Mesh Saliency and Human Eye Fixations

YOUNGMIN KIM, AMITABH VARSHNEY, DAVID W. JACOBS, and FRANÇOIS GUIMBRETIÈRE
Department of Computer Science, University of Maryland

As the 3D data size grows, it is more important to integrate principles of saliency with geometric processing of meshes. Mesh saliency has been proposed as a computational model of perceptual importance for meshes and it has been used in graphics for abstraction, simplification, segmentation, illumination, rendering, and illustration. Even though this technique is inspired by models of low-level human vision, it has not yet been validated. Here we present a user study that compares the previous mesh saliency approach with human eye movements. Our results show that the current computational model of mesh saliency can model human eye movements significantly better than what can be expected purely by chance.

Categories and Subject Descriptors: I.3.5 [**Computer Graphics**]: Computational Geometry and Object Modeling; I.3.m [**Computer Graphics**]: Perception

General Terms: Algorithms, Human factors, Verification

Additional Key Words and Phrases: Visual perception, mesh saliency, eye-tracker

1. INTRODUCTION AND RELATED WORK

When people examine an image, their eyes tend to fixate on certain points, then jump quickly, with *saccades*, to new points. Although viewers may attend to portions of an image on which they do not fixate, a good deal of evidence suggests that viewers tend to move their eyes to parts of an image that have attracted their attention (see [Palmer 1999], Chapter 11, for a brief review). For this reason, many models of visual attention and saliency have been evaluated by their ability to predict eye movements. It is not realistic to expect any model to perfectly predict eye movements, because of the variability between human subjects and even for the same subject at different times. However, recent research demonstrates that there is a significant correlation between existing models and human eye fixations. For example, Privitera and Stark [2000] compare points of fixation by subjects to clusters formed by the most salient regions predicted by a large number of simple models of 2D image saliency. They compare this with the degree to which fixations agree between subjects. They have found that for a particular class of images (paintings), algorithms based on simple operators including symmetry, center-surround, and discrete wavelet transform cohere very well with human data and approach

Email: {ymkim,varshney,djacobs,francois}@cs.umd.edu

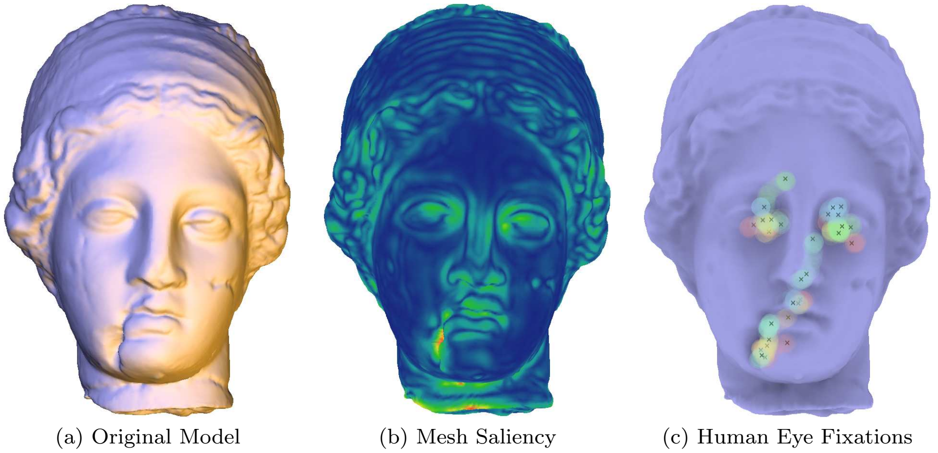

(a) Original Model (b) Mesh Saliency (c) Human Eye Fixations

Fig. 1. *Image (a) shows the Igea model. Image (b) shows its computed mesh saliency and image (c) shows the human eye fixations color-coded for six subjects.*

the coherence among fixations across subjects. Parkhurst *et al.* [2002] measure the saliency at points of fixation and show that the model of 2D image saliency of Itti *et al.* [1998] is more predictive of fixation points than a random model. Previous research also makes the useful methodological points that bottom-up models can better predict the first fixations, which are less influenced by top-down knowledge ([Parkhurst et al. 2002]), and that the exact order of fixations is highly variable and difficult to predict [Privitera and Stark 2000].

Lee *et al.* [2005] have proposed a model of mesh saliency as a measure of regional importance. Their method for computing mesh saliency uses a center-surround mechanism that is inspired by the human visual system. Similar mechanisms have been widely used in models of 2D image saliency [Itti et al. 1998] [Koch and Ullman 1985]. Previous research in vision has assumed that visual search simply relies on two-dimensional aspects of an image. However, Enns and Rensink [1990] have shown that 3D information can play an important role in making salient objects pop-out in a cluttered image. As far as we know, there has been no work comparing models of 3D saliency to eye movements, although many experiments have measured eye movements as subjects examine 3D objects [Howlett and O'Sullivan 2005] [Kim and Varshney 2006] [Cole et al. 2006] [Lu et al. 2006].

2. EXPERIMENT

The computational model of mesh saliency uses a center-surround mechanism that is inspired by the human visual system. Our goal is to explore whether the mesh saliency model has better correlation with human eye fixations than a purely random model. To gather objective evidence of this correlation, we have carried out an eye-tracking-based user study and have quantified the similarity between the model and human eye fixations.

2.1 Physical Setup

We used the ISCAN ETL-500 eye-tracker which can record eye movements continuously at 60 Hz. The study was carried out on a 17-inch LCD display with a resolution of 1280 × 1024, placed at a distance of 24 inches, subtending a visual angle of approximately 31.4 degrees horizontally. The subjects had a chin rest to minimize head movements and to maintain calibration. Our experimental setup is shown in Figure 2.

2.2 Eye-tracker Calibration and Subject Selection

The standard calibration of ETL-500 eye-tracker was performed with 4 corner points and one center point. However, this was not sufficiently accurate for our purposes due to non-linearities in the eye-tracker-calibrated screen space. Therefore we used the second calibration step which involves a more densely-sampled calibration phase similar to [Parkhurst et al. 2002] with 13 additional points. For this we asked the subjects to successively look at and click on 13 points presented on the screen. This gave us an accurate correspondence between the eye-tracker space and the monitor space for that subject. After this we tested the accuracy of the calibration by asking the subjects to look at 16 randomly selected points on the screen. Of the 9 subjects participating for pay, 6 were able to successfully calibrate to within an accuracy of 30 pixels (about .75 degree) for each of the 16 points. We proceeded with our study using these 6 subjects with normal or corrected-to-normal vision. Our subjects were not familiar with the goals of this study. The subjects were told to freely view the images with no assigned goal.

2.3 Stimuli

There were a total of 5 models and each model was shown from 10 different views. We have used the Armadillo, Dinosaur, Igea, Isis, and Male models shown in Figure 7 for our study.

Since the computational model of mesh saliency relies only on geometric properties (curvature values), we would like to see whether it actually correlates with the human eye fixations from any viewing direction. For this purpose, we generated images from 10 different views and used them in our study. As shown in Figure 3, we have generated ten (five right-side up and five upside down) views for each model. We manually choose the first view. We rotate this model -30°, -15°, 15°, and 30°

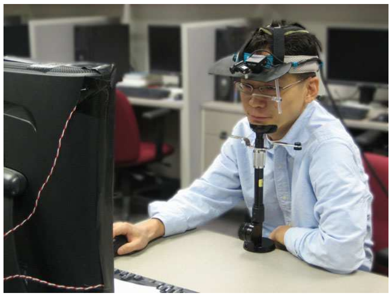


Fig. 2. *Our experimental setup for the user study with the ISCAN ETL-500 eye-tracker.*

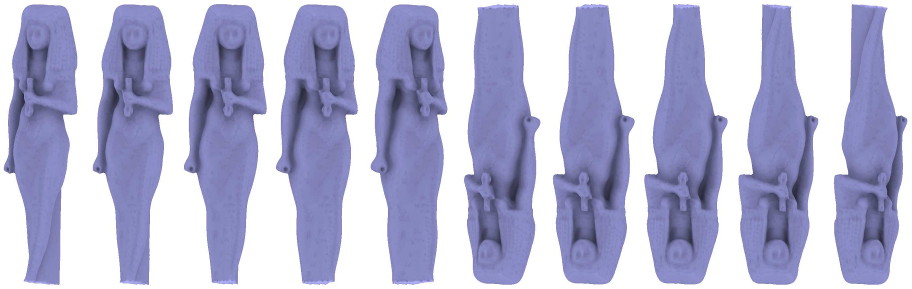
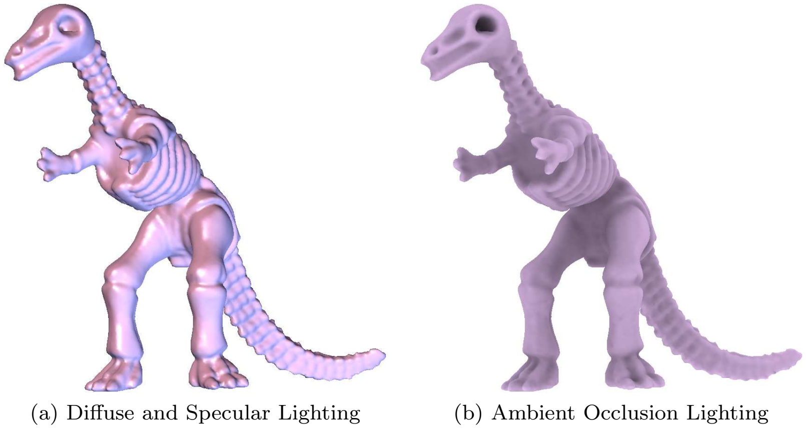


Fig. 3. *Ten different views of the Isis model. There are five right-side up and five upside down views, and these views are rotated 15 degrees apart along the vertical axis.*



(a) Diffuse and Specular Lighting

(b) Ambient Occlusion Lighting

Fig. 4. *The Dinosaur model under different lighting conditions. Local illumination models such as specular lighting and diffuse lighting can result in high contrast regions depending on the normal direction, the light position, and the viewer position while ambient occlusion lighting minimize this effect.*

along the vertical axis to generate four more views of the model. Then we turn each of these views upside down to generate the remaining five views.

Image Ordering: Each user saw 50 images (5 models × 10 views). When we ordered the images for each user, we minimized differential carryover effects by placing similar images far apart. Alleviating differential carryover effect was very important because each user looked at 5 similar images ($-30°$, $-15°$, $0°$, $15°$, and $30°$).

Image Synthesis Consideration: In our experiments for validating mesh saliency, we wanted to minimize the influence of lighting on the human perception of the rendered images. The easiest solution is to use a simple ambient term, but this approach leads to indiscriminate flatness. Instead, we use ambient occlusion [Landis 2002] [Zhukov et al. 1998], in which each illumination at a vertex is proportional to the fraction of the environment that it can see. We preprocess the view-independent occlusion information using ray casting and then use this information at runtime for ambient occlusion lighting. The effect of local illumination models and ambient occlusion is shown in Figure 4.

The user study had 52 trials (images). The data captured for the first two images was discarded as they were intended to give the subjects a sense of the duration. Each trial started with the subject seeing a blank screen with a cross at the center of the screen. The subject was asked to look at the cross before clicking the mouse to bring up the next image. This ensured that each trial started with the subject's eyes fixated at the center of the image. Each image was shown for 5 seconds. Each study took about 5 minutes.

2.4 Hypothesis

Our hypothesis is that the computational model of mesh saliency has better correlation with human eye fixations than a random model regardless of viewing direction for the first few seconds after stimulus onset.

As we have explained in Section 2.3, there are three independent variables in our experiment: models, rotations of models, and right-side up vs. upside down views.

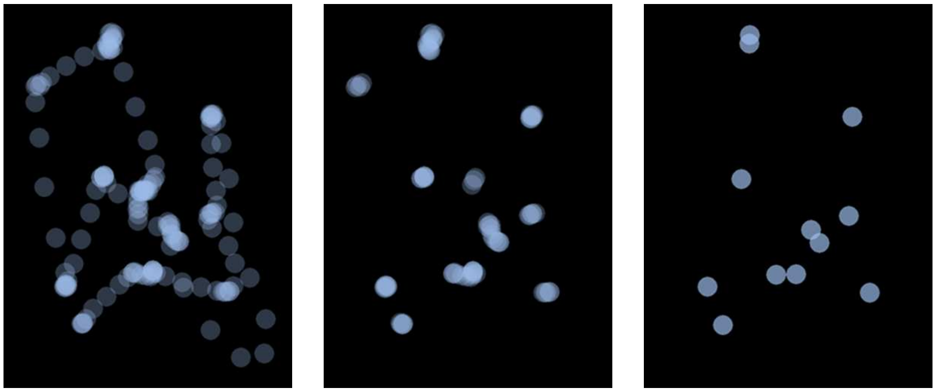
Models and rotations: Parkhurst *et al.* [2002] have observed that fixations for the subjects are usually biased towards the center. In our experiment, each subject is asked to look at the cross at the center of the screen before each trial. By using different models and rotating the models, we change the distances from the center to the high saliency regions.

Right-side up vs. upside down views: Recent work on gaze control has focused on two attentional models: bottom-up stimulus-based information and top-down memory-based knowledge [Henderson 2003]. The 3D models we used in our study were not absolutely knowledge-free as they were scanned models of animals or humans. Subjects could use generic semantic and spatial knowledge even though we are measuring their eye movements for short time (the first five seconds). Parkhurst *et al.* [2002] support this argument by showing that stimulus dependence is greatest for early fixations in their work. We include upside down views to slow down the onset of top-down attentional effects.

3. DATA ANALYSIS

3.1 Fixation Points

We divide the raw data points from the eye-tracker into two groups – *fixation points* which correspond to a user looking at a single location and *saccade points* which correspond to fast eye movements from one fixation point to the next. We followed an approach similar to the one suggested by Stampe [1993] to identify fixations and saccades. Figure 5 shows a two step process to extract fixation points from the raw data points. We considered data points that had a velocity greater than $15°/sec$ as saccade points and removed them. We then averaged consecutive eye locations that were within 15 pixels and classified them as a single fixation point. Some researchers have advocated discarding short (exploratory) fixations in measuring the attention of the viewer [Henderson and Hollingworth 1998]. We ignored the brief fixations below the threshold of 133ms. This corresponds to 8 consecutive points in the ISCAN ETL-500 eye-tracking device.



(a) Raw Data Points (b) Saccades removed (c) Fixation Points

Fig. 5. *Image (a) shows all the raw data points from the eye-tracking device. Image (b) shows the points remaining after removing saccade points. Image (c) shows final fixation points after removing brief fixations and combining consecutive points if they are spatially close.*

3.2 Normalized Chance-adjusted Saliency

3.2.1 *Chance-adjusted Saliency.* Parkhurst *et al.* [2002] introduced the notion of chance-adjusted saliency to quantify the correlation between stimulus saliency and fixation locations for an image. They compute the chance-adjusted saliency as follows. Consider a collection of images I_i, $1 \le i \le N$. A subject is asked to look at each image in turn. This generates a set of fixation points f_{ij}, $1 \le j \le F_i$ for each image I_i. Let us consider the k-th fixation points f_{ik} across all the images I_i. Let s_{ik} be the saliency value at the k-th fixation point f_{ik} in the image I_i. They compute the mean fixation saliency for the k-th fixation points as $\bar{s}^f_k = \frac{1}{N}\sum_{i=1}^{N} s_{ik}$. To compute the mean random saliency, they first generate F_i random points r_{ij} over each image I_i, where $1 \le i \le N$ and $1 \le j \le F_i$. Then, the mean random saliency $\bar{s}^r_k$ is computed as the average saliency over the k-th random point r_{ik} across all the images I_i, $1 \le i \le N$. Finally, they define the chance-adjusted saliency (s^c_k) for the k-th fixation points as the difference between the mean fixation saliency ($\bar{s}^f_k$) and the mean random saliency ($\bar{s}^r_k$): $s^c_k = \bar{s}^f_k - \bar{s}^r_k$.

3.2.2 *Normalized Chance-adjusted Saliency.* We observed three shortcomings in using the previously defined chance-adjusted saliency to quantify the correlation between human eye fixations and the model of mesh saliency.

(1) The chance-adjusted saliency was developed for images in which there is a well-defined saliency at every pixel. We are trying to measure the correlation between a mesh saliency approach and the fixation points on the mesh but not the fixations on the entire rendered image. Therefore, we should only consider the foreground pixels that are covered by projected triangles of the mesh. This ensures fairer comparisons between a random model and the mesh saliency model for 3D rendered images because excluding the background pixels would prevent lowering the average saliency values in a random model. Figure 6 shows the points considered in chance-adjusted saliency and normalized chance-

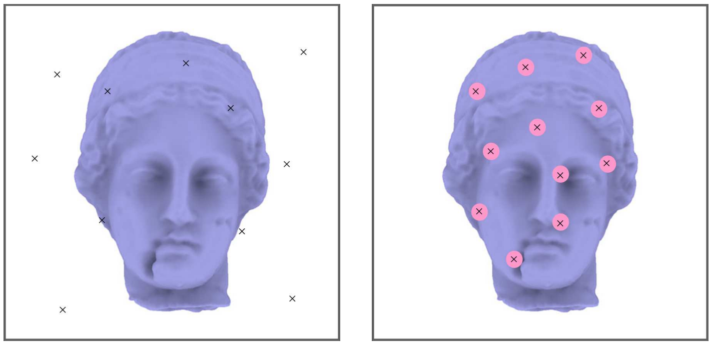


Fig. 6. *The left image shows the random points in chance-adjusted saliency computation. These points are scattered all over the image. The right image shows the points that we consider in normalized chance-adjusted saliency computation. We only include the foreground pixels that are covered by projected triangles of the mesh. For each fixation point represented as a cross, we also take into account the eye-tracker accuracy of* 20 *pixels, which is represented as a circle.*

adjusted saliency.

(2) The chance-adjusted saliency does not consider eye-tracker accuracy. Since the fixation point acquired from the eye-tracker can differ from the actual pixel that a user looked at, we have to consider the eye-tracker accuracy as shown in Figure 6(b) when we assign the mesh saliency value to the fixation point.

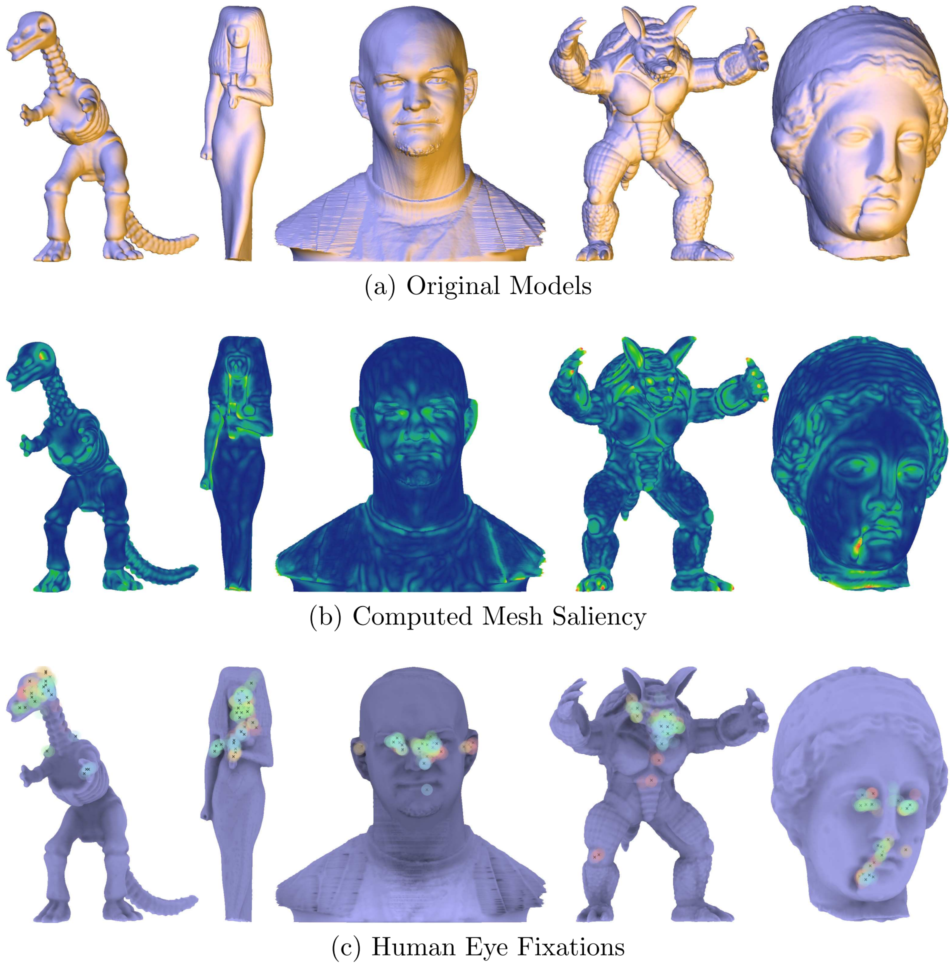
(3) The chance-adjusted saliency is defined over a collection of images. This restricts the analysis of the effect of different models and viewing directions. We need a method that normalizes saliency on a per-image basis.

We define *normalized chance-adjusted saliency* in this section. First, we consider the eye-tracker accuracy ϵ which depends on both the accuracy of the eye-tracking device and the calibration steps. We have used $\epsilon = 20$ pixels, subtending a visual angle of approximately 0.5 degree horizontally. Note that a fixation point and a pixel share the same coordinate system. Let us consider the the pixel p_{ij} on which a fixation point f_{ij} falls. Instead of taking the saliency value on a fixation point f_{ij}, we compute the error-adjusted saliency s^{ϵ}_{ij} as the maximum of the saliency values within a radius of $\epsilon = 20$ pixels around p_{ij} in the image I_i, $1 \le i \le N$: $s^{\epsilon}_{ij} = \max_{k \in \mathcal{N}^{\epsilon}_j} s_{ik}$, where $\mathcal{N}^{\epsilon}_j = \{k | dist(p_{ij}, k) \le \epsilon\}$. For each rendered image I_i, we compute the mean $(\bar{s}^{\epsilon r}_i)$ of the saliency s^{ϵ}_{ij} of the pixels j that are covered by the rendered mesh. Then, we define our normalized chance-adjusted saliency for the fixation point f_{ik} as $s^{n}_{ik} = s^{\epsilon}_{ik} / \bar{s}^{\epsilon r}_i$.

3.3 Results

Figure 7 shows the fixation points and computed mesh saliency for one viewpoint of each model. Fixation points are color-coded for six subjects. We observe that most fixations are close to warm-colored salient regions computed by the model of mesh saliency.

The results of our normalized chance-adjusted saliency values can be seen in Table I and Figure 8. We have shown the average of the normalized chance-adjusted



(a) Original Models

(b) Computed Mesh Saliency

(c) Human Eye Fixations

Fig. 7. *The comparison between the model of mesh saliency and human eye fixations. The first row shows the original Dinosaur, Isis, Male, Armadillo, and Igea models. The second row shows the computed mesh saliency for each model. Here warm colors indicate high saliency regions while cool colors indicate low saliency regions. The third row shows the human eye fixations from our eye-tracking-based user study. Fixation points are color-coded for six subjects.*

Table I. *Normalized chance-adjusted saliency values for each model.*

Model	Average	StdDev
Armadillo	1.4497	0.3206
Dinosaur	1.5144	0.5212
Igea	2.0149	0.3735
Isis	2.2343	0.3927
Male	1.9786	0.5157

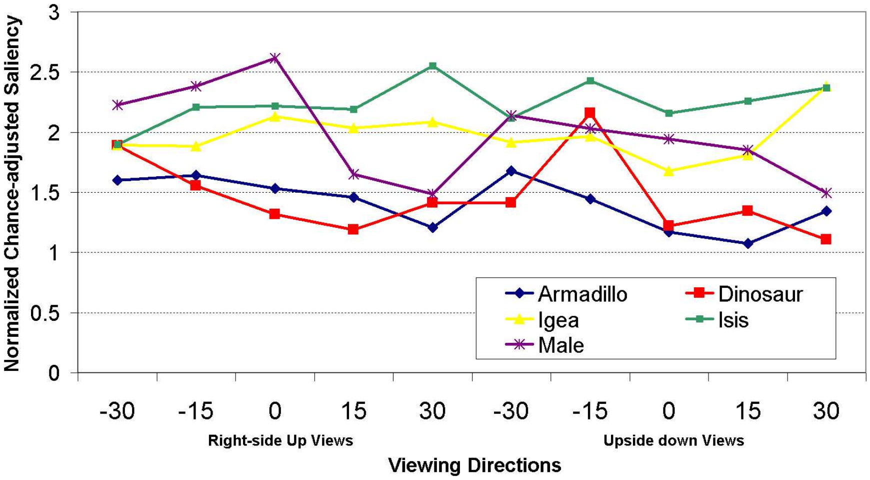


Fig. 8. *Average normalized chance-adjusted saliency value across subjects for each viewing direction for each model. For all the cases, the values are higher than 1, which is the value that can be expected by chance.*

saliency values across all views of each model in Table I. In general, we have observed that the computational model of saliency has a higher correlation with human eye fixations than a random model as the normalized chance-adjusted saliency values are higher than 1, the value that can be expected purely by chance. However, there is considerable variability depending on the models and rotations of models as shown in Figure 8.

To observe the influence of models and viewing directions, we first carried out a two-way ANOVA on the normalized chance-adjusted saliency values with two conditions: different models and viewing directions. For both models and viewing directions, we have observed significant differences: $F(4, 250) = 54.90, p < 0.001$ and $F(9, 250) = 3.698, p < 0.001$, respectively. The difference between viewing directions is especially interesting. The result indicates that even though saliency of each vertex in a model can be computed once and does not change, the projection of a model could result in different behavior in drawing viewers' attention depending on viewing directions. The possible causes include occlusion amongst salient and non-salient regions and distance changes from the center to the salient regions (as we explained in Section 2.4).

Next we carried out a two-way ANOVA on the normalized chance-adjusted saliency values with two conditions: different models and right-side up vs. upside down views. We have observed significant differences, $F(4, 290) = 41.73, p < 0.001$, for different models as expected. We have observed there are no significant differences between right-side up and upside down views, $F(1, 290) = 1.99, p = 0.159$. This indicates that regardless of inversion of right-side up views, there is significant correlation between eye fixation points and the model of mesh saliency. Turning the model upside down is likely to reduce the effect of semantics, and therefore this result indicates that the effect of semantics is not high enough to influence the

Table II. *List of pairwise t-tests (two-tailed).*

Model	t-Value	p-Value
Armadillo	-8.37	< 0.0001
Dinosaur	-7.20	< 0.0001
Igea	-18.26	< 0.0001
Isis	-17.18	< 0.0001
Male	-12.54	< 0.0001

correlation between eye fixations and the mesh saliency model during the first five seconds.

Since the standard deviations are high and there are high variance among models in Table I, we have carried out a pairwise t-test on the average saliency values between fixation points and random points for each model (this is the only condition in the test). We compared the model of saliency to the random model as in Parkhurst *et al.* [2002]. We compute the average saliency values from randomly chosen locations lying on the foreground instead of the observed fixation locations. Table II shows that there are significant differences in the average saliency values for all the 3D models between the mesh saliency model and the random model.

The results validate that the mesh saliency model has significantly higher correlation with human eye fixations than a random model regardless of viewing direction.

4. DISCUSSION, CONCLUSIONS, AND FUTURE WORK

We have used a few devices to reduce the effect of semantics in this paper. The first one is including upside down views. The second one is limiting the time to the first five seconds. Others [Parkhurst et al. 2002] [Santella and DeCarlo 2004] have also used similar durations. However, even five seconds could be considered too long since semantic interpretation starts increasing right after the stimulus onset. We plan to study more about the effect of semantics as we vary the durations. Another thing we can do for reducing the effect of semantics is to experiment on semantic-free objects such as some man-made objects or close-up views of scanned models.

According to the results of the two-way ANOVA in Section 3.3, there is no significant difference between right-side up and upside down views. However, the lack of significance might be simply due to the lack of power (i.e. running more subjects will change the results). We plan to perform a user study with more subjects in the future.

In this paper, we have taken the first steps towards validating the existing mesh saliency model through an eye-tracking based user study. We have introduced the notion of normalized chance-adjusted saliency which is a robust measure of success of a mesh saliency model. We have observed significant correlations between the model of mesh saliency and human eye fixations regardless of the viewing direction. Having a validated model of mesh saliency will be greatly useful for several contexts. For example, it could be helpful for identifying the role of 3D information in visual search task as Enns and Rensink [1990] have explored in their work. In addition, we can now build further saliency-based systems for tasks such as visual enhancement. Conversely, our carefully designed user studies can be also helpful for designing a better visual saliency model which better models human eye movements.

At present we have analyzed the effect of turning each model upside down, but more dynamic analysis can be performed in the future. For example, we can analyze if people fixate on same or different locations as the model rotates. Our normalized chance-adjusted saliency can give us a general correlation between human eye fixations and the mesh saliency model. However, this measure cannot let us directly compare the mesh saliency model to other algorithms or to human eye movements. By clustering the most salient regions predicted by a computational model of mesh saliency, we can compare the agreement between the mesh saliency model and eye movements with the inter-subject agreement of fixations as Privitera and Stark [2000].

Another interesting future work will be comparing the mesh saliency model to other 3D measures such as the magnitude of curvature. It will also be interesting to compare and contrast the 3D mesh saliency with 2D image saliency.

ACKNOWLEDGMENTS

We would like to thank Stanford 3D Scanning Repository and Cyberware for 3D models. This work has been supported in part by the NSF grants: IIS 04-14699, CCF 04-29753, CNS 04-03313, and CCF 05-41120.

REFERENCES

Cole, F., DeCarlo, D., Finkelstein, A., Kin, K., Morley, K., and Santella, A. 2006. Directing gaze in 3D models with stylized focus. In *Eurographics Workshop/ Symposium on Rendering*. 377–387.

Enns, J. and Rensink, R. 1990. Sensitivity to three-dimensional orientation in visual search. *Psychological Science 1,* 5, 323–326.

Henderson, J. M. 2003. Human gaze control during real-world scene perception. *Trends in Cognitive Science 11,* 7, 498–504.

Henderson, J. M. and Hollingworth, A. 1998. *Eye movements during scene viewing: An overview. In Eye Guidance in Reading and Scene Perception.* Elsevier Science Ltd.

Howlett, S. and O'Sullivan, C. 2005. Predicting and evaluating saliency for simplified polygonal models. *ACM Trans. on Applied Perception 2,* 3, 286 – 308.

Itti, L., Koch, C., and Niebur, E. 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. on Pattern Analysis and Machine intelligence 20,* 11, 1254–1259.

Kim, Y. and Varshney, A. 2006. Saliency-guided enhancement for volume visualization. *IEEE Transactions on Visualization and Computer Graphics (Proceedings of IEEE Visualization) 12,* 5, 925–932.

Koch, C. and Ullman, S. 1985. Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology 4*, 219–227.

Landis, H. 2002. Production-ready global illumination. *Course 16 notes, SIGGRAPH 2002*.

Lee, C. H., Varshney, A., and Jacobs, D. 2005. Mesh saliency. *ACM Trans. on Graphics (Procs. ACM SIGGRAPH) 24,* 3, 659 – 666.

Lu, A., Maciejewski, R., and Ebert, D. 2006. Volume composition using eye tracking data. In *Eurographics/IEEE VGTC Symposium on Visualization*. 147–154.

Palmer, S. 1999. *Vision Science: Photons to Phenomenology.* MIT Press, Cambridge, MA.

Parkhurst, D., Law, K., and Niebur, E. 2002. Modeling the role of salience in the allocation of overt visual attention. *Vision Research 42,* 1, 107–123.

Privitera, C. M. and Stark, L. W. 2000. Algorithms for defining visual regions-of-interest: Comparison with eye fixations. *IEEE Trans. on Pattern Analysis and Machine Intellignece 22,* 9, 970–982.

Santella, A. and DeCarlo, D. 2004. Visual interest and NPR: an evaluation and manifesto. In *Proceedings of NPAR*. 71–150.

Stampe, D. 1993. Heuristic filtering and reliable calibration methods for video-based pupil tracking systems. *Behavior Research Methods, Instruments and Computers 25*, 137 – 142.

Zhukov, S., Inoes, A., and Kronin, G. 1998. An ambient light illumination model. *Proceedings of Eurographics Rendering Workshop*, 45–56.