$

where σ_X and σ_Y are the standard deviation of X and Y, respectively, and $\mathrm{cov}(X,Y)$ is the covariance between the two variables. Note that one of the things captured by LCC is whether two distributions have peaks at the same place, but this is limited by the fact that LCC is greatly influenced by the shapes of these peaks. Despite its limitations, we include LCC since it is a popular metric for measuring linear relationships between distributions.

Computational models that are the closest to ground-truth have high AUC, NSS and LCC scores. We use the Wilcoxon rank-sum test [Wilcoxon 1945], a non-parametric alternative to the two-samples t-test, at a 0.05 significance level to report statistically significant differences between saliency performances of competing methods.

3.3 Selected saliency methods

We selected three computational models from the literature discussed in Section 2, based on an informal assessment of their likely quality. We either obtained source code from the authors ($\mathbf{MS}$, $\mathbf{CS}$) or implemented the method proposed in their papers ($\mathbf{LS}$):

Saliency of large point sets ($\mathbf{LS}$) [Shtrom et al. 2013]

Mesh saliency via spectral processing ($\mathbf{MS}$) [Song et al. 2014]

Cluster-based point set saliency ($\mathbf{CS}$) [Tasse et al. 2015]

To those, we added:

PCA-based saliency ($\mathbf{PS}$)

In addition to the above computational saliency models, we evaluate the following two baseline models:

Chance ($\mathbf{RS}$): We test saliency performance when saliency values are randomly assigned. Computational saliency models should have better performance than a random model. For each point on a mesh, we choose at random a value between 0.0 and 1.0 to set its saliency value.

Human performance ($\mathbf{HS}$): We are interested in how well one or more human participants' predictions differ from the consensus of all the participants. We investigate how well saliency data collected from one participant predict ground-truth saliency. We can say that a computational saliency model predicts saliency as well as a human, if its performance is similar to the performance of the average human.

The six saliency models we are evaluating vary significantly in how saliency is distributed over the mesh, with some models having more salient regions than others. For a fair comparison, the histogram of each saliency map is matched to that of the ground-truth, similar to the previous work on 2D saliency evaluation [Judd et al. 2012]. We refer to the average histogram of a ground-truth saliency map as the reference histogram. Given a saliency map, we find the discrete mapping that optimally transforms its cumulative distribution function towards that of the reference histogram. This ensures that all saliency maps have the same distribution of saliency values.

4 Experimental results

We present the performance of the six selected saliency models, under the AUC, NSS and LCC metrics. We did not report timings since various models were implemented in different languages.

4.1 Model performances

AUC is the area under the ROC curve generated by plotting true positive rate against false positive rate. Figure 2 (top) shows the performance of the selected saliency models, under the AUC metric. All models are significantly better, on average, than chance. Saliency models based on the FPFH descriptor all perform better than both human performance HS and spectral mesh saliency MS. Among these descriptor-based methods, LS has the highest AUC performance. There is no statistically significant difference between the other two descriptor-based techniques PS and CS.

Table 1: *AUC performance per shape class in SHREC07. For each class, we average the saliency AUC scores of all shapes belonging to that class. The table shows, for instance, that it is easier to detect saliency on a* mechanic *shape than a* spring. *It also shows that no tested saliency model can yet detect saliency on a* teddy *shape better than a human.*

Class	LS	PS	CS	HS (1 vs all)
mechanic	0.69 ± 0.03	$\mathbf{0.71 \pm 0.03}$	0.70 ± 0.03	0.64 ± 0.02
fish	0.67 ± 0.02	$\mathbf{0.67 \pm 0.02}$	0.67 ± 0.01	0.61 ± 0.01
armadillo	$\mathbf{0.67 \pm 0.01}$	0.66 ± 0.01	0.66 ± 0.01	0.62 ± 0.01
airplane	$\mathbf{0.67 \pm 0.02}$	0.64 ± 0.02	0.63 ± 0.02	0.60 ± 0.01
table	$\mathbf{0.67 \pm 0.03}$	0.62 ± 0.02	0.63 ± 0.03	0.61 ± 0.02
vase	0.62 ± 0.02	$\mathbf{0.63 \pm 0.02}$	0.62 ± 0.02	0.58 ± 0.01
buste	0.63 ± 0.03	$\mathbf{0.64 \pm 0.03}$	0.62 ± 0.03	0.59 ± 0.01
four-legs	$\mathbf{0.62 \pm 0.03}$	0.61 ± 0.03	0.60 ± 0.03	0.59 ± 0.01
cup	$\mathbf{0.62 \pm 0.02}$	0.61 ± 0.02	0.61 ± 0.02	0.57 ± 0.01
hand	$\mathbf{0.63 \pm 0.02}$	0.60 ± 0.02	0.61 ± 0.02	0.58 ± 0.01
chair	$\mathbf{0.66 \pm 0.02}$	0.58 ± 0.03	0.59 ± 0.03	0.62 ± 0.02
ant	$\mathbf{0.63 \pm 0.01}$	0.58 ± 0.01	0.61 ± 0.01	0.60 ± 0.01
bearing	0.64 ± 0.03	0.63 ± 0.03	$\mathbf{0.65 \pm 0.03}$	0.55 ± 0.01
bird	$\mathbf{0.62 \pm 0.02}$	0.60 ± 0.02	0.60 ± 0.02	0.57 ± 0.01
plier	$\mathbf{0.62 \pm 0.01}$	0.58 ± 0.01	0.60 ± 0.01	0.56 ± 0.01
human	0.59 ± 0.02	$\mathbf{0.59 \pm 0.02}$	0.57 ± 0.02	0.58 ± 0.01
octopus	$\mathbf{0.62 \pm 0.02}$	0.57 ± 0.02	0.55 ± 0.02	0.60 ± 0.01
teddy	0.56 ± 0.01	0.57 ± 0.01	0.57 ± 0.01	$\mathbf{0.60 \pm 0.01}$
glasses	$\mathbf{0.57 \pm 0.01}$	0.55 ± 0.01	0.52 ± 0.02	0.56 ± 0.01
spring	$\mathbf{0.55 \pm 0.01}$	0.53 ± 0.02	0.55 ± 0.02	0.54 ± 0.01

NSS is the average of saliency values at human-selected keypoints. Figure 2 (middle) presents performance under this metric. Similarly to AUC, all models perform better than chance under NSS, with spectral mesh saliency MS having a statistically lower mean score than others. LS performs as well as human performance, with no statistically significant difference between their two mean scores. CS is not statistically different from PS. Note that under the NSS metric, HS is one of the top performing models, in contrast with its low AUC score. This is because AUC is influenced by the ordering of saliency values within a ground-truth saliency map. NSS uses the user-selected keypoints directly in its formulation with no consideration of how many times a point was selected by participants.

LCC estimates the strength of the relationship between the distributions of a shape saliency map and its ground-truth saliency. Results of models' performance under LCC are presented in Figure 2 (bottom). Models' rankings under this metric are similar to rankings by NSS scores. The key difference is that there is no significant difference between CS and human performance HS. Thus CS performs better under LCC.

In summary, LS [Shtrom et al. 2013] achieves the best performance under all metrics. There is little difference between the performance of the other two FPFH-based saliency models, namely CS [Tasse et al. 2015] and PCA-based saliency. It is important to note that the PCA approach to 3D saliency, first introduced in this paper, uses dramatically less data in its comparisons, since it only uses the first principal axis of the FPFH PCA. Yet, it produces competitive results compared to the state-of-the-art. Of course, while PS uses only a single value in its comparison, it comes at the expense of a preprocessing step: running PCA on an $n \times 33$ matrix to get the single saliency value, where n is the number of points; however it does not need to compare FPFHs, unlike LS and CS.

4.2 Performance per shape classes

We analyse how the selected saliency models perform for each of the 20 classes in SHREC07. This gives us insight into which classes have the worst saliency detection and thus could benefit from future work in the field. Table 1 shows AUC scores for each class and saliency model. For the class *teddy*, human performance (HS) outperforms all other saliency models. In this particular case, human participants know that real-life shapes from this class have facial features and thus consider the face to be salient even if the 3D shape presented to them has no discriminating features. This is an example of humans using their prior experience in saliency detection, which is not yet possible for unsupervised saliency models. Man-made shapes such as *mechanic*, *airplane* and *chair* are easier for saliency detection due to their sharp features and simple structure. In the case of categories such as the feature-less *spring*, a method based on supervised learning could be more successful.

5 Conclusion

We introduce the first saliency evaluation framework for 3D shapes, based on three performance metrics. We compare six saliency models including chance, human performance and a PCA-based

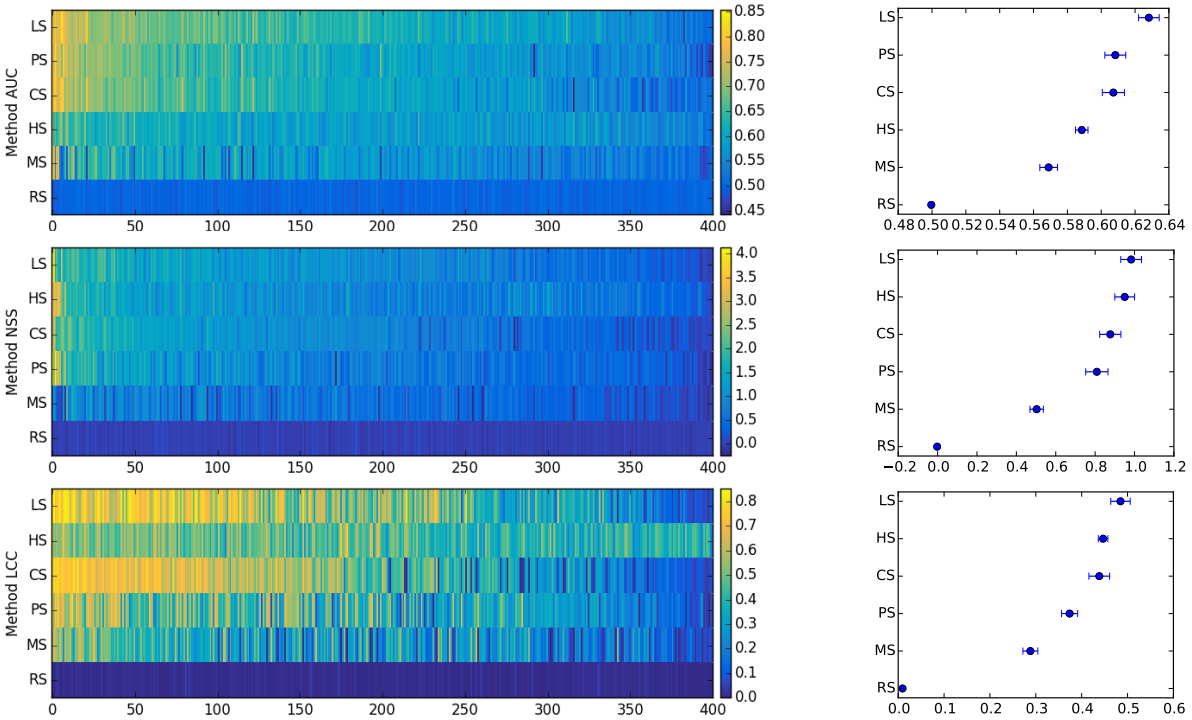

Figure 2: *Saliency performance evaluation. The figure has three parts, one for each metric: AUC (top), NSS (middle), and LCC (bottom). Left: a color map with the x-axis showing shapes ordered by their average score and the color scale representing the actual metric value. Right: saliency models ordered by their average score, with blue error bars representing 95% confidence intervals.*

approach. Results showed that all models performed better than chance, and Shtrom et al.'s method [2013] has, on average, the best scores. PCA-based saliency performed as well as cluster-based point set saliency [Tasse et al. 2015], and significantly better than spectral mesh saliency [Song et al. 2014]. Using our proposed evaluation framework, previous methods not tested here and future saliency techniques can be evaluated objectively.

References

BORJI, A., SIHITE, D. N., AND ITTI, L. 2013. Quantitative analysis of human-model agreement in visual saliency modeling: a comparative study. *IEEE Trans. Im. Proc. 22*, 1, 55–69.

CASTELLANI, U., CRISTANI, M., FANTONI, S., AND MURINO, V. 2008. Sparse points matching by combining 3D mesh saliency with statistical descriptors. *Computer Graphics Forum 27*, 2, 643–652.

CHEN, X., SAPAROV, A., PANG, B., AND FUNKHOUSER, T. 2012. Schelling points on 3D surface meshes. *ACM Trans. Graph. 31*, 4 (July), 29:1–29:12.

DUTAGACI, H., CHEUNG, C. P., AND GODIL, A. 2012. Evaluation of 3D interest point detection techniques via human-generated ground truth. *Vis. Comput. 28*, 9 (Sept.), 901–917.

GAL, R., AND COHEN-OR, D. 2006. Salient geometric features for partial shape matching and similarity. *ACM Trans. Graph. 25*, 1 (Jan.), 130–150.

GIORGI, D., BIASOTTI, S., AND PARABOSCHI, L., 2007. Shape retrieval contest 2007: Watertight models track.

JUDD, T., DURAND, F., AND TORRALBA, A. 2012. A benchmark of computational models of saliency to predict human fixations. In *MIT Technical Report*.

LE MEUR, O., AND BACCINO, T. 2013. Methods for comparing scanpaths and saliency maps: strengths and weaknesses. *Behavior Research Methods 45*, 1, 251–266.

LEE, C. H., VARSHNEY, A., AND JACOBS, D. W. 2005. Mesh saliency. *ACM Trans. Graph. 24*, 3 (July), 659–666.

LIU, X., LIU, L., SONG, W., LIU, Y., AND MA, L. 2016. Shape context based mesh saliency detection and its applications: A survey. *Computers & Graphics 57*, 12 – 30.

PAULY, M., KEISER, R., AND GROSS, M. H. 2003. Multi-scale feature extraction on point-sampled surfaces. *Comput. Graph. Forum 22*, 3, 281–290.

RUSU, R. B., BLODOW, N., AND BEETZ, M. 2009. Fast point feature histograms (FPFH) for 3D registration. In *ICRA*, 1848–.

SHTROM, E., LEIFMAN, G., AND TAL, A. 2013. Saliency detection in large point sets. In *ICCV 2013*, 3591–3598.

SONG, R., LIU, Y., MARTIN, R. R., AND ROSIN, P. L. 2014. Mesh saliency via spectral processing. *ACM Trans. Graph. 33*, 1 (Feb.), 6:1–6:17.

TASSE, F. P., KOSINKA, J., AND DODGSON, N. 2015. Cluster-based point set saliency. In *ICCV*, 163–171.

WILCOXON, F. 1945. Individual comparisons by ranking methods. *Biometrics Bulletin 1*, 6, 80–83.