A Comparative Perceptual Study of Soft-Shadow Algorithms

MICHAEL HECHER, Vienna University of Technology
MATTHIAS BERNHARD, Vienna University of Technology
OLIVER MATTAUSCH, University of Zurich
DANIEL SCHERZER, HS Ravensburg-Weingarten
MICHAEL WIMMER, Vienna University of Technology

We performed a perceptual user study of algorithms that approximate soft shadows in real time. While a huge body of soft-shadow algorithms have been proposed, to our knowledge this is the first methodical study for comparing different real-time shadow algorithms with respect to their plausibility and visual appearance. We evaluated soft-shadow properties like penumbra overlap with respect to their relevance to shadow perception in a systematic way, and believe that our results can be useful to guide future shadow approaches in their methods of evaluation. In this study we also capture the predominant case of an inexperienced user observing shadows *without* comparing to a reference solution, e.g., when watching a movie or playing a game. One important result of this experiment is to scientifically verify that real-time soft-shadow algorithms, despite having become physically based and very realistic, can nevertheless be intuitively distinguished from a correct solution by untrained users.

Categories and Subject Descriptors: I.3.7 [**Computer Graphics**]: Three-Dimensional Graphics and Realism—*Shadowing*

General Terms: Measurement, Human factors

Additional Key Words and Phrases: Soft Shadows, Perception

ACM Reference Format:

Hecher, M., Bernhard, M., Mattausch, O., Scherzer, D., and Wimmer, M. 2014. A Comparative Perceptual Study of Soft-Shadow Algorithms. *ACM Trans. Appl. Percept.* 0, 0, Article 0 (0), 20 pages.
DOI: http://dx.doi.org/10.1145/0000000.0000000

1. INTRODUCTION

Real-time algorithms for soft shadows have been a major branch of real-time rendering research for many years. The goal is to find clever ways to provide a *plausible* approximation of the appearance of physical soft shadows within a time budget of 60 frames per second. In this matured field, the difference between competing algorithms is often quite subtle (Figure 1). Traditionally, shadow algorithms have been evaluated by comparing the output to a reference solution in a couple of selected scenes, and highlighting the per-pixel differences. Furthermore, the judging authority is usually a group of experts on shadows. However, this is not the way a typical user, who does not have a reference solution to compare to and is not an expert, experiences shadows.

This work has been supported by the EU FP7 People Programme (Marie Curie Actions) under REA Grant Agreement no. 290227 and the Austrian Science Fund (FWF) contract no. P23700-N23. Author's address: M. Hecher; email: hecher@cg.tuwien.ac.at;

DOI: http://dx.doi.org/10.1145/0000000.0000000

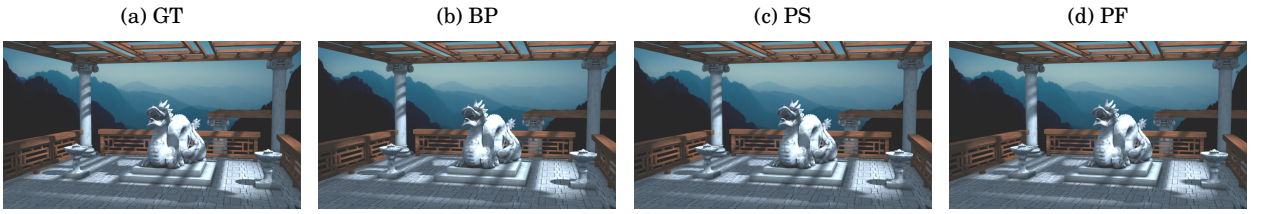

Fig. 1: A scene rendered with four different soft-shadow algorithms. Although results look very similar, there are subtle differences in soft-shadow regions. The question is whether users are able to perceive these differences, and if they are able to differentiate the ground truth solution (GT) from physically-based real-time approximations like Backprojection (BP), or simple non-physical solutions like Percentage-Closer Soft Shadows (PS), and Percentage-Closer Filtering (PF).

Hence, it has been a major motivation of this work to question how *plausible* a given shadow algorithm appears to the *typical* user as compared to an experienced user, an expert user, or somebody with a reference solution, and whether a casual user is able *in principle* to judge the perceptual differences between two qualitatively different classes of shadow algorithms. These questions have to be evaluated *methodically* with respect to a variety of relevant shadow properties. In particular, we have to examine the question of how *sensitive* human perception is to certain features of a soft-shadow algorithm in a particular environment.

Surprisingly, the ability to correctly estimate the softness of shadows (e.g., for judging the current light conditions) has to our knowledge not been subject to scientific evaluation. This is quite remarkable considering that the importance of shadows as an essential depth cue and a means to make a scene look realistic to an observer is well known.

The contributions of this paper are the following:

- The first comprehensive study that *systematically* compares classes of soft-shadow algorithms with respect to their ability to approximate certain key properties of physical penumbrae within various carefully chosen scene categories.
- New insights about the perception of soft shadows of non-experienced, experienced, and expert users, which could be used as guidelines for game designers and researchers. For example, the overall preference for physically-based models indicates that striving for higher accuracy in real time might have practical relevance.
- New facts about soft-shadow study design and visual soft-shadow experiments in general. For instance, how different questions can effect the user's judgment of soft shadows. Also, which scene categories and soft-shadow features are suitable for soft-shadow comparisons, and which are not (surprisingly, it turned out that vegetation scenes are not).

1.1 Study Scope and Limitations

Our main goal was to explore a manageable range of independent variables and investigate them thoroughly in a worst-case scenario. In particular, we judge how well a soft-shadow algorithm can capture the physical properties of the penumbra regions, and put as much emphasis as possible on the penumbra differences. We avoid low-level rendering artifacts of the chosen algorithms as much as possible, to offer a fair comparison of their penumbra approximation quality. Only static soft shadows are investigated, as the evaluation of animations requires additional degrees of freedom which would have been difficult to handle adequately within the scope of a single study. We do not use textures in all but one of the scenes in order to avoid potential masking of wrong soft shadows [Ferwerda et al. 1997]. We did, however, include a single realistic game-type scene with textures into our study to get

an outlook on the factors involved. Another limitation of this study is that in each scene a light source of the same size which was visible in the rendered image (except in the game-type scene) was used. However, different sizes and shapes of the light source scale the penumbra size in the entire shadow by the same factor and the current results indicate that users better understand relative differences in the shadow, i.e., the way the penumbra changes.

2. STUDY OVERVIEW

We conducted a perceptual study using the *method of pairwise comparison* to reveal whether and to which extent users are able to perceive differences between approximations of varying quality. Another aim of the experiment was to find out which features in a soft shadow significantly improve the perceived quality. Since soft shadows are a complex lighting phenomenon, the comparison task is challenging and can be facilitated by showing a ground truth reference. However, comparing to a ground truth reference is an artificial setting and the result likely overestimates a typical user's ability to judge soft shadows. On the other hand, we suspected that the performance in this task is a matter of user experience. An inexperienced user might have a much lower sensitivity compared to an experienced one with knowledge about soft shadows. Thus we investigated three levels of user experience, which we defined as follows: the *inexperienced*, the *experienced*, and the *cognizant* user. These three levels were investigated by a study paradigm in which the behavior of each user type was simulated in one of three consecutive phases. The initially inexperienced users became experienced through a learning phase, where a ground truth reference was provided. To find out how these results compare to genuine experts of soft-shadow rendering, we included people with scientific publications in the field of soft shadows and/or global illumination and compare their results to non-expert users.

As differences between individual soft-shadow approximations can be rather subtle, particularly in the perception of an inexperienced user, we found that the commonly used concept of a two-option forced-choice comparison can have disadvantages which can limit the informative value of the study by decreasing *user consistency*. We therefore use an alternative concept which leaves a neutral option.

3. RELATED WORK

3.1 Perception of shadows and illumination

Shadow perception was the topic of several psychological studies. Wanger et al. showed that shadows are an important visual clue for the spatial relationship of objects [Wanger 1992; Wanger et al. 1992]. Knill et al. [1997] were able to show that shadows can provide information about the direction of the light source and the shape of the underlying surface. Other research focused on the role of shadows [Hu et al. 2000] and interreflections [Madison et al. 2001] for perceiving object contact. Jarabo et al. [2012] evaluated how factors like geometric complexity and color affect the perceived fidelity of illumination in virtual scenes with dynamic crowds.

There has been previous research on how to accelerate soft-shadow rendering in regions where the human sensitivity is low and hence a lower quality shadow is sufficient [Sattler et al. 2005; Vangorp et al. 2006; Schwarz and Stamminger 2008b]. However, these works are complementary as they address the question of scaling a particular type of algorithm, whereas we provide a study about comparing the perception of distinct classes of soft-shadow algorithms and whether handling different features of soft shadows increases the perceived quality.

3.2 Study design

As computer graphics has matured as a science, there has been an increased interest to evaluate the benefit of various graphical effects in comprehensive perceptual studies, like caustics [Gutierrez

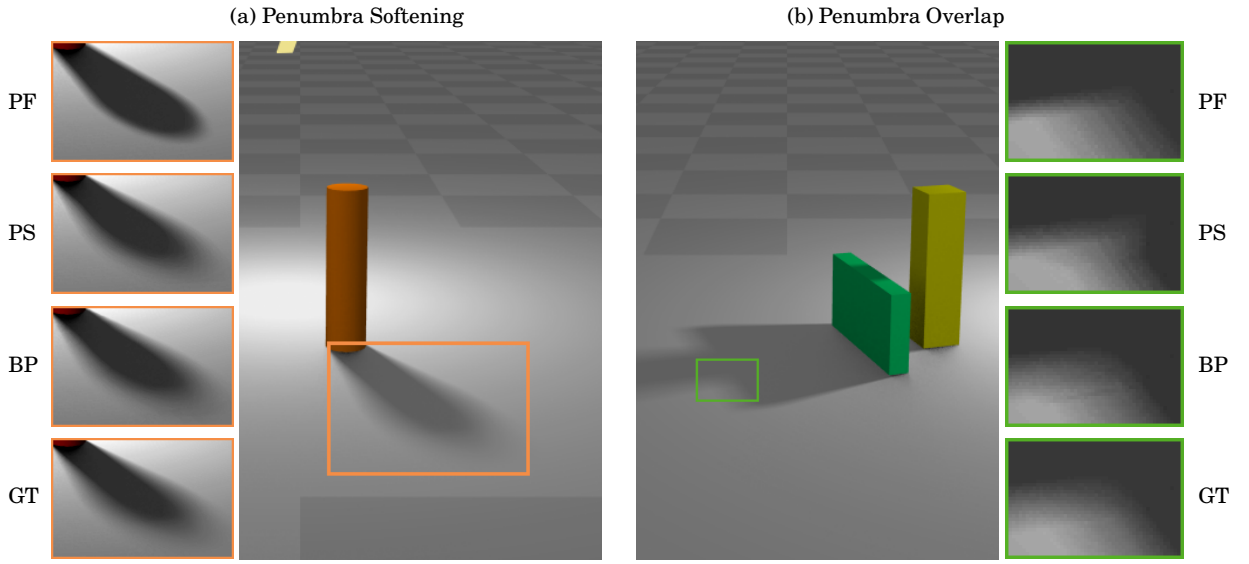

Fig. 2: Visual comparison of soft-shadow algorithms. Closeups are shown of the two properties that are most difficult to fake by heuristics and hence account for most of the visual difference between the shadow algorithms: *penumbra softening* and *penumbra overlap*. Note that the contrast of enlarged images was adjusted to better visualize differences.

et al. 2008], tone-mapping [Ledda et al. 2005], geometric distortions [Setyawan and Lagendijk 2004], image-retargeting algorithms [Rubinstein et al. 2010], and global illumination [Čadík et al. 2012]. The structure of our study is naturally related to these, and we discuss similarities as well as differences with respect to them in the following.

In the experiment of Ledda et al. [2005], LDR images produced by different tone-mapping operators were shown to participants on two LCD monitors, while showing the original image on a HDR monitor as a reference solution. This approach is useful when the performance of algorithmic processing has to be tested against the optimal solution and provides a worst case scenario where participants know exactly what a stimulus should look like. In contrast to Ledda et al., the large-scale user study of Rubinstein et al. [2010] uses pairwise comparison *without* providing a reference solution.

In our study we include *both* scenarios, first with and then without a reference solution, and use them to emulate different degrees of user expertise to evaluate soft-shadow algorithms. This is akin to Čatík et al. [2012], who evaluated image quality metrics for detection of global illumination and rendering artifacts. In addition to these two phases, we study the learning effect achieved from showing the reference solution in another study phase, again without reference.

An interesting alternative method to study the perception of rendering methods has been used by Yu et al. [2009], who investigated the required accuracy in visibility calculations to generate a plausible global illumination. They used the so-called *paired comparison plus category* [Scheffe 1952] paradigm. Instead of responding with preferences, participants scored the *similarity* of two images on a scale of predefined categories ranging from low to high similarity. To obtain an estimate of similarity in our study as well, we incorporate the number of neutral responses into the evaluation.

Setyawan and Lagendijk [2004] analyze the quality of responses of single participants and the agreement between all participants. We adopt this concept in our study to evaluate the performance of participants, and whether the learning phase has been effective. Following Akyüz et al. [2007], we use Tukey's HSD method [Steel et al. 1997] to find groups of algorithms for which we could not measure

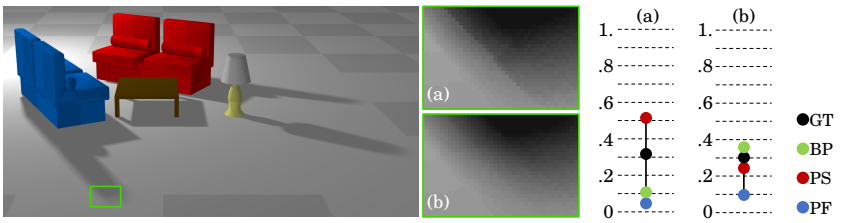

Fig. 3: The probability to select an algorithm in a scene (a) with and (b) without artifacts in the BP algorithm. The absence of artifacts led to noticeably different results. Note that the contrast of enlarged images was adjusted to better visualize differences.

a significant difference and hence judged them to be of comparable quality. To better evaluate and understand the effect of experience on the outcome, we introduce a new measure which we denote as *degree of indifference* (explained in detail in Section 7), and compute *worth parameters* [Hatzinger and Dittrich 2012] from the pairwise comparison results.

4. SELECTED SOFT-SHADOW ALGORITHMS

Because of the huge variety of algorithms that approximate soft shadows in real time [Eisemann et al. 2011], the selection of a suitable subset was a crucial process in the preparation of this study. On the one hand, our goal was to cover the whole range from simple empirical to costly but fully physical methods. On the other hand, we had to restrict ourselves to a feasible number of algorithms we can investigate in a single study. We selected three algorithms that form *representative classes* of real-time soft-shadow approaches, with respect to their capability of approximating the real penumbra of a physical shadow using available visibility information, and add a brute-force ground-truth solution as comparison. We argue that most methods used for real-time soft-shadow rendering in practice are variations of these basic forms, or tweaks that aim to hide certain *artifacts*. Also, only algorithms based on shadow mapping have been considered because of their predominance in all major real-time applications. A visual comparison of all our algorithms can be seen in Figure 2.

We decided to avoid artifacts like *overshadowing* or *staircase artifacts*, because their presence resulted in a clear preference for artifact-free images in preceding tests (see Figure 3). The used algorithms are:

Percentage-Closer Filtering (PF). PF was introduced as an anti-aliasing method for hard shadow borders by filtering the binary shadow map test results [Reeves et al. 1987]. However, it can also be used to simulate the penumbra regions of soft shadows by setting an appropriate filter kernel size. However, as the kernel size is a global parameter and local penumbra size variations are not possible, the method cannot reproduce *penumbra softening* as shown in Figure 2a. This is a cheap non-physical algorithm, and our expectation was that it would emerge as the least physical plausible in the study.

Percentage-Closer Soft Shadows (PS). PS [Fernando 2005] is an extension of PF where the size of the penumbra is computed dynamically for each fragment, based on the relative distance of a receiver and the blocker geometry from the light source. It uses a single shadow map for approximating the distance of the blocker geometry from the light, and hence is known as a single-sample soft-shadow algorithm. This simple empirical model accounts for penumbra size variation, but the non-physical nature of the algorithm causes problems in regions where penumbrae with different degree of softness overlap (Figure 2b). A recent method achieves higher quality by combining PS with temporal coherence [Schwärzler et al. 2013], which however only works flawlessly as long as the coherence does not break.

Table I. : The capabilities for soft-shadow approximation supported by the algorithms investigated in this study.

	pen. size variation	physically-based pen.	physically-correct pen.
PF	–	–	–
PS	✓	–	–
BP	✓	✓	–
GT	✓	✓	✓

Backprojection (BP). Backprojection [Atty et al. 2006] estimates the amount of light reaching a point in the scene by identifying small blocker patches from a shadow map and backprojecting them on the light source. This method is physically based except for the fact that visibility is estimated with a single sample only. The algorithm can handle both penumbra size variation as well as complex penumbra overlap more accurately than the simpler PS method. Unfortunately, BP is prone to light leaks and overshadowing artifacts (i.e., penumbrae can become slightly too dark). We chose to use a gap-filling approach [Guennebaud et al. 2006] to reduce light leaks with parameters chosen carefully so as not to lead to overshadowing artifacts.
Note that several methods with the goal to make BP more robust have been published since, usually alleviating the single-sample artifacts. However, they all come with the cost of more complex visibility evaluations in the form of multiple shadow-map layers [Schwarz and Stamminger 2008a; Bavoil et al. 2008] or shadow-map samples from multiple view-points [Yang et al. 2009], essentially relaxing the real-time requirements for improved accuracy.

Ground Truth (GT). We use a brute-force method as the reference solution, which samples the area light source multiple times and then blends together the resulting shadow maps. As this method converges to the physical ground truth for a sufficient number of point lights (disregarding global illumination effects), we call it the ground-truth solution within the scope of this study. In our experiments, the solution converged after 1024 samples (i.e., there were no more changes of pixel intensities). This approach is slow but produces correct penumbra softening without overlapping artifacts.

The approximation quality each of these methods can achieve is fundamentally different and increases when going from PF to GT. While PF can only model uniform penumbrae, PS has the capability for local *penumbra size variation*. BP goes a step further by computing *physically-based* penumbrae, as it uses a correct physical model of integrating the light-source coverage. GT calculates a *physically correct* penumbra in the sense that it converges to the real solution when 1024 samples are used. See Table I for a summary of the algorithm capabilities.

5. INDEPENDENT VARIABLES OF SHADOW PERCEPTION

Many factors potentially influence the perception of soft shadows, like observer and object movement, texture color, light angle, number of lights, and scene complexity. In this study, we focus on the following two factors: the spatial relation of light and objects and the resulting penumbra categories, and the degree of user experience. We evaluate these factors with a meaningful set of test scenes.

5.1 Penumbra Categories

We identified two properties of soft shadows which cause quality differences between soft-shadow algorithms: 1) *Penumbra overlaps* in the case of multiple shadow casters, which can lead to complex shadowing effects that are difficult to fake using empirical models (e.g., correct occluder fusion), and 2) *penumbra softening*, which occurs when the distances between shadow caster points and their corresponding points on the shadow receiver have sufficient variations along the shadow casting surface

(e.g., consider a point on the top and bottom of the thin occluder in Figure 2a). Figure 2 demonstrates a situation with visible differences between all used shadow algorithms with respect to these 2 properties. Both phenomena can be present in a scene (or not), resulting in 4 possible combinations denoted as *penumbra categories*. If neither of these effects is present, the soft-shadow algorithms are basically indistinguishable.

5.2 User Experience

Another goal of this study is to evaluate the perception of soft-shadow approximations at different levels of user experience. We chose to make use of the knowledge participants gain during the study by designing multiple phases in which we evaluate three levels of experience. This design has the additional benefit of being able to study the learning behavior of a typical user. In particular, each user participates in three consecutive phases. In the first phase, we measure the ability to distinguish soft-shadow approximations expected from an *inexperienced* user, in the second the maximum capability we expect from a *cognizant* user, and in the third phase, the performance of an *experienced* user. These phases are discussed in detail in Section 6. Additionally we included a group of soft-shadow *experts* who performed the same experiment under identical conditions.

5.3 Scene Categories

To produce meaningful results, we need to include a variety of object and shadow complexities in the chosen set of scenes. In particular, we want our study to be *systematic* to cover a wide variety of possible scene configurations. However, there are many factors that potentially influence shadow perception, like the number of objects in the scene, their respective complexity, and their arrangement. The challenge was to find a small set of categories that systematically cover most possible scene configurations encountered in practice, while avoiding the curse of dimensionality when investigating all combinations of factors. The chosen *scene categories* are:

Simple Objects. A scene is made up of simple, regular objects like boxes and cylinders. It is an important limit case, since users should have the strongest conception about the final shadow shapes. Most complex objects can be constructed from these basic building blocks. In applications, such objects are for instance used to represent roofless houses or columns in urban environments.

Complex Objects. Complex objects are assemblies of objects from the first category which form more complex arrangements, but do not show too much regularity.

Structured Objects. Structured objects show a high amount of regularity and symmetry, e.g., they show translational symmetry in 1 or 2 dimensions [Pauly et al. 2008]. For example, fences can be seen as a regular arrangement of boxes in one dimension, and the crossbars of the fence as regular alignment of boxes in the horizontal dimension. It is essential to have this as a separate category since humans will recognize structure also in the shadow. Additionally, the regular parts must be connected so that their shadows are not perceived as belonging to separate entities.

Vegetation. Vegetation denotes an irregular and seemingly random arrangement of smaller objects like leaves, where statistical properties like the leaf density become more important than individual primitives. We chose this as a separate category since it constitutes another limit case due to the minimal amount of regularity that can be used by an observer to compare the shadow with the shadow caster. Also, it constitutes a difficult case for soft-shadow approximations because of highly complex silhouettes.

To arrive at the configurations to actually present to the participants, we observe that simply combining scene and penumbra categories would give 16 different combinations. However, there are some special cases. For fences, grids and plants, penumbra overlap will always occur as individual parts

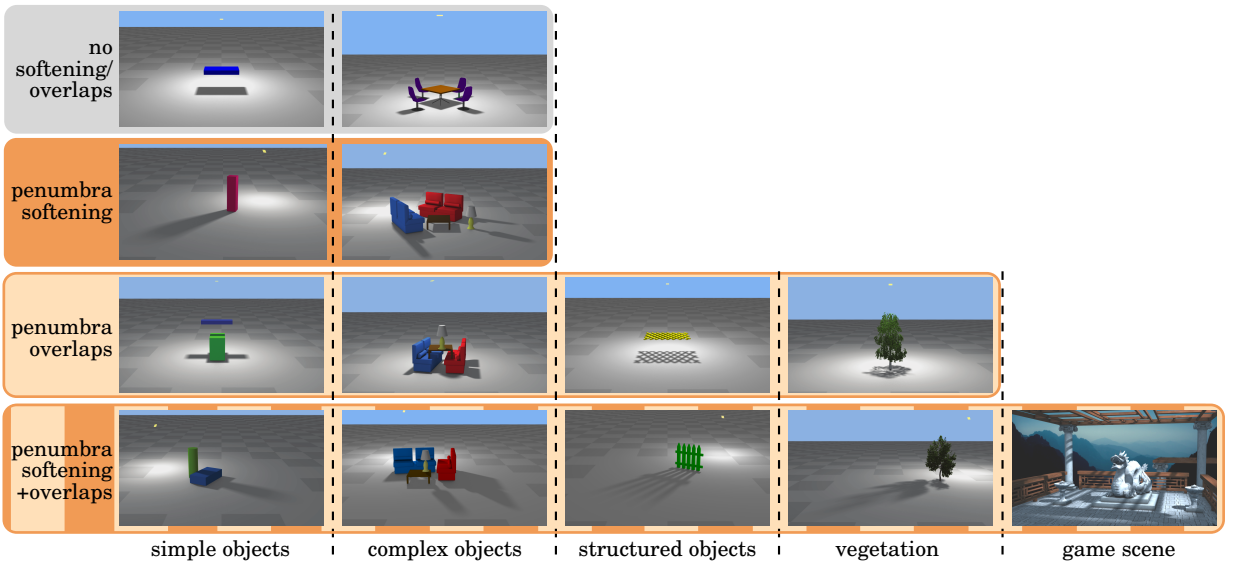

Fig. 4: This table shows the investigated combinations of *penumbra* and *scene categories*. Rows denote *penumbra categories* and colors *scene categories*. An empty field marks a category that is covered by another (see Section 5.3 for details).

of such objects have to be connected. Typically, this is also true for random objects like trees, which reduces the possible combinations to 12. Finally, to also investigate a less artificial setting, we added another category (a game scene), with softening and overlapping penumbrae that are partially cast on uneven surfaces, resulting in a total of 13 combinations. We represent each of these combinations by 2 different scenes, leading to 26 test scenes overall. Figure 4 shows a selection of these scenes and classifies them by penumbra (rows) and scene categories (columns).

Other Scene Characteristics. For each of the variables for which we could not introduce another dimension to sample, we decided to use the configuration that corresponds most to a "worst-case" condition, i.e., one where a user is most sensitive to soft-shadow inaccuracies: Using static view conditions maximizes the ability to inspect shadows thoroughly. A plain white surface as shadow-receiving geometry enables optimal perception of shadow properties without masking effects. Using a white color for the shadow-casting light maximizes the contrast of the shadow. Furthermore, we increased the plasticity of the shadow-casting objects to facilitate the recognition of their geometry for the user. This was achieved by computing ambient occlusion with a commercial offline renderer, and diffuse shading on a per-pixel basis. To emphasize the distinction of different objects, shadow-casting objects were textured with primary and secondary colors (red, yellow etc.). Distance perception was facilitated by using a checkerboard texture for regions surrounding the white shadow-receiving area. To produce a pleasant but serious atmosphere, ambient light was enabled and the background was colored light blue.

6. EXPERIMENTAL PROCEDURE

We conducted an experiment with 48 participants (age mean 26.27, stdev 7.57, 25 female, 23 male) who were recruited by advertising in university-related online forums. No artists, heavy gamers, or persons with background in computer graphics were allowed to participate in the study to ensure inexperienced participants. Additionally, we recruited 8 experts (age mean 33.75, stdev 7.76, all male) who have one or more scientific publications dealing with shadows and/or global illumination. People

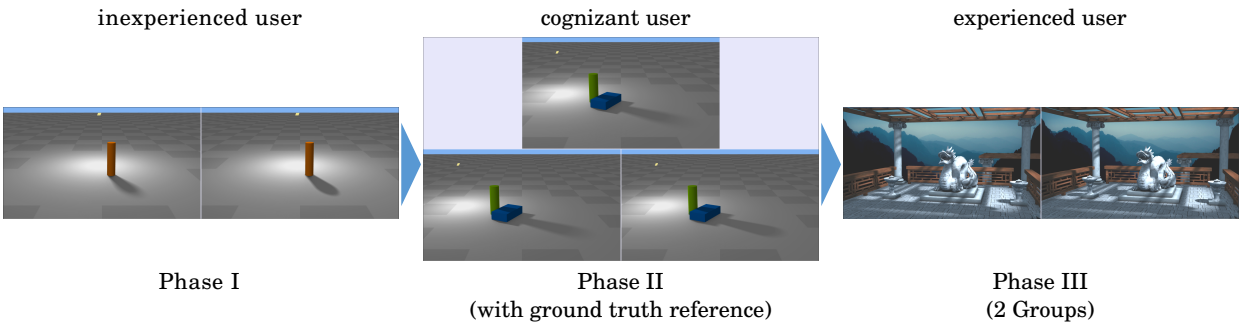

Fig. 5: The three-phase design of our experiment. In the first phase users were inexperienced and performed a pairwise comparison task. In the second phases they had a reference solution to compare to, and in the last phase they again did pairwise comparison but were randomly assigned to one of two groups with slightly different tasks.

were randomly assigned to one of two groups (Group A: male 12, female 13, experts 4; Group B: male 11, female 12, experts 4). In total, 14,976 votes were cast.

The experiment was carried out in a room with blacked-out windows to ensure the same purely artificial light conditions. On average, the experiment took 46 minutes for a non-expert user and 62 minutes for an expert user (not including regeneration breaks). To assure a sufficient level of motivation, each participant was paid a reimbursement. The experiment was divided into three phases (separated by short recreational breaks), each of which comprised 91 trials in total (see Figure 5).

6.1 Comparison Task

Participants were instructed that they would see two images in each trial, which they would have to compare by judging the soft shadows cast on the floor of the scene. They were informed that images could be identical. They were also told that the light source producing the shadow is an area light source visualized as a yellow square on its respective position. Each test subject had to participate in all three phases in the specified order (Figure 5).

Phase I. In this phase, an initially inexperienced user had to compare two images rendered with different shadow algorithms without a reference. The participant was asked to choose the image where he/she thinks the soft shadow *looks better*. We chose this question because we observed in pilot testing that this is more intuitive for inexperienced users than asking to choose the "physically more plausible" soft shadow.

Phase II. In the second phase, a reference image was shown on top of the image pair to be compared. The participants were informed that this is the physically correct simulation. They were then asked to judge which of the two images below the reference had the physically more plausible soft shadows.

Phase III. In the last phase, the comparison task was performed again without a reference. Participants were divided into two groups, with female and male subjects being evenly distributed over both. *Group A* received the same question as in Phase I (which shadow looks better), and *Group B* the same question as in Phase II (which shadow is physically more plausible).

Responses were given with a slider, where participants could score the degree of preference for the left or the right image of a pair. The slider had a continuous scale. On the left and right ends, labels were placed defining a clear preference for the corresponding image. At the center of the slider, which was located between both images, there was a mark defining no preference.

The decision against a 2AFC design was motivated by our concern for cases where the actual perceivable difference between the results of two algorithms was low. In these cases, a neutral option

avoids forced decisions based on other factors than those investigated (e.g., a general preference for sharp edges) causing contradicting responses that lower the consistence of the results (see Section 7). Moreover, counting the neutral responses provides us a measure for the perceived similarity in the quality of the compared algorithms in particular scene categories (see Section 7).

Note that a continuous response scale was chosen instead of an ordinal one to reduce the issue of predominantly neutral or nearly neutral responses. The continuous scale is supposed to give participants the illusion of specifying the magnitude of their preference, whereas the data was ultimately analyzed on an ordinal scale ("left preferred", "no preference", "right preferred").

There was no time limitation for a comparison. The next trial was launched after the response to the previous one and was preceded by a black fixation cross on a white background displayed for two seconds.

6.2 Image Pairs

The combination of $\binom{4}{2} = 6$ possible pairings of the 4 soft-shadow algorithms with 13 penumbra-scene category combinations gives 78 pairwise comparisons per participant. To this, we added one pair of identical images for each penumbra-scene combination, rendered with a randomly chosen algorithm, giving in total 91 image pairs. Identical image pairs were used to obtain a baseline for participants' responses and their agreement for identical image pair conditions.

To have a minimum variety in each penumbra-scene combination, we created two scenes for each of the 13 combinations, which were obtained by different arrangements or by using different models. All images used in the experiment were generated offline by rendering each scene with the four soft-shadow algorithms. All four methods utilized a shadow map of 1024^2 pixels with 32 bit floating-point precision. Some configurations like the color and intensity of the light source and the ambient term were kept equal for all rendering methods. Other rendering parameters were adjusted such that the resulting images resembled the ground truth as much as possible.

To counterbalance time-related effects, all participants saw an individual randomized order of image pairs throughout all phases. The images of each comparison pair were randomly assigned to the left or right position. In Phase I, participants saw one of the two scenes we created for each category and in Phase II the other. Which of both scenes was selected for which phase was also randomly determined for each category and participant. In Phase III, image pairs were picked randomly for both scenes of a category such that a participant saw 50% of the scenes in the first phase and 50% in the second, respectively.

7. ANALYSIS METHODOLOGY

After transforming slider responses into categories "left preferred" (slider$<$0), "no preference" (slider=0), and "right preferred" (slider$>$0), we obtain a 4×4 matrix A. In each cell A_{ij} where $i \neq j$, we have an entry specifying the number of preferences of algorithm i over j (and vice versa in A_{ji}). The amount of "no preference" responses A^0_{ij} when comparing i and j can be obtained by subtraction from the number of participants ($A^0_{ij} = N - A_{ij} - A_{ji}, i > j$).

To analyze the quality of the results in terms of consistence within comparison responses and agreement between different participants, we compute the so-called *coefficient of consistence* and *coefficient of agreement* [Kendall and Gibbons 1990], as proposed by Setyawan and Lagendijk [2004]. Since we are also interested in the informative value given by the amount of neutral responses, we define a new measure which we denote as *degree of indifference*. This measure should give us an estimate for the perceived similarity of soft shadows computed by different algorithms.

For the interpretation of the results, we were further interested in the "worth" of a particular algorithm, which we obtain by transforming pairwise comparison results into a one-dimensional quality

measure using the Bradley-Terry Model. For a scientifically reliable interpretation of the results, we further need to evaluate the results in terms of their statistical significance. To this end, we perform *post-hoc testing* and group algorithms were our experiment could not measure a significant difference.

We will extend on these analysis tools in the following:

Coefficient of Consistence (ξ). The coefficient of consistence ξ is a measure to score the amount of contradictory responses. It is computed by counting or estimating the amount of so-called *circular triads* which occur when algorithm i is preferred over j, j over k, and k over i (see [Setyawan and Lagendijk 2004] for details).

Low consistence is an indicator for *systematic* wrong responses due to an inconsistent judgment strategy (e.g., random responses) or confounding factors due to a bad selection of test objects. This measure is computed per participant and is suited to analyze the performance of individual participants. Thus it is useful to identify and exclude participants with an unacceptably low level of compliance.

The coefficient of consistence can reach a maximum of 1.0 when there are zero circular triads in the data and a minimum of 0, which is expected for random responses. Figure 6 visualizes the results for each participant and will be discussed in Section 8.

Coefficient of Agreement ($\boldsymbol{u}$). To measure the agreement between users in preference votes, we compute the coefficient of agreement u. This measure is computed by summing over all comparisons the number of "participant pairs" $\binom{A_{ij}}{2}$ which agree in their choice [Setyawan and Lagendijk 2004]. It is a measure to evaluate the ability of participants to make "good" judgments in the comparison task (e.g., to identify the more physically correct shadow).

This coefficient can reach a maximum of 1.0 when there is total agreement, that is, each entry in A_{ij} is either 0 or N. The minimum is -1.0, which occurs when $N = 2$ and each A_{ij} contains a 1. Rules of thumb are $\leqslant 0.0$ no agreement, $]0, 0.2)$ slight agreement, $]0.2, 0.4)$ fair agreement, $]0.4, 0.6)$ moderate and > 0.6 high agreement.

Degree of Indifference (ι). We also make use of the information obtained by analyzing the amount of no-preference responses. We therefore define the new measure *Indifference* ι, which we compute from the proportion of no-preference responses A^0_{ij} among all (N) comparisons being performed with pairs of different images. We normalize this score with the expected maximum of no-preference answers, which we obtain empirically from the no-preference responses A^0_{ii} in the N_0 comparisons where the participants were shown pairs of identical images:

$$\iota = \frac{\frac{1}{N}\sum_i \sum_{j>i} A^0_{ij}}{\frac{1}{N_0}\sum_i A^0_{ii}}$$

This measure is useful to estimate the perceived similarity of the images shown for one scene category. The measure has a maximum of 1, which we get when responses are equivalent to those observed for identical image pairs, and a minimum of 0, which we get in case of zero neutral responses.

Post-Hoc Testing. To group algorithms for which our experiment could not measure a significant difference, we perform post-hoc testing using Tukey's HSD test, as described by Akyüz et al. [2007]. To this end, we run a procedure which performs iteratively multiple comparisons with an ANOVA test to find means that are significantly different from each other under a significance level α, which we set to 5%. The algorithm groups found through post-hoc testing are presented in the top row of Figure 7 for all results pooled together (Figure 7a) and individual penumbra categories (Figure 7b-7e).

Obtaining Worth Parameters. To obtain a quality measure on a one-dimensional scale from a matrix of pairwise comparison results, we fit the so-called loglinear Bradley-Terry (LLBT) model to matrix A.

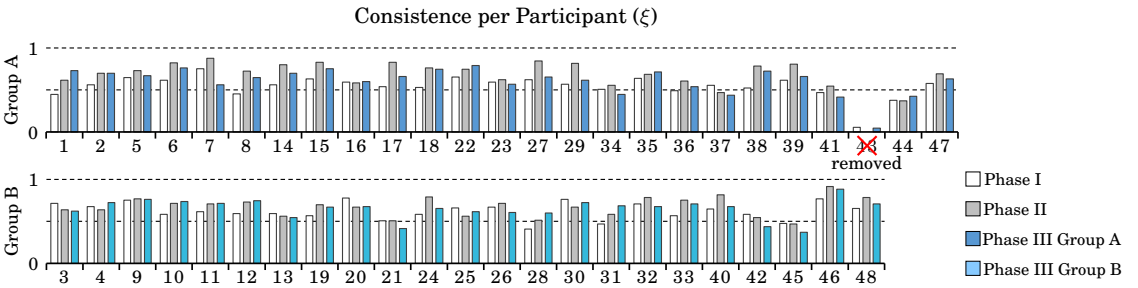

Fig. 6: Consistence values per participant. Numbers from 1 to 48 denote participants. For each participant the consistence for Phase I, II, and III are listed in consecutive order. The votes of participant 43 were removed as consistence is suspiciously low.

Since the basic LLBT model does not assume the possibility of neutral responses, we use an extension [Davidson and Beaver 1977] which accommodates ties from neutral answers. The model assumes so-called *object parameters* λ_i and λ_j and the *undecided effect* γ. They predict the probabilities p_{ijk} for a comparison between objects i and j to result in one of three outcomes k (1: "i preferred", 2: "j preferred", and 3: "no preference"):

$$p_{ij1} = \frac{e^{\lambda_i - \lambda_j}}{\Omega} \qquad p_{ij2} = \frac{e^{\lambda_j - \lambda_i}}{\Omega} \qquad p_{ij3} = \frac{e^{\gamma}}{\Omega} \qquad \Omega = e^{\lambda_i - \lambda_j} + e^{\lambda_j - \lambda_i} + e^{\gamma}$$

We compute the fit to the extended LLBT model by using the *prefmod* R-package published by Hatzinger and Dittrich [2012]. The outcome are object parameter estimates λ_i for each algorithm i, which can be transformed into *worth parameters* π_i ($\pi_i = \frac{e^{2\lambda_i}}{\sum_k e^{2\lambda_k}}$). Worth parameters specify the probability for users to prefer one object (algorithm) i over the others. We use worth parameters to render *worth plots*, which provide a convenient comparative visualization of perceived algorithm qualities. They can be found in Figure 7 below the results of post-hoc testing. This combination is a powerful tool to better understand significant differences between algorithms and groups of algorithms in post-hoc testing results.

8. RESULTS

The data collected in this experiment is a matrix of preference votes which we process using the tools described in the previous section. In order to analyze the expressiveness of our results, we compute the coefficient of agreement, indifference scores, perform post-hoc testing to group algorithms that are not significantly different, and compute worth plots.

To get first an insight about the quality of the results, we computed the coefficient of consistence on the response data of each participant as shown in Figure 6. While most participants have a quite acceptable and balanced score, participant 43 gave many contradictory responses and was thus excluded from a further analysis. An overview of the most important results is shown in Figure 7, where the results were pooled for all scenes (Figure 7a) or grouped by penumbra categories (Figure 7b–7e).

With respect to the scene categories, we conclude that the results obtained for vegetation scenes are not consistent with those from other scenes. A reason for this behavior is that we were not able to avoid/control all factors that influence shadow perception which may be caused by the high complexity of a tree shadow pattern. We therefore excluded the vegetation scenes from the analysis of the pooled results in Section 8.1 and 8.2 and we discuss instead the results in isolation in Section 8.4.

In the following, we proceed with an analysis for non-expert users. To study the effect of user experience, we pooled the results of all scenes (Section 8.1). In Section 8.2, we investigate the perception

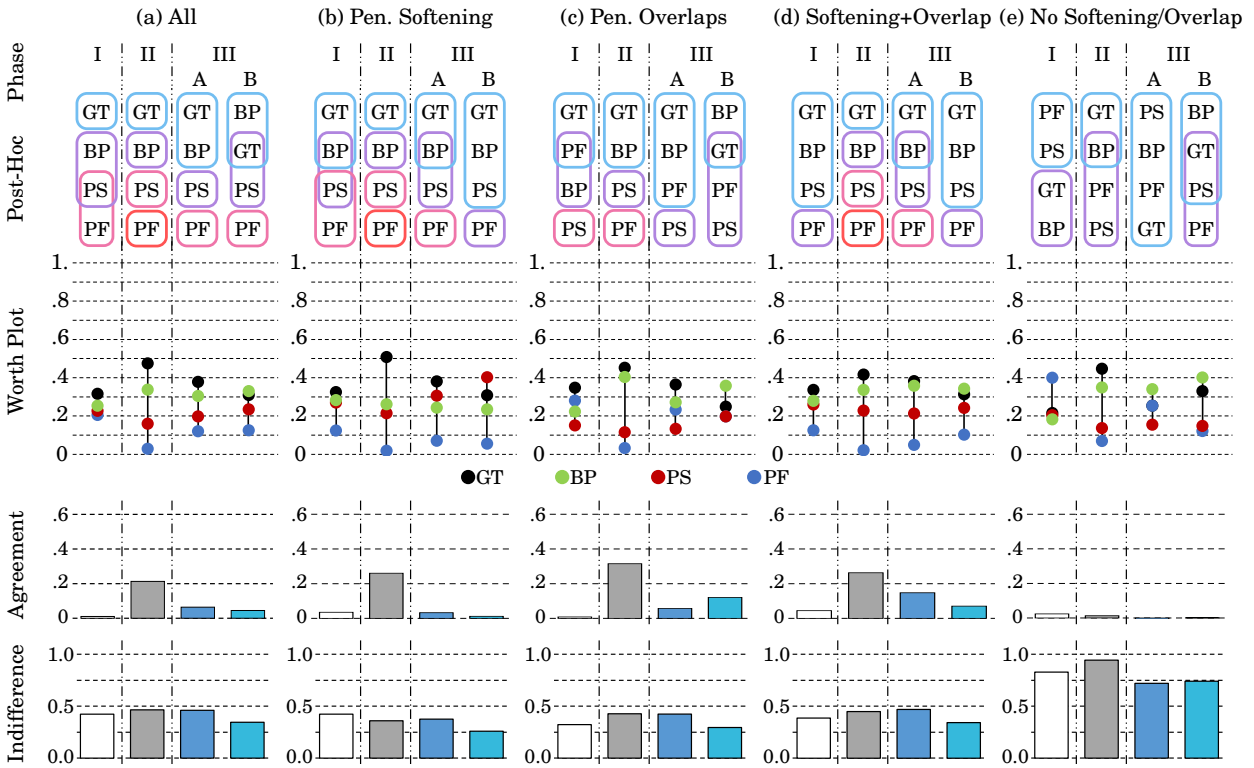

Fig. 7: The results of post-hoc testing and the worth plots for all scene categories (except vegetation), with (a) all penumbra categories pooled together, (b) penumbra softening only, (c) penumbra overlaps only, (d) both penumbra softening and overlaps, and (e) neither penumbra softening nor overlaps.

of differences between algorithms by means of a differential analysis. We will then discuss the results for the game scenes (Section 8.3) and the vegetation scenes (Section 8.4) in more detail, since both turned out to constitute cases which raise new questions to be answered in future work (Section 8.4). In Section 8.5 we then relate the observations for normal users to the results of experts.

8.1 Effect of User Experience

First we analyze the characteristics of the three phases of user experience according to Figure 7a (i.e., the combined results over all scene categories and penumbra categories):

8.1.1 *Phase I.* We expected that participants who are inexperienced with the task of judging soft shadows will perceive no or at best a subtle difference between soft shadows rendered with different algorithms. Supporting our initial expectation, there is a low agreement between participants (Figure 7a). However, the ordering of the worth values corresponds to the quality of the physical model of the respective algorithm, and post-hoc testing revealed differences between GT, BP, and PF (Figure 7a). This shows that even though inexperienced users exhibit a high uncertainty in their judgments, the results converge to a preference for more physically motivated soft shadows when many votes are collected. We therefore believe that users have a slight preference for more physically correct soft shadows which is rooted in intuition.

8.1.2 *Phase II.* This phase served as a training block where participants ought to learn to understand soft shadows. It also constitutes a corner-case scenario where users have the maximum knowledge about how the physically correct solution should look like. Hence, we expected the highest agreement and lowest indifference for this condition. Indeed, the agreement between users' preference votes was raised considerably from 0.01 to 0.21 (Welch's t-test[1] on agreement Phase I vs. Phase II: $t = 1.55$, $df = 13.6$, $p < 0.01$). To our surprise, however, we were not able to show that providing the ground truth yielded in a decrease of indifference (Welch's t-test on indifference Phase I vs. Phase II: $t = -0.66$, $df = 19.41$, $p = 0.74$) as can be seen in Figure 7a.

8.1.3 *Phase III.* Looking at the changes between Phase I and Phase III, we see an increase of agreement (Welch's t-test Phase I vs. Phase III: $t = 2.86$, $df = 75.87$, $p < 0.01$) together with more significant differences in the post-hoc results in Phase III (Figure 7a), which confirms a learning effect. Comparing the two groups of Phase III, we found that in general asking "what looks better" in Group A produced more significant outcomes during post-hoc testing compared to Group B (Figure 7a–7e). Additionally we observed that rankings in the worth plots of Group A better correlate with the physical plausibility of algorithms (Figure 7a–7e) and the differences between worth scores are more pronounced. Moreover, agreement scores are slightly higher in Group A (Figure 7a). Our intuition is that although asking users what looks better may cause more indifferent responses (Welch's t-test on indifference Phase III/A vs. Phase III/B: $t = 1.62$, $df = 19.82$, $p = 0.06$), this question seems to yield in more useful responses from users. We thus recommend for future studies about similar computer-graphics related topics to ask for an *emotional response* like "what looks better", because this type of question yields higher agreement and more significant results.

8.2 Differential Analysis

The scenes used for this experiment were designed in order to control the two penumbra phenomena we believed to mostly influence the perception of soft shadows (also shown in Figure 2), 1) the presence of penumbra softening, 2) the occurrence of penumbra overlaps. In Figure 7b–7e, we show the results pooled by the four respective penumbra categories.

As expected, the last category, where 1) and 2) are absent (Figure 7e), yields less useful results, having low agreement, high indifference, few significant differences, and a less meaningful ranking of algorithms. Thus, we continue to focus our analysis on scenes where phenomena 1), 2), or both occur (Figure 7a–7d). As shown in Table I, each algorithm has a different set of capabilities (penumbra size variation, physically-based penumbra, and physically-correct penumbra) which we believed to become particularly apparent when 1), 2), or both phenomena are present in a soft shadow.

To study whether users perceive differences in the way algorithms handle penumbra phenomena, we ask three representative questions, and formulate the conditions that have to be met for the answer to be positive. We then check the post-hoc results (Figure 7a–7e) whether a condition is met with respect to a penumbra category and summarize the outcome in Table II (using "✓" if the conditions are met). In particular, we look for an answer to the following 3 questions:

When do users prefer methods that model penumbra size variation? If GT, BP, and PS (all supporting penumbra size variation) are preferred over PF, we can conclude that this is in fact the case. Since the ability to render a varying penumbra becomes most visible through penumbra softening, we expected penumbra size variation to have a stronger effect when *penumbra softening* is present. This can be confirmed by looking at Table IIa (there are only "✓" in rows where penumbra softening occurs) and worth parameters in Figure 7 (the predicted probability to select PF is lower

[1]We used Welch's t-test, because the equality of variances asumption was violated.

Table II. : Analysis of the penumbra perception with respect to a particular penumbra category, by asking if users (I: inexperienced; II: cognizant; III: experienced) prefer a) penumbra size variations, b) physically-based penumbrae, c) physically-correct penumbrae. If a positive answer is supported with significance in post-hoc results of Fig. 7, the slot is marked with a "✓".

	(a) penumbra size variation				(b) phys.-based penumbra				(c) phys.-correct penumbra			
	I	II	III A	III B	I	II	III A	III B	I	II	III A	III B
All Scenes (Fig. 7a)	–	✓	✓	✓	✓	✓	✓	✓	✓	✓	–	–
Pen. Softening (Fig. 7b)	–	✓	✓	✓	✓	✓	✓	–	–	✓	–	–
Pen. Overlaps (Fig. 7c)	–	–	–	–	–	✓	–	✓	–	–	–	–
Softening+Overlaps (Fig. 7d)	✓	✓	✓	✓	–	✓	✓	–	–	✓	–	–
No Softening/Overlaps (Fig. 7e)	–	–	–	–	–	✓	–	–	–	–	–	–

when penumbra softening is present). For such situations, we find that in our experiment cognizant (Phase II) and experienced users (Phase III) significantly prefer algorithms that model penumbra size variation, while inexperienced users (Phase I) do not seem as sensitive to this. However, we find a significant difference between {GT,BP,PS} and PF for inexperienced users when overlaps occur in addition to penumbra softening ("Softening+Overlaps"). We conclude that missing penumbra size variation is more noticeable to inexperienced users if softening occurs in combination with overlaps.

When do users prefer a physically-based method? For this to be true, we expect that GT or BP, which are both based on a physical model, will be preferred over the empirical algorithms PS and PF. We also check that neither algorithm is worse than PS or PF. Looking at Table IIb, we observe that cognizant users (Phase II) are able to distinguish physically-based methods very well in our experiment. For inexperienced users (Phase I), a significant difference can be found for scenes with penumbra softening and for all scenes combined. On top of that, experienced users (Phase III) are also more sensitive when penumbra overlaps (Group B) and penumbra softening in combination with penumbra overlaps occur (Group A). Thus, we conclude that a physical model is preferred by all user groups because it supports penumbra softening, and that experienced users have learned to recognize other cues to whether a physically-based method has been used or not.

When do users prefer physically correct penumbrae? For this to be the case, we expect that the GT is preferred over all real-time algorithms (PF, PS, BP). Looking at the results of Figure 7 and Table IIc, we find that for cognizant users (Phase II) physically correct penumbrae pay off in all situations where penumbra softening is present. Surprisingly, when combining all scenes inexperienced users (Phase I) seem to be more sensitive than experienced ones (Phase III). Although this appears to be counterintuitive, we assume that the increased preference for a physically-based method like BP (Table IIb) may have reduced the perceived difference between BP and the GT. Overall we conclude that it is difficult to pinpoint a single penumbra feature which allows recognizing the often subtle differences to GT (e.g, with respect to BP), although there seems to be at least some intuition for it.

8.3 Game Scenes

We used this type of scene to add one less controlled but more representative and ecologically valid case with a high amount of visual detail, including textured shadow receivers. Since we used only two similar scenes for this category, the results are probably too specific to generalize for other complex scenes. Our intention was rather to gain a first clue how visual complexity and textures can affect users' perception, which should serve as a hint for future work.

Since textures may mask features of a shadow [Ferwerda et al. 1997], we expected a lower sensitivity for differences in shadows for all user categories. This expectation appears to hold true for the case of an inexperienced user (Phase I), where our experiment did not yield in statistical differences in the

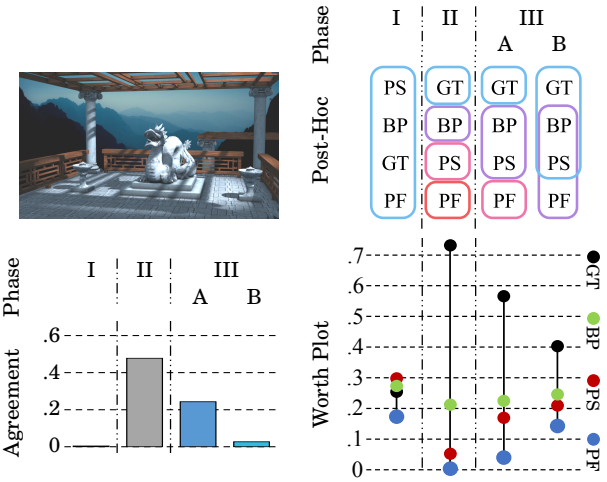

Fig. 8: Results for the game scenes.

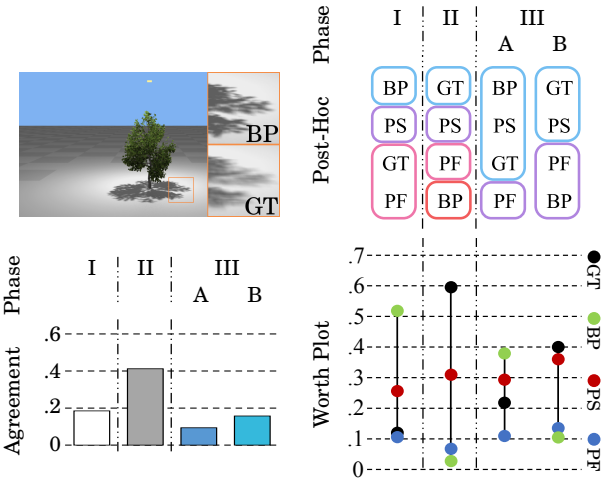

Fig. 9: Comparison of BP and GT and accompanying results for vegetation scenes.

post-hoc results of Figure 8. However, against our expectations, in Phase II and III/A the agreement scores and contrasts between worth values are exceptionally high compared to the results in Figure 7a. It hence seems like experienced users of Group A (Phase III/A) are very sensitive to the difference between a physically correct (GT) and an approximate physically-based solution (BP) in the game scenes (see post-hoc results of Figure 8). However, due to the variety of uncontrolled factors potentially influencing a user's perception, it is difficult to infer an explanation.

Another interesting observation is the pronounced difference between Group A and B in Phase III. The predicted probability for an experienced user of Group A to prefer GT is 17% higher, while the probability to prefer PF is 10% lower than for Group B (see worth plot in Figure 8). This is also reflected in the results of post-hoc testing. It indicates that asking for aesthetic properties in Group A encouraged a more accurate judgment strategy, probably because the users were less concerned and hence less distracted by complex details.

8.4 Vegetation Scenes

The trees in the vegetation scenes are complex occluders, and our expectation was that the huge number of randomly arranged leaves and branches overstrain a user's ability to understand the shadow being cast on the ground. Nevertheless, we find, particularly in Phase I, an above-average agreement and a more pronounced contrast in worth values than for other scenes discussed before (Figure 9). Due to the order of the worth values, which contradicts in Phases I and III how physically-based the algorithms are, these results have to be analyzed with utmost care.

One reason for this outcome is that it was practically impossible to configure the BP method such that artifacts resulting from gap filling were fully avoided. Interestingly, however, these artifacts can also be interpreted as additional details which improve the shadow appearance. In Phase III, where we have a strong contradiction between Group A and B, the worth values of BP indicate that users who judge "what looks better" (Group A) seem to prefer the artifacts of BP, while users who judge physical plausibility (Group B) seem to identify artifacts as errors which reduce physical plausibility.

We think that this result at least reveals that the perception of artifacts is a very important and interesting issue, particularly for light interaction in vegetation and other very complex scene config-

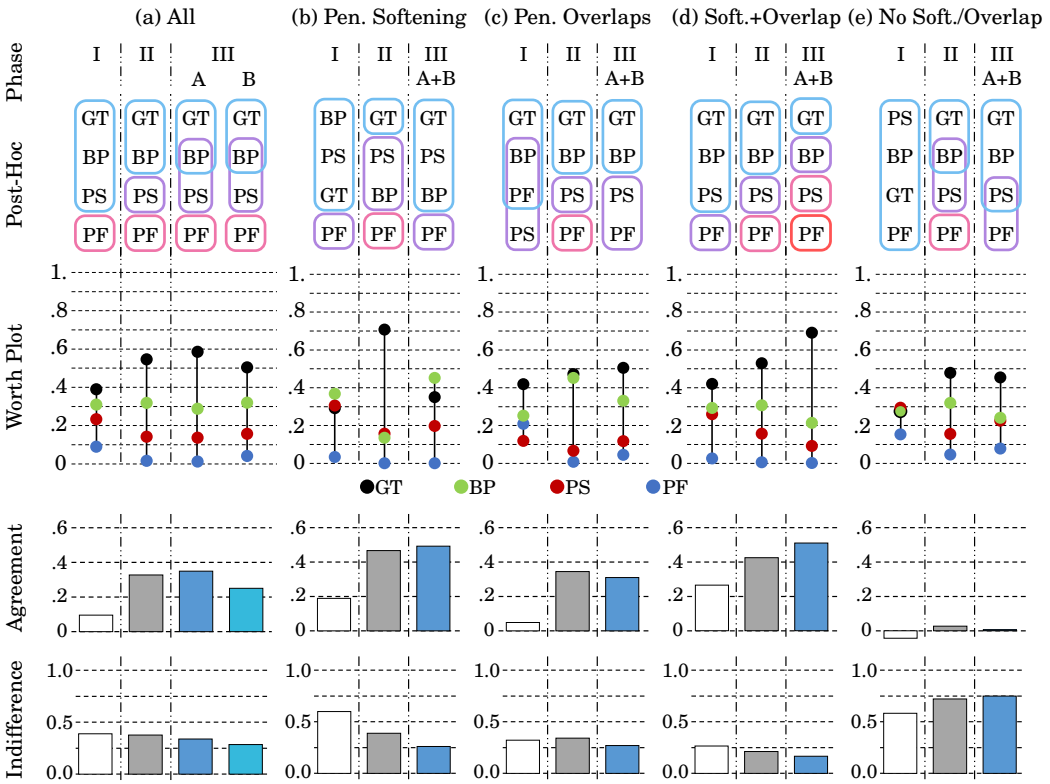

Fig. 10: The results of expert users for all scene categories except vegetation (compare with Figure 7).

urations. Masking and perceived noisiness occurring in shadow patterns of trees could be exploited to render shadows in vegetation scenes much more efficiently, e.g., by faking the appearance of the patterns through noise synthesis.

Another explanation for the unexpected result when asking about "what looks better" is that participants might not be used to vegetation (a predominantly outdoor phenomenon) being lit by an area light source that casts large penumbrae. Instead they may have expected shadows that resemble those produced by sunlight. Thus the results indicate that the validity of evaluating soft-shadow algorithms with vegetation scenes, although very common since they constitute a challenging case [Guennebaud et al. 2006], should be reconsidered.

8.5 Comparison With Expert Users

The expert group performed the experiment under identical conditions as normal users tested previously. The results for this group were analyzed separately and are shown in Figure 10. Please note, that we expect post-hoc testing results to be overly conservative. Due to the relatively low sample size (8 participants) the statistical power is low and thus the probability for type II errors (i.e., falsely accepting H_0 "there is no difference") is high. Due to the issue of sample size, we decided to merge the data of Group A and B for the results shown in Figures 10b–10e where data was split by penumbra category. However, the effect of the two groups can be seen in the overall results depicted in Figure 10a, where

we kept the data for both groups separated. This figure shows that we again observe that Group A votes with higher agreement and the contrast in the worth values is also slightly higher for this group.

Overall, the results demonstrate that indeed expert users can distinguish correct soft shadows better than normal users. This becomes apparent in the agreement scores which are higher in the expert group with a weak significance (Welch's t-test on agreement experts vs. users: $t = 1.55$, $df = 11.49$, $p = 0.07$). Expert users start in Phase I with an agreement twice as high as for experienced users (i.e., Phase III from the non-expert experiment). An interesting observation is that experts could also significantly improve their judgment skills through training with a reference in Phase II (Welch's t-test on agreement Phase I vs. Phase III: $t = 2.34, df = 19.90, p = 0.01$). After this "warm-up", experts of Group A vote with similarly high agreement (0.33 in Phase II compared to 0.35 in Phase III/A), and worth values are also quite similar.

Besides comparing expert and normal users, we were also interested which penumbra properties are important for expert users to judge soft shadows. When penumbra softening is the only shadow cue, experts clearly prefer a varying penumbra since PF is rated significantly lower than all other methods (Figure 10b post-hoc). The difference between physically-based methods and empirical models (PS) can be best identified in the presence of overlaps. However, only after the training received in Phase II this becomes statistically significant – presumably due to the small number of expert participants.

If trained experts see shadows with both overlaps and penumbra softening, the results show the highest agreement together with the most plausible rating of the algorithms. In this condition, we observe significant results proving that experts are able to correctly distinguish correct from approximate physical simulations. In the absence of penumbra softening and overlaps, even experts have a very low agreement in their votes and the majority of responses are indifferent.

9. SUMMARY AND CONCLUSIONS

In this work, we investigated how non-expert users perceive soft shadows and to which extend they are sensitive to simplifications in the simulation of this phenomena. We did this by studying user preferences with the method of pairwise comparison on a set of four different soft-shadow simulation methods. The methods were selected such that they represent four distinct quality levels.

Since soft shadows are a global lighting phenomenon, casting a preference vote requires a user to analyze the interaction of light source, shadow caster and shadow receiver. We expected that understanding of soft shadows is an ability that can be learned by normal users and hence used the three phases in this study to constitute three corner cases of user experience. Additionally we included experts in the study and compared their results with non-experts.

Our *main conclusions* from the results of this study are the following:

Inexperienced users have an intuition to prefer correct soft shadows. Regardless of the low agreement, surprisingly the results converge to a significant preference for the physically correct model.

Users can learn to understand features of physically-based soft shadows. Experienced and cognizant users prefer algorithms that can produce locally varying penumbrae over those without this capability, and physically-based solutions over faked soft shadows when dissimilar penumbrae overlap.

Asking about "what looks better" provides more significant results. We found that it is favorable to ask which shadow looks better (Group A) than which shadow is physically more plausible (Group B), because we obtained more significant and meaningful results for both non-experts and experts.

Take-away messages for game designers. As artifact suppression for BP requires a lot of fine-tuning and implementation effort, simpler methods like PS might be a good compromise if a game does not strive for high realism. For high-profile games on the other hand, it might pay off to put more resources into shadow computations, since experienced users prefer *phys. based* shadows like BP.

Take-away messages for researchers. The surprisingly clear preference for the ground truth (GT) suggests that there is a practical potential for further improving the best real-time methods in terms of correctness, along with robustness and efficiency, even for the benefit of inexperienced users that only *casually* encounter soft-shadow approximations.

10. FUTURE WORK

We see this work as a starting point to explore the complex, multi-dimensional space that constitutes all aspects of soft shadows. Some results, e.g., those for textured game scenes, indicate that there are other not yet investigated factors which may influence a user's perception. Based on the proposed study design framework, we plan to study factors like rendering artifacts, masking effects of textures, visual complexity in vegetation shadows, the role of display devices, and of course the influence of animation.

An interesting question is whether animations constitute cases where users are more sensitive to wrong shadows or whether the opposite is true, since it has been observed by others that movement actually blurs the human sensitivity to shadows and thus a lower quality can be used for dynamic shadows [Scherzer 2009]. Our intuition is that it is important to study complex animations, such as used for leaves and trees, for some surprising new insights into shadow perception.

Furthermore, it would be interesting to study perception of shadows cast by multiple lightsources and how light sources which vary in shape and size relative to each other would affect penumbra perception in such a scene.

And finally, as a general preference for physically-based models has been observed, it would be interesting to extend our study to methods which fill the gap between single-sample real-time methods like BP and the ground-truth method.

Acknowledgements

We would like to thank Michael Schwarz for kindly providing us his code for the BP algorithm and Anita Mayerhofer-Sebera for her help during the organization of the study. Finally, we are grateful to the anonymous reviewers for their valuable suggestions and comments.

REFERENCES

AKYÜZ, A. O., FLEMING, R., RIECKE, B. E., REINHARD, E., AND BÜLTHOFF, H. H. 2007. Do hdr displays support ldr content?: a psychophysical evaluation. *ACM Trans. Graph. 26,* 3.

ATTY, L., HOLZSCHUCH, N., LAPIERRE, M., HASENFRATZ, J.-M., HANSEN, C., AND SILLION, F. X. 2006. Soft shadow maps: Efficient sampling of light source visibility. *Computer Graphics Forum 25,* 4, 725–741.

BAVOIL, L., CALLAHAN, S. P., AND SILVA, C. T. 2008. Robust soft shadow mapping with backprojection and depth peeling. *Journal of Graphics Tools 13,* 1, 19–30.

ČADÍK, M., HERZOG, R., MANTIUK, R., MYSZKOWSKI, K., AND SEIDEL, H.-P. 2012. New measurements reveal weaknesses of image quality metrics in evaluating graphics artifacts. *ACM Trans. Graph. 31,* 6, 147:1–147:10.

DAVIDSON, R. AND BEAVER, R. 1977. On extending the bradley-terry model to incorporate within-pair order effects. Biometrics Series, vol. 33. International Biometric Society, 693–702.

EISEMANN, E., SCHWARZ, M., ASSARSSON, U., AND WIMMER, M. 2011. *Real-Time Shadows*. A.K. Peters.

FERNANDO, R. 2005. Percentage-closer soft shadows. In *ACM SIGGRAPH Sketches*.

FERWERDA, J. A., SHIRLEY, P., PATTANAIK, S. N., AND GREENBERG, D. P. 1997. A model of visual masking for computer graphics. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*. SIGGRAPH '97. ACM Press/Addison-Wesley Publishing Co., New York, NY, USA, 143–152.

GUENNEBAUD, G., BARTHE, L., AND PAULIN, M. 2006. Realtime soft shadow mapping by backprojection. *Computer Graphics Forum*, 227–234.

GUTIERREZ, D., SERON, F. J., LOPEZ-MORENO, J., SANCHEZ, M. P., FANDOS, J., AND REINHARD, E. 2008. Depicting procedural caustics in single images. *ACM Trans. Graph. 27,* 5, 120:1–120:9.

HATZINGER, R. AND DITTRICH, R. 2012. prefmod: An R package for modeling preferences based on paired comparisons, rankings, or ratings. *Journal of Statistical Software 48,* 10, 1–31.

HU, H. H., GOOCH, A. A., THOMPSON, W. B., SMITS, B. E., RIESER, J. J., AND SHIRLEY, P. 2000. Visual cues for imminent object contact in realistic virtual environment. In *Proceedings of the conference on Visualization'00*. IEEE Computer Society Press, 179–185.

JARABO, A., EYCK, T. V., SUNDSTEDT, V., BALA, K., GUTIERREZ, D., AND O'SULLIVAN, C. 2012. Crowd Light: Evaluating the Perceived Fidelity of Illuminated Dynamic Scenes. *Computer Graphics Forum 31,* 2, 565–574.

KENDALL, M. AND GIBBONS, J. D. 1990. *Rank Correlation Methods* 5 Ed. A Charles Griffin Title.

KNILL, D. C., MAMASSIAN, P., AND KERSTEN, D. 1997. Geometry of shadows. *Journal of the Optical Society of America A 14*, 3216–3232.

LEDDA, P., CHALMERS, A., TROSCIANKO, T., AND SEETZEN, H. 2005. Evaluation of tone mapping operators using a high dynamic range display. *ACM Trans. Graph. 24,* 3, 640–648.

MADISON, C., THOMPSON, W., KERSTEN, D., SHIRLEY, P., AND SMITS, B. 2001. Use of interreflection and shadow for surface contact. *Perception and Psychophysics 63,* 2, 187–194.

PAULY, M., MITRA, N. J., WALLNER, J., POTTMANN, H., AND GUIBAS, L. J. 2008. Discovering structural regularity in 3d geometry. *ACM Trans. Graph. 27,* 3, 43:1–43:11.

REEVES, W. T., SALESIN, D. H., AND COOK, R. L. 1987. Rendering antialiased shadows with depth maps. *SIGGRAPH Comput. Graph. 21,* 4, 283–291.

RUBINSTEIN, M., GUTIERREZ, D., SORKINE, O., AND SHAMIR, A. 2010. A comparative study of image retargeting. *ACM Trans. Graph. 29*, 160:1–160:10.

SATTLER, M., SARLETTE, R., MÜCKEN, T., AND KLEIN, R. 2005. Exploitation of human shadow perception for fast shadow rendering. In *Proceedings of the 2nd symposium on Applied perception in graphics and visualization*. APGV '05. ACM, New York, NY, USA, 131–134.

SCHEFFE, H. 1952. An analysis of variance for paired comparisons. *Journal of the ASA 47,* 259, 381–400.

SCHERZER, D. 2009. Applications of temporal coherence in real-time rendering. Ph.D. thesis, Institute of Computer Graphics and Algorithms, Vienna University of Technology.

SCHWARZ, M. AND STAMMINGER, M. 2008a. Microquad soft shadow mapping revisited. In *Eurographics Annex to the Conference Proceedings (Short Papers)*. 295–298.

SCHWARZ, M. AND STAMMINGER, M. 2008b. Quality scalability of soft shadow mapping. In *Proceedings of graphics interface*. GI '08. 147–154.

SCHWÄRZLER, M., LUKSCH, C., SCHERZER, D., AND WIMMER, M. 2013. Fast percentage closer soft shadows using temporal coherence. In *Proceedings of the ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*. I3D '13. ACM, New York, NY, USA, 79–86.

SETYAWAN, I. AND LAGENDIJK, R. L. 2004. L.: Human perception of geometric distortions in images. In *Proceedings of SPIE, Security, Steganography and Watermarking of Multimedia Contents VI*. SPIE.

STEEL, R., TORRIE, J., AND DICKEY, D. 1997. *Principles and procedures of statistics: a biometrical approach*. McGraw-Hill series in probability and statistics. McGraw-Hill.

VANGORP, P., DUMONT, O., LENAERTS, T., AND DUTRÉ, P. 2006. A perceptual heuristic for shadow computation in photo-realistic images. In *ACM SIGGRAPH Sketches*.

WANGER, L. 1992. The effect of shadow quality on the perception of spatial relationships in computer generated imagery. In *Proceedings of the 1992 symposium on Interactive 3D graphics*. I3D '92. ACM, New York, NY, USA, 39–42.

WANGER, L. C., FERWERDA, J. A., AND GREENBERG, D. P. 1992. Perceiving spatial relationships in computer-generated images. *IEEE Comput. Graph. Appl. 12*, 44–51, 54–58.

YANG, B., FENG, J., GUENNEBAUD, G., AND LIU, X. 2009. Packet-based hierarchal soft shadow mapping. *Computer Graphics Forum 28,* 4, 1121–1130.

YU, I., COX, A., KIM, M. H., RITSCHEL, T., GROSCH, T., DACHSBACHER, C., AND KAUTZ, J. 2009. Perceptual influence of approximate visibility in indirect illumination. *ACM Trans. Appl. Percept. 6,* 4, 24:1–24:14.

Received August 2013; revised November 2013; accepted February 2014