On Predicting Visual Popping in Dynamic Scenes

Michael Schwarz[1,2] * Marc Stamminger[2]

[1]Max Planck Institute for Informatics
[2]University of Erlangen-Nuremberg

Abstract

Popping is a major source of visual artifacts in dynamic scenes. To alleviate or avoid it, usually some temporal smoothing scheme is employed or levels of detail are chosen conservatively based on geometric deviation measures. In this paper, we consider the actual perceptibility of popping artifacts and its prediction. We first discuss several issues affecting popping perception, pointing out its complexity. Introducing some simplifying assumptions, we then present a practical perceptually-motivated predictor for popping. It makes heavy use of a spatio-velocity color vision model and aggregates the model output in a novel and useful way. We demonstrate the predictor's application to concrete examples, and discuss a conducted user study which indicates the validity of our approach.

CR Categories: I.3.7 [Computer Graphics]: Three-Dimensional Graphics and Realism

Keywords: popping, visual perception, temporal artifacts, vision models

1 Introduction

When rendering dynamic scenes and hence the image content changes over time, several artifacts can arise in the temporal domain. Apart from aliasing and its most disturbing manifestation, flickering, many of these artifacts are due to *popping*. Popping occurs if the renderer uses two different representations or parameter sets, referred to as levels of detail (LOD), for at least one scene entity in two consecutive frames and this results in an abrupt change that gets noticed by the user. Often, such switches are adaptations of an object's geometric LOD [Luebke et al. 2002], which in general not only influences outer and inner silhouettes but also shading. Popping may also be caused by transitioning from a geometric to an image-based or a point-based representation and vice versa, or by updating an impostor once parallax error or sampling density mismatch exceeds a threshold [Schaufler 1995]. Other examples comprise changing the complexity of employed shaders [Pellacini 2005] and picking a different set of virtual point lights for approximating indirect illumination.

Many techniques have evolved over time to avoid popping or at least alleviate its severity. One class combats popping by smoothing the transition over several frames, with image-space blending [Giegl and Wimmer 2007] and geomorphing [Hoppe 1996] being the most notable approaches. On the downside, however, they introduce additional overhead, often countering the motivation for changing the LOD, increasing performance. Even more problematic, these approaches may cause the transition itself or some of the intermediate states to be perceived as unnatural or even disturbing, essentially trading one problem for a different one.

Another well-adopted option is to utilize a deviation metric and only switch LOD if the predicted deviation stays below a threshold considered acceptable. For geometric LOD, many of these metrics operate on bounds derived in object space and then project them into screen space. The resulting error bound is often considered appropriate in order to avoid popping if it is at most half a pixel. Examples include the texture deviation metric [Cohen et al. 1998] for mesh simplification and the geometric approximation error for adaptive tessellation of higher-order surfaces [Guthe et al. 2005]. Most of these metrics only consider an object's geometry and basically ignore its surface signal. Consequently, they may be too conservative because, for instance, even a large texture space distortion often remains imperceptible if the texture is uniformly colored or in shadow. On the other hand, small changes in geometry may result in major shading variations like jumping or vanishing highlights.

To overcome some of these shortcomings, several researchers suggested metrics motivated by perceptual considerations. For instance, the contrast and the change in spatial frequency content induced by a geometric LOD change were employed to predict visual detectability, also accounting for texture content [Williams et al. 2003]. Using a more involved and computationally expensive vision model, the visual masking potential of an object's surface signal may additionally be regarded [Qu and Meyer 2006]. Unfortunately, such approaches mainly target a best-effort simplification and don't directly account for popping. On the other hand, Hamill et al. [2005] conducted psychophysical experiments for models of buildings and humans, deriving thresholds for the pixel-to-texel ratio at which a change from impostor to geometric representation can be carried out without (disturbing) popping.

In most cases, a change potentially causing popping is executed because the affected scene entity is moving relative to the viewer. Depending on how fast and in which direction an entity moves, the perceptibility of the switch it is subjected to can significantly differ. However, this temporal aspect of popping is basically ignored by all metrics for choosing an appropriate LOD. Although few approaches exist which take object movement into account to select coarser geometric LODs for fast-moving objects [Reddy 1997], they don't consider the switch among two LODs.

Because of the practical importance of popping and the absence of reliable solutions which are not over-conservative, there is a certain need for a computational model for predicting whether and where popping occurs in dynamic scenes. One potential application is the derivation of optimal LOD transition points for prerecorded paths in walk-through and fly-over scenarios. Moreover, such an automatic metric may serve as oracle when optimizing parameters or testing LOD schemes. It could also help identifying screen regions where popping is likely to be perceivable. Note, however, that it is not intended for a per-frame on-the-fly application to guide LOD selection in real-time rendering settings.

In this paper, we propose a first solution towards the elusive goal of reliably predicting popping.

- First, we review and discuss several aspects involved in per-

*e-mail: mschwarz@mpi-inf.mpg.de

ceiving popping, highlighting the complexity of this phenomenon and why a reliable prediction is extremely hard to achieve (Sec. 2).

- Second, we present a perceptually-motivated predictor for the perceptibility of popping artifacts which tackles some of the involved issues (Sec. 3). Our approach makes some simplifying assumptions and hence only targets a certain but important class of transitions. We describe the employed vision model and introduce a novel and meaningful way of condensing the model output, yielding popping regions.
- Third, we applied our predictor to concrete examples (Sec. 4) and conducted two experiments within a user study (Sec. 5) to evaluate the predictor's performance. The results indicate that our approach makes predictions which are well in line with the subjects' perception.

Although this work presents some promising first steps, it is still far from offering a complete and general solution. We believe, however, that it helps highlighting several of the involved challenges, and hope that it spurs further research in this important topic.

2 Aspects of perceiving popping

The perception of popping turns out to be a very complex phenomenon that is influenced by several factors, many of which are far from being completely understood. In general, popping is perceived if a temporal discontinuity in the image signal occurs that is large enough to be captured by the human visual system (HVS) and that is then actually detected by the viewer.

Consequently, attention plays a significant role in perceiving popping. Even strong popping may go unnoticed if the viewer's attention is not directed towards the region where it occurs. There is experimental evidence [Schütz et al. 2007] that in case a moving object is pursued, the attention is both focused on this target and its direction, causing a loss of sensitivity for both peripheral objects and motion opposite to the pursuit direction. Attention is guided by both a bottom-up process, which is stimulus-driven and attracted by salient image features, and a top-down process, which is task-dependent. Popping itself may be highly salient and hence attract attention; however this is mainly true for large-scale popping involving multi-pixel geometric deviations which can usually easily be identified by classic non-perceptual metrics. While computational models for visual attention exist [Itti and Koch 2001], incorporating the viewer's experiences and identifying and modeling her adopted task remains a challenge.

Motion perception involves higher-level visual mechanisms and depends partly on more abstract image features like surfaces and objects [Wandell 1995]. Motion introduces some spatial uncertainty about the future location of such features, and may also occlude and reveal scene elements, which constitutes another source of uncertainty. Moreover, the HVS has only limited resources and hence each of its receptive fields is sensitive to a range of spatial and temporal frequencies, causing an uncertainty in measuring spatio-temporal signals [Gepshtein et al. 2007]. Spatial and temporal integration is performed, i.e. each receptive field computes a kind of weighted average of local image signals over a small space-time region, effectively causing a blurring [Wandell 1995]. When the motion flow field is processed, these uncertainties are factored in and the higher-level feature information is taken into account and gets updated. In case of inconsistencies of or temporal discontinuities in the flow field, popping may be detected.

Concerning vision, the sensitivity of contrast detection and discrimination shows both intra- and inter-observer variations. and degrades with age [Hardy et al. 2005]. Hence, like with most perception-based aspects, a perfect prediction that applies to everybody is impossible. Only if the sensitivity is high enough, a luminance change or a chromatic shift due to popping can be noticed. Regarding modeling this sensitivity, the higher the desired accuracy, the more dependencies have to be considered. However, too many parameters make a model hard to apply as several parameter values are difficult to provide. Moreover, experimental data is usually only acquired for a small number of parameters.

Other artifacts like aliasing and in particular flickering can also influence the perception of popping. Not only may they attract the viewer's attention and hence divert it from a region where popping occurs, but they may also mask the actual popping. That is, despite noticing the popping, it is not perceived as popping but attributed to another artifact.

Finally, the display device impacts popping perception. Most notably, the now ubiquitous LCD displays typically suffer from motion blur, mainly because of the employed sample-and-hold technique but also due to their response times [Pan et al. 2005]. While techniques like flashing backlights are able to alleviate this problem [Feng 2006], they are not widely utilized yet. Other display characteristics like the chosen white level, as well as light reflected off of the screen influence visual sensitivity, thus affecting the perceptibility of popping artifacts, too.

3 Perceptually-motivated popping predictor

Considering all the factors involved in popping perception, an accurate prediction appears to be a goal extremely hard to achieve. In particular, higher-level mechanisms play an important role but are challenging to account for. While their influence might be modeled in general, doing so reliably at the pixel level essentially is an open problem presumably requiring a lot of further vision research.

To make the prediction task more tractable, we hence introduce several simplifying assumptions. Most notably, we ignore temporal integration and consider only the single frame where the popping-prone switch of LOD occurs. To detect temporal discontinuities, we compare the actually rendered frame against a prediction of what the user might expect for this frame by means of a vision model. Differences above a certain magnitude then indicate popping. The predicted frame content is obtained by rendering the frame again but utilizing the previous LOD. We hence assume that this way of extrapolating the image content and motion of the previous frames is a good enough approximation for identifying perceived temporal discontinuities resulting in popping artifacts.

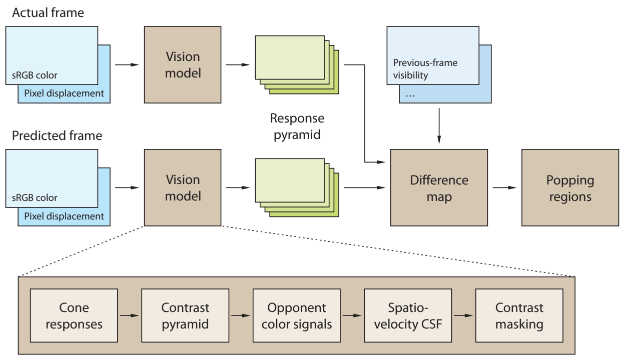

Figure 1: *Overview of our popping predictor.*

An overview of our predictor is shown in Fig. 1. As input both the actually rendered frame and the predicted frame are provided. Each frame input comprises a color image in sRGB space and a map storing the screen-space displacement of each pixel center with respect to the previous frame. The frame data is subjected to a color vision model, detailed in Sec. 3.1, which takes retinal velocity derived from the pixel displacement into account. The model outputs a contrast response pyramid, with its levels corresponding to the spatial frequency decomposition performed by the vision model. Next, the pixel-wise difference between the two input frames' response pyramids is determined across levels and color channels, yielding a difference map. During its computation, additionally provided input for down-weighting differences, like the previous-frame visibility of each pixel which allows identifying disoccluded pixels, is processed. Finally, connected regions where popping may be perceived, referred to as *popping regions*, are extracted. The whole model output aggregation scheme and its predictive utility are further elaborated on in Sec. 3.2.

Discussion Although higher-level visual mechanisms are not explicitly modeled due to their complexity, the rather simple approach of comparing the actual with the predicted frame accounts for them to a certain degree by indirectly factoring in shape and shading information. Nevertheless our approach is clearly not appropriate in all cases. For instance, regarding impostor updates, using the previous impostor texture usually doesn't correspond to the user's expectation of how the previous frame evolved; in contrast it will probably be a worse match than the new impostor texture due to its larger distortion. On the other hand, transitions from one geometric LOD to another one are rather well captured by our approximation. We believe that while such LOD switches which are amenable to our approach constitute only a subset of all popping-prone LOD changes, they still form a large class of practical importance.

Since we are not modeling most of the uncertainty involved in motion perception, our predictor is slightly too conservative and sometimes wrongly reports a temporal discontinuity which actually gets smoothed out by the visual system. For instance, imagine an object with a curved horizontal silhouette that is approaching the viewer, where every few frames the number of pixels in a scan line covered by the silhouette increases. If this increase is postponed by one frame, often no popping is perceived while the comparison of our input frames may suggest a popping artifact.

Even though attention is of high importance for perceiving popping, we are not accounting for it. Our algorithm just outputs screen regions where popping is predicted to be perceptible if attention is directed towards it. Note, however, that in principle these regions can easily be checked against the output of a computational attention model. Similarly, we refrain from regarding motion blur inherent to LCD displays, which may lead to some erroneously predicted popping artifacts.

3.1 Spatio-velocity color vision model

A computational vision model processes the visual input and yields a response that scales roughly with the perceptibility of the visual contrast stimuli. By comparing the responses for two different inputs, visual differences can be determined. A multitude of vision models were developed for static images, operating either only on luminance [Daly 1993; Lubin 1995] or also on color [Bolin and Meyer 1999; Pattanaik et al. 1998; Lovell et al. 2006]. Some models for dynamic images which in addition to luminance (but not color) take motion speed into account were also devised [Yee et al. 2001; Myszkowski 2002]. Our vision model is influenced by these approaches and extends them as required by our problem domain, while being comparably cheap to execute.

As input, the model expects a color image in sRGB space as well as a pixel displacement map. First, the color image is converted to absolute CIE XYZ tristimulus values, taking the display's black and white level luminances into account. Then a transformation to Hunt-Pointer-Estevez cone fundamentals is performed [Fairchild 2004]:

$$\begin{pmatrix} L \\ M \\ S \end{pmatrix} = \begin{pmatrix} 0.40024 & 0.70760 & -0.08081 \\ -0.22630 & 1.16532 & 0.04570 \\ 0.0 & 0.0 & 0.91822 \end{pmatrix} \begin{pmatrix} X \\ Y \\ Z \end{pmatrix}.$$

Next, we construct a Gaussian pyramid [Burt and Adelson 1983] with levels G_i, utilizing a binomial filter kernel of size 5×5. From this, a contrast pyramid is built which stores local band-limited contrast [Lubin 1995], i.e. level i is computed as $(G_i - G_{i+1})/G_{i+2}$, where coarser levels are appropriately upsampled. Subsequently, the contrast values are converted to Hunt's opponent color space [Fairchild 2004]:

$$\begin{pmatrix} A \\ a \\ b \end{pmatrix} = \begin{pmatrix} 2 & 1 & 1/20 \\ 1 & -12/11 & 1/11 \\ 1/9 & 1/9 & -2/9 \end{pmatrix} \begin{pmatrix} L \\ M \\ S \end{pmatrix}.$$

A represents the achromatic response; a and b correspond to the red–green and yellow–blue opponent signals, respectively.

The contrast pyramid is then normalized by multiplication with the spatio-velocity contrast sensitivity function (CSF). Since sensitivity for fast-moving contrast stimuli is often lower than for static ones, this stage accounts for the observation that visual differences leading to popping artifacts are usually harder to spot for moving objects. The employed CSFs, which are further detailed below, depend on the spatial frequency ρ (in cycles per degree, cpd), the retinal velocity v (in deg/s) and the local adaptation luminance L. For each contrast pyramid level, which essentially represents a spatial frequency band, we take its peak frequency for ρ. The raw velocity v_s is computed from the input pixel displacement map, using parameters of the viewing setup like viewing distance, screen size and resolution. It is then subjected to Daly's model [1998] of unconstrained eye movements, which accounts for the eye's tracking behavior, to obtain a conservative estimate of the retinal velocity:

$$v = \left| v_s - \min\{0.82\, v_s + 0.15 \text{ deg/s}, 80.0 \text{ deg/s}\} \right|.$$

Note that due to drift eye movements, the minimum retinal velocity is in general non-zero. Finally, the adaptation luminance is derived from the Gaussian pyramid level where one pixel roughly corresponds to one degree of visual field.

In a last step, we account for contrast masking by applying the transducer function

$$T(c) = \text{sign}(c) \cdot \frac{|c|}{\left(1 + (|c|^{0.3})^{10}\right)^{0.1}}$$

to the normalized contrast values c. Note that T converges to a simple power law for sub-Weber behavior at suprathreshold contrast levels [Legge 1981].

Achromatic CSF For the achromatic channel A, we employ Daly's refinement [1998] of Kelly's model [1979]

$$\begin{aligned} \text{csf}_{\text{A}}(\rho, v) = {} & \left(6.1 + 7.3\,|\log_{10}(c_2 v/3)|^3\right) c_0\, c_2 v \\ & \cdot (2\pi c_1 \rho)^2 \, \exp\left(-4\pi c_1 \rho (c_2 v + 2)/45.9\right) \end{aligned}$$

where $c_0 = 1.14$, $c_1 = 0.67$, $c_2 = 1.7$ for a luminance level of about 100 cd/m^2. As shown in Fig. 2, for increasing velocities the CSF's band-pass shape moves towards lower spatial frequencies and the peak sensitivity eventually drops.

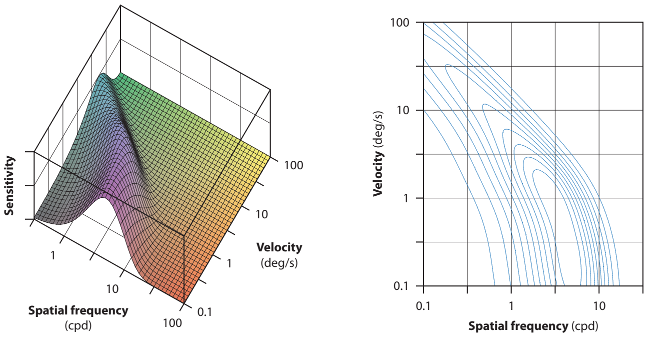

Figure 2: *Achromatic CSF at luminance level $L = 100$ cd/m^2.*

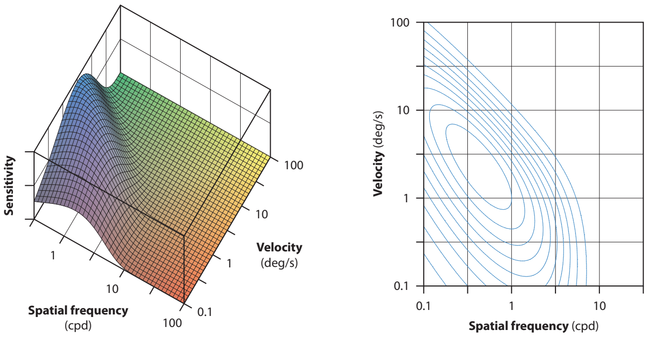

Figure 3: *Chromatic CSF.*

To model the CSF's dependence on the adaptation luminance level L, we resort to Barten's spatial CSF [Barten 2003]

$$\mathrm{csf_B}(\rho, L) = \frac{5200\, e^{-0.0016\rho^2(1+100/L)^{0.08}}}{\sqrt{(2+0.64\rho^2)\left(63/L^{0.83} + 1/(1-e^{-0.02\rho^2})\right)}},$$

adapting both the peak frequency scale factor

$$c_0(L) = 1.14 \cdot \frac{\max_\rho \mathrm{csf_B}(\rho, L)}{\max_\rho \mathrm{csf_B}(\rho, 100)}$$

and the spatial frequency scale factor

$$c_1(L) = 0.67 \cdot \frac{\arg\max_\rho \mathrm{csf_B}(\rho, 100)}{\arg\max_\rho \mathrm{csf_B}(\rho, L)}$$

which controls the shift of peak sensitivity along the frequency axis.

Chromatic CSF For both chromatic channels a and b, we adopted Kelly's spatio-temporal CSF [1983], which is a linear combination of two space-time-separable low-pass functions modeling a center and a surround component. With increasing velocity, the low-pass nature of the CSF becomes more pronounced and its peak sensitivity rises as depicted in Fig. 3.

Again, the CSF doesn't model dependence on the luminance level. However, experiments indicate that the threshold contrast increases proportionally to the square root of the retinal illuminance [van der Horst and Bouman 1969], with retinal illuminance being related to luminance by the pupil's area. To be consistent with the assumptions in Barten's luminance CSF used to adapt the achromatic CSF to varying light levels, we compute the pupil's diameter by Le Grand's approximation [1968]:

$$d(L) = 5 - 3\tanh(0.4 \log_{10} L).$$

The chromatic CSF is then scaled by the square root of the ratio of the retinal illuminances corresponding to L and the reference luminance (roughly 35 cd/m^2).

Contrast pyramid levels Each level of the contrast pyramid is tuned to a certain band of spatial frequencies, which results from subtracting two band-limited levels of the Gaussian pyramid. Note however that the repeated filtering with a fixed-size Gaussian and downsampling does not exactly yield the frequency response of Gaussian filtering with a spread being doubled every level. Therefore, and especially because the finest level of the Gaussian pyramid is only band-limited by the sampling frequency, the common assumption that the peak frequencies of the contrast pyramid levels halve with every coarser level is not true. Most notably, the finest level's peak frequency is almost four times as high as that of the second-finest level. In our implementation we account for both this irregularity as well as the amplitude loss due to filtering. We choose the number of levels such that the coarsest level has a peak frequency of at least 0.5 cpd, which results in a five-level contrast pyramid for our viewing setup.

3.2 Popping regions

For both the actual and the predicted frame, the vision model yields a contrast response pyramid. To derive visual differences, we subtract these pyramids and collapse the resulting difference pyramid, combining differences across levels and channels (A, a, b) by Minkowski summation with an exponent of 2.4 [Lubin 1995; Watson and Ahumada 2005]. The obtained difference map indicates for each pixel the probability of being able to detect a difference in units of just noticeable differences (JNDs).

Recall that factors like disocclusion introduce uncertainty and hence differences between the actual frame and its prediction that occur at pixels affected by such uncertainties are less likely to be detected. To account for this, we take a practical approach and scale down the corresponding values in the difference map by weights provided as additional input to our predictor. For instance, the previous-frame visibility of the current frame's pixel centers may constitute one such weight.

While a difference map is of certain utility itself, the contained information should be aggregated in a meaningful way for further analysis. Standard measures like number of pixel differences above threshold, maximum difference, average and variance are usually of limited use because they are too coarse-grained. We hence adopt a different approach which is based on the observation that not only difference magnitude but also spatial context is important for detection [Bonneh and Sagi 1998]. Intuitively, even smaller visual differences may be easily detected if the affected pixels are clustered together and cover a larger screen region. On the other hand, if a visual difference occurs at an isolated pixel, its magnitude must be rather large to spot the difference.

To model this, we first identify all pixels where the two input color images differ and the visual difference map reports a value of at least 2 JNDs. We then start growing regions around these seed pixels, successively considering all eight direct neighbor pixels and including those with visual difference values of again at least 2 JNDs. The empirically chosen threshold of 2 JNDs accounts for the fact that differences are harder to detect in complex images than in case of simple gratings (typically employed in vision experiments for determining sensitivity). This procedure finally yields a number of popping regions, identifying those parts of the image where popping artifacts can be expected. For each region, we acquire statistics like its size in number of pixels and determine the Minkowski sum (with an exponent of 2.4) of the visual difference values at its pixels. The magnitude of this sum is a good indicator of how severe a popping occurs in the region. If we further subject it to the empirically derived mapping $R(\Sigma) = \ln(0.375\Sigma)/\ln(2.25)$, we obtain a simple rating R where values of $R < 1$ predict rarely visible popping and $R > 3$ suggests easily detectable popping.

An in-context visualization of the popping regions colored according to their rating values (see Fig. 4 for an example) allows for fast identification of where popping artifacts of which degree can be expected. Moreover, the popping region information is well-suited for further automatic processing. For instance, given a screen region of high importance, possibly provided by a computational attention model, it could be checked whether any popping regions are located in this screen region and if so how many pixels they cover and what their rating is. Based on this, an informed decision whether the potential popping artifacts can be considered acceptable or not for the given application can automatically be made.

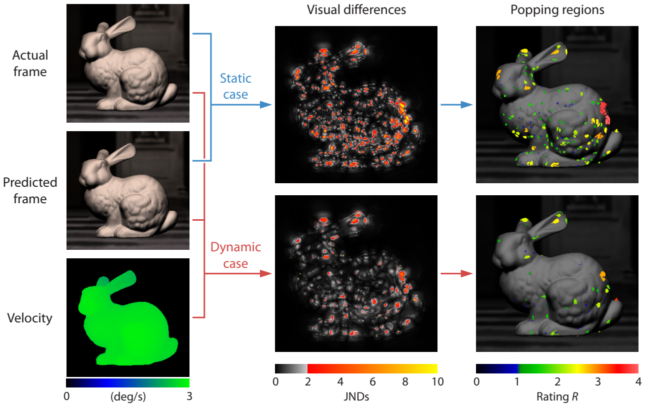

Figure 4: *The popping predictor applied to a concrete example. From the visual difference map, more meaningful popping regions are extracted.*

Fig. 4 shows a concrete example. Please recall that differences are harder to spot when viewing images side by side. Thanks to the selective aggregation performed, popping regions are a useful tool for analyzing the visual difference map and for identifying and rating popping artifacts. Moreover, note that the visual differences' magnitude is clearly affected by fast motion.

4 Application to real-world examples

We applied our popping predictor to two different examples chosen to be representative of possible real-world applications: an object-wise geometric LOD and a simple terrain LOD (see Fig. 5). In the first example, we constructed coarser concrete LODs via the progressive mesh implementation of Direct3D 9 and manually specified distances at which to switch LODs. To obtain the required input for our predictor, we render the frame in question twice, once with the new and once with the previous LOD. Apart from the pixel color, we further derive for each fragment the previous-frame screen location of its corresponding point and store the resulting screen-space displacement. To account for disocclusion, for both of the involved LODs we determine the depth map for the previous frame and perform a depth comparison with percentage-closer filtering to derive a fractional previous-frame visibility factor. Finally, being conservative, we take the pixel-wise maximum of these factors and provide the resulting weight map as further input to the predictor.

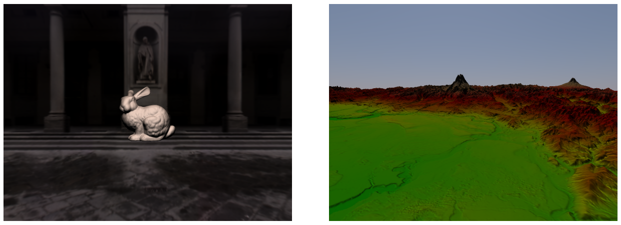

Figure 5: *Screenshots of two examples, object-wise geometric LOD and terrain LOD, to which we applied our predictor.*

For the terrain application, we adopted a simple chunk LOD approach [Ulrich 2002] where coarser-level terrain tiles are generated by regular subsampling. LOD switches are controlled by the screen-projected maximum height deviation from the finest-level terrain geometry. The input data for our popping predictor is obtained like in the first example.

However, since terrain fly-overs often suffer from strong flickering mainly at distant mountain ranges that can mask popping, we additionally incorporate a map for weighting down visual differences in flicker-affected regions. To detect flickering, we resort to a simple heuristic. For each pixel, we compute its shading for the previous-frame setup and compare it against the color obtained by sampling the previous frame's color image, taking depth discontinuities into account. If these two colors have a CIELAB ΔE^*_{94} [Fairchild 2004] difference value that is roughly as large as or even larger than the color difference between the two LOD renderings of the current frame, we assume flickering to occur, unless the pixel's inter-frame screen-space displacement is large.

5 User study

Given the simplifying assumptions and empirical choices made, we consider it important to conduct a user study to investigate the plausibility of our approach and its predictions. However, experimental validation turns out to be challenging for multiple reasons. In practice, a LOD change usually results in several popping regions. But because popping occurs at a single point in time, a subject can only spot and attend to at most one region (or maybe a few small and closely clustered ones), but misses processing all the other ones. Moreover, it is hard to determine where a subject directed its attention to. On the other hand, attention can only be guided to a certain degree and accuracy, especially in case of complex stimuli. Consequently, validating all predicted popping regions directly in ecological settings is an elusive task.

Another major obstacle is the huge space of possible LOD transitions that could result in popping, and its high dimensionality. In particular, perceptibility of popping artifacts is influenced by the involved objects (shape and its complexity, material), their environment (lighting, complexity), the LODs used, the chosen transition point (e.g. certain distance), and the kind, direction and speed of the object's movement relative to the camera. Therefore, any test necessarily has to concentrate on few samples of the LOD transition space.

We address these challenges by two different experiments. In the first one (Sec. 5.1), we seek to directly evaluate the predictive power of single popping regions. To this end, we focus on a simple object that allows directing a participant's attention to a specific region. Testing all combinations of two LOD sets, multiple transition points and two movement speeds while fixing all other degrees of possible variation, we densely sample a small subspace of all LOD transitions. In contrast, the second experiment (Sec. 5.2) deals with

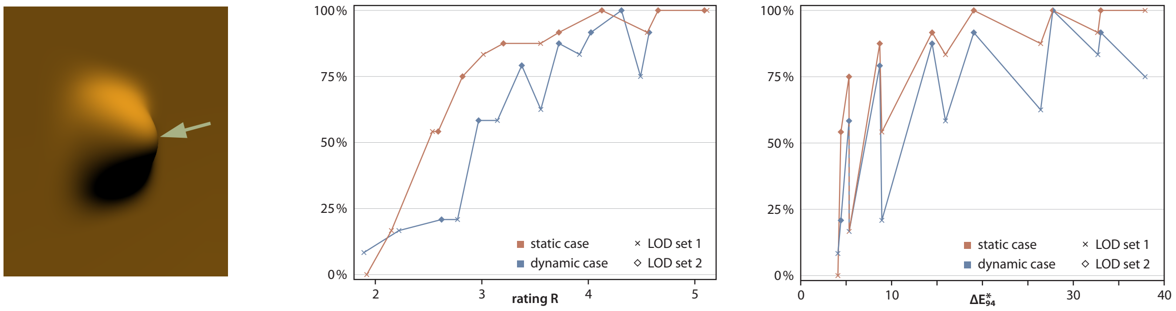

Figure 6: *Experiment I: Patch used, with arrow indicating the tip to which to attend for popping (left). The graphs show the average observed popping detection rate as a function of our predictor's rating R (center), and of the maximum CIELAB ΔE^*_{94} value (right).*

a larger subspace of the LOD transition space, but samples it only sparsely. Utilizing our two example applications, it also considers more natural and complex situations. Since attention cannot really be controlled in the setup, no validation of single popping regions is possible this time. Instead, an indirect evaluation of the overall prediction of all popping regions is performed.

5.1 Experiment I: direct evaluation with simple object

For the experiment we adapt our geometric LOD example, using a simple bicubic B-spline patch (see Fig. 6) instead of a mesh. Different LODs are obtained by varying the tessellation level. Normals are computed per pixel by directly evaluating the patch's derivatives. We show 4.5-second sequences of the camera approaching the patch, during each of which the LOD is changed exactly once. Two LOD sets (tessellation factor $2 \rightarrow 3$ and $3 \rightarrow 9$) and seven distances at which the LOD is switched are considered. In addition to showing these scenarios as an animation (dynamic case), we also present just the frame where the LOD transition occurs, initially with the old and then with the new LOD (static case). Each sequence is shown three times, as well as one extra time without any change in LOD to verify that a subject indeed perceives popping and is not giving random answers.

The stimuli are presented on a dark-grey background in the central 1024×768 region of a 20" FSC P20-2 LCD display with a resolution of 1600×1200. The participant, being sat at a viewing distance of 60 cm, is instructed to attend to the tip of the patch. After each sequence, he is asked whether popping was noticed. In total, 112 sequences are presented in random order with the dynamic scenarios preceding the static ones; a whole session lasts less than 20 minutes. To make the subject familiar with the task and the voting interface, an exercise session is run before the actual experiment.

Eight subjects with normal or corrected-to-normal vision, all of them members of our institution, participated in the experiment. The mean of the subjects' average popping detection rates is 67.71% (stdev: 12.01%) when LOD switches occurred but only 4.46% (4.58%) in absence of a LOD transition (and hence popping), indicating the plausibility of the answers. A repeated-measures ANOVA applied to the cases where the LOD was changed shows a main effect of the employed LOD set ($F(1,7) = 98.926$; $p < 0.0001$) and of the switch distance ($F(6,42) = 30.226$; $p < 0.0001$), as well as an interaction of these two factors ($F(6,42) = 3.139$; $p = 0.012$). Moreover, there is also a main effect of whether the stimulus is static or dynamic ($F(1,7) = 29.377$; $p < 0.001$).

To evaluate our predictor, we consider the single popping region at the tip of the patch. For the static case, this region's rating value R shows an almost monotonic relationship with the average popping detection rate (cf. Fig. 6). This is also reflected in a high Kendall rank correlation coefficient $\tau_c = 0.933$. To compute the Pearson product-moment correlation coefficient r, we first clamp R to the lowest value where a 100% detection performance is encountered, accounting for the detection rate's upper bound of 100%. The value of $r = 0.915$ indicates a rather highly linear relationship.

As a sanity check, a comparison with another metric's performance is desirable. In particular, it is advisable to test whether the complexity of the predictor is justified and the obtained results are really superior compared to a much simpler and cheaper approach. However, being not aware of any alternative popping predictor, the best we can do is to adopt our approach of comparing the actual frame with its prediction using the previous LOD and employ some image-space metric to compute their difference. Unfortunately, it is also not obvious how to reasonably aggregate such a metric's pixel-wise output. Therefore, opting for simplicity, we chose the maximum CIELAB ΔE^*_{94} [Fairchild 2004] difference value. It shows a weaker correlation ($\tau_c = 0.723$, $r = 0.806$) to the subjects' average detection rates. In particular, unlike our predictor, this metric only reasonably orders the scenarios using the same LOD set but fails to correctly rank them across LOD sets (cf. Fig. 6).

In the dynamic case, our predictor's output still shows a high correlation to the average detection rate ($\tau_c = 0.808$, $r = 0.932$). It also performs far better than maximum ΔE^*_{94} ($\tau_c = 0.561$, $r = 0.747$) which doesn't account for object motion. Compared to the static case, however, our predictor slightly overestimates the subjects' detection performance. We partially attribute this to attentional effects; tracking of the patch's tip has to be performed, which complicates focusing attention. Also, LCD motion blur, which we don't account for, might lower the perceptibility. Nevertheless, when considering both static and dynamic cases together, correlation is still reasonably high ($\tau_c = 0.790$, $r = 0.902$) and better than in case of maximum ΔE^*_{94} ($\tau_c = 0.589$, $r = 0.786$).

5.2 Experiment II: indirect evaluation with real-world examples

In the second experiment, we show four-second sequences of the two example applications where exactly one LOD switch occurs in each. Unless otherwise noted, we use the same setup and procedure as in the first experiment. Twelve subjects with normal or corrected-to-normal vision, again members of our institution, participated. In total, each of them is shown 64 stimuli in randomized order; a whole session lasts about 25 minutes.

Object-wise geometric LOD In the first part of the experiment, we use the Stanford bunny for the object-wise geometric LOD. We consider eight different scenarios, varying the movement velocity and employed LODs (the coarser LOD features between 3% and

88% less triangles than the finer LOD). Moreover, different initial locations and movement directions were chosen. The bunny moves either horizontally across screen or towards the user. The LOD is switched after a certain horizontal distance has been covered, or the distance to the camera has fallen below a threshold, respectively. As in experiment I, for each scenario we additionally include the corresponding static case, and, for testing for subject reliability, consider each sequence also without changing the LOD. Altogether, each scenario is hence presented to a subject in four instances (dynamic/static, with/without LOD change).

The participant is essentially freely viewing the object and not told a specific region to which to direct its attention. To increase the chance that the subject attends a region where popping occurs, we show each dynamic scenario instance three times and each static one twice, with a one-second gray interval between the repetitions. After all repetitions of a stimulus have been presented, the subject is asked to vote whether popping was perceived and, if yes, to rate the strongest detected popping artifact on a three-level scale (hardly ... clearly visible).

The mean of the subjects' average popping detection rates is 61.46% (stdev: 16.61%) when the LOD was changed and 4.17% (4.87%) in case of no LOD switches, suggesting that popping artifacts were indeed perceived by all subjects. For the cases with LOD changes, a repeated-measures ANOVA shows a main effect of whether the stimulus is static or dynamic ($F(1, 11) = 5.337$; $p = 0.041$) and of the scenario ($F(7, 77) = 18.222$; $p < 0.0001$) on detection performance.

Concerning the evaluation of our predictor's output, comprising several popping regions for each scenario, note that we cannot predict whether one of them is attended to and especially not which one. However, assuming that our predictor works, we can reasonably expect that the chance of attending to any of the predicted popping regions increases as the object's coverage with popping regions grows. Moreover, the larger the ratings R of the predicted regions, the higher the chance of detecting popping when attending to a popping region. Adopting this reasoning, we assign to each output of our predictor both a coverage score (four levels: tiny, small, large, huge) and a rating score (four levels: very low, low, high, very high) reflecting the average rating R of the most highly rated popping regions. We then derive an integer detection score (1 ... 5) according to a rule table. For example, tiny coverage and low rating yields a detection score of 1, small but high maps to 3, and huge and high to 5.

For the obtained detection score, we observe a high rank correlation to the subjects' average detection rate for the static case ($\tau_c = 0.833$), the dynamic case ($\tau_c = 0.917$), as well as both cases together ($\tau_c = 0.830$). Treating the score as interval-scaled, Pearson's r suggests a highly linear relationship (static: $r = 0.922$; dynamic: $r = 0.929$; both: $r = 0.927$).

Further analysis shows that there is also a lower, but still distinct correlation between coverage score and detection rate (static: $\tau_c = 0.750$, $r = 0.816$; dynamic: $\tau_c = 0.792$, $r = 0.860$; both: $\tau_c = 0.750$, $r = 0.848$). In addition, the rating score correlates well to the subjects' average rating of how strong they perceived a detected popping artifact (static: $\tau_c = 0.938$, $r = 0.916$; dynamic: $\tau_c = 0.750$, $r = 0.829$; both: $\tau_c = 0.781$, $r = 0.878$). Overall, we reckon that these distinct relationships are an encouraging indication that our predictor works well.

Terrain LOD In the second part of the experiment, our terrain LOD example is used, showing a fly-over. We again consider eight scenarios with varying flying speeds; in each a different terrain region is subjected to a LOD switch. The LOD is changed after a certain distance has been covered, affecting only a single terrain tile. As before, each scenario is presented in four different instances (dynamic/static, with/without LOD change). We also adopt the same stimulus presentation and voting procedure as in the first part. However, since the terrain sequences are much more complex and popping can only occur at a rather small region (a single tile), the likelihood that the subject directs its attention towards the occurrence of popping is far too low, as also indicated by a pretest. To address this, we highlight a circular region with a radius of about 50 pixels on average, to which the user should attend to, for two seconds before each trial.

The mean of the subjects' average detection rates is 67.71% (stdev: 33.59%) for the instances with a LOD change and 3.13% (4.21%) otherwise, again indicating that no random answers were given. A repeated-measures ANOVA for the cases with LOD changes shows a main effect of whether the stimulus is static or dynamic ($F(1, 11) = 5.337$; $p = 0.004$) on detection rate, but not of the scenario ($p = 0.170$).

In the static case, our predictor's output always yields a coverage score of at least "large" and a rating score of at least "high". On the other hand, the mean of the scenarios' average detection rates is 75.0% (stdev: 8.33%), and the mean of the average severity ratings on the three-level scale is 2.20 (stdev: 0.24). That is, well-visible popping didn't get noticed by everyone, which we attribute to inattentional blindness. For instance, even a whole mountain tip popping in got missed by two subjects. Therefore, we feel that the overall voting result is well captured by our predictor's output.

In the dynamic case, we observe a larger variation of coverage and rating scores as well as in subject response. However, given that attention was guided to the affected terrain region and that there are three repetitions to detect popping, we expect coverage score to play a minor role. It is thus not surprising that coverage score is only weakly correlated to the subjects' average detection rate ($\tau_c = 0.234$, $r = 0.285$) whereas, on the other hand, the rating score shows a distinct relationship with the detection rate ($\tau_c = 0.750$, $r = 0.811$).

Overall, both the presented indirect evaluation of our predictor for the second experiment and the direct evaluation of a single popping region in the first experiment indicate that our approach yields plausible and useful predictions of popping perceptibility. In particular, we consider the good correlations between our predictions and the subjects' votings to be very encouraging, especially given the simplifying assumptions made and the influence of attentional effects.

6 Conclusion

In this paper, we focused on the perception of visual popping, mostly in dynamic scenes, discussing several of the aspects having a major influence. Ignoring some of the involved complexity, we introduced an approach for predicting popping that employs a spatio-velocity color vision model to detect differences between an actually shown frame and its prediction obtained with the previous LOD. From the visual differences, meaningful popping regions are derived which predict where popping artifacts of which severity occur.

Constituting a promising first step, with encouraging results from a conducted user study, our approach suffers from several limitations. We basically only support popping due to transitions among geometric LODs, owing to the crucial simplification of considering just a single frame. Since we don't take all relevant factors like motion blur inherent to LCD displays into account, our prediction may be too conservative. Finally, assessing and incorporating attention remains a major challenge.

Acknowledgements

This work was partially supported by the European Union within the CROSSMOD project (EU IST-014891-2).

References

BARTEN, P. G. J. 2003. Formula for the contrast sensitivity of the human eye. In *Proc. SPIE*, vol. 5294, 231–238.

BOLIN, M. R., AND MEYER, G. W. 1999. A visual difference metric for realistic image synthesis. In *Proc. SPIE*, vol. 3644, 106–120.

BONNEH, Y., AND SAGI, D. 1998. Effects of spatial configuration on contrast detection. *Vision Research 38*, 22, 3541–3553.

BURT, P. J., AND ADELSON, E. H. 1983. The Laplacian pyramid as a compact image code. *IEEE Trans. Communications 31*, 4, 532–540.

COHEN, J., OLANO, M., AND MANOCHA, D. 1998. Appearance-preserving simplification. In *Proc. ACM SIGGRAPH 98*, 115–122.

DALY, S. J. 1993. The visible differences predictor: An algorithm for the assessment of image fidelity. In *Digital Images and Human Vision*. MIT Press, ch. 14, 179–206.

DALY, S. J. 1998. Engineering observations from spatiovelocity and spatiotemporal visual models. In *Proc. SPIE*, vol. 3299, 180–191.

FAIRCHILD, M. D. 2004. *Color Appearance Models*, 2nd ed. John Wiley & Sons.

FENG, X.-F. 2006. LCD motion-blur analysis, perception, and reduction using synchronized backlight flashing. In *Proc. SPIE*, vol. 6057, 213–226.

GEPSHTEIN, S., TYUKIN, I., AND KUBOVY, M. 2007. The economics of motion perception and invariants of visual sensitivity. *Journal of Vision 7*, 8, Article 8.

GIEGL, M., AND WIMMER, M. 2007. Unpopping: Solving the image-space blend problem for smooth discrete LOD transitions. *Computer Graphics Forum 26*, 1, 46–49.

GUTHE, M., BALÁZS, Á., AND KLEIN, R. 2005. GPU-based trimming and tessellation of NURBS and T-spline surfaces. *ACM Trans. Graphics 24*, 3, 1016–1023.

HAMILL, J., MCDONNELL, R., DOBBYN, S., AND O'SULLIVAN, C. 2005. Perceptual evaluation of impostor representations for virtual humans and buildings. *Computer Graphics Forum 24*, 3, 623–633.

HARDY, J. L., DELAHUNT, P. B., OKAJIMA, K., AND WERNER, J. S. 2005. Senescence of spatial chromatic contrast sensitivity. I. Detection under conditions controlling for optical factors. *JOSA A 22*, 1, 49–59.

HOPPE, H. 1996. Progressive meshes. In *Proc. ACM SIGGRAPH 96*, 99–108.

ITTI, L., AND KOCH, C. 2001. Computational modelling of visual attention. *Nature Reviews Neuroscience 2*, 3, 194–203.

KELLY, D. H. 1979. Motion and vision. II. Stabilized spatio-temporal threshold surface. *JOSA 69*, 10, 1340–1349.

KELLY, D. H. 1983. Spatiotemporal variation of chromatic and achromatic contrast thresholds. *JOSA 73*, 6, 742–750.

LE GRAND, Y. 1968. *Light, Colour and Vision*, 2nd English ed. Chapman and Hall.

LEGGE, G. E. 1981. A power law for contrast discrimination. *Vision Research 21*, 4, 457–467.

LOVELL, P. G., PÁRRAGA, C. A., TROSCIANKO, T., RIPAMONTI, C., AND TOLHURST, D. J. 2006. Evaluation of a multiscale color model for visual difference prediction. *ACM Trans. Applied Perception 3*, 3, 155–178.

LUBIN, J. 1995. A visual discrimination model for imaging system design and evaluation. In *Vision Models for Target Detection and Recognition*. World Scientific Publishing, 245–283.

LUEBKE, D., REDDY, M., COHEN, J. D., VARSHNEY, A., WATSON, B., AND HUEBNER, R. 2002. *Level of Detail for 3D Graphics*. Morgan Kaufmann.

MYSZKOWSKI, K. 2002. Perception-based global illumination, rendering, and animation techniques. In *Proc. SCCG 2002*, 13–24.

PAN, H., FENG, X.-F., AND DALY, S. 2005. LCD motion blur modeling and analysis. In *Proc. ICIP*, II–21–24.

PATTANAIK, S. N., FERWERDA, J. A., FAIRCHILD, M. D., AND GREENBERG, D. P. 1998. A multiscale model of adaptation and spatial vision for realistic image display. In *Proc. ACM SIGGRAPH 98*, 287–298.

PELLACINI, F. 2005. User-configurable automatic shader simplification. *ACM Trans. Graphics 24*, 3, 445–452.

QU, L., AND MEYER, G. W. 2006. Perceptually driven interactive geometry remeshing. In *Proc. I3D 2006*, 199–206.

REDDY, M. 1997. *Perceptually Modulated Level of Detail for Virtual Environments*. PhD thesis, University of Edinburgh.

SCHAUFLER, G. 1995. Dynamically generated imposters. In *Modeling – Virtual Worlds – Distributed Graphics*, 129–135.

SCHÜTZ, A. C., DELIPETKOS, E., BRAUN, D. I., KERZEL, D., AND GEGENFURTNER, K. R. 2007. Temporal contrast sensitivity during smooth pursuit eye movements. *Journal of Vision 7*, 13, Article 3.

ULRICH, T., 2002. Rendering massive terrains using chunked level of detail control. In: *ACM SIGGRAPH 2002 Course Notes*.

VAN DER HORST, G. J. C., AND BOUMAN, M. A. 1969. Spatiotemporal chromaticity discrimination. *JOSA 59*, 11, 1482–1488.

WANDELL, B. A. 1995. *Foundations of Vision*. Sinauer Associates.

WATSON, A. B., AND AHUMADA, JR., A. J. 2005. A standard model for foveal detection of spatial contrast. *Journal of Vision 5*, 9, 717–740.

WILLIAMS, N., LUEBKE, D., COHEN, J. D., KELLEY, M., AND SCHUBERT, B. 2003. Perceptually guided simplification of lit, textured meshes. In *Proc. I3D 2003*, 113–121.

YEE, H., PATTANAIK, S., AND GREENBERG, D. P. 2001. Spatiotemporal sensitivity and visual attention for efficient rendering of dynamic environments. *ACM Trans. Graphics 20*, 1, 39–65.