Subjective Validation Study for Video Quality Assessment for Computer Graphics Applications

Tunç Ozan Aydın* Martin Čadík* Karol Myszkowski* Hans-Peter Seidel*

MPI Informatik

1 Introduction

In this supplementary material we discuss the subjective validation study (summarized in the main publication) in more detail. The goal of the study was to examine the correlation between the objective quality predictions computed by the proposed video quality metric, and the subjective responses obtained by the experimental procedure described below in Section 2. The calibration procedure (described in Section 4 of the main publication) and the validation study are complementary, in the sense that the former involves simple stimuli at near threshold visibility to match the sensitivity of the metric to that of an average observer, and the latter involves more complex, application oriented stimuli for validating that the individual components of the metric work well in concert.

Two important properties of the proposed metric were influential while designing the validation study: (i) the capability of assessing the quality of HDR videos, as well as comparing HDR videos with LDR videos and vice versa, and (ii) the outcome of the metric in the form of distortion maps that show quality prediction as a function of spatial position, which is especially important for applications in computer graphics. To that end the subjective study has the following novelties over previous studies on video quality assessment:

- The test set includes LDR-LDR, HDR-HDR, and HDR-LDR reference-test video pairs with various types of distortions.
- A BrightSide DR37-P HDR display (max. luminance ≈ 3000 cd/m^2) was used for displaying the videos.
- The subjects are not asked to assess only an overall quality of the video, but to mark the regions where they see differences between test and reference videos, resulting in distortion maps similar to the metric outcome.

In the remainder of this document we will describe our experimental setup and procedure (Section 2), present (Section 3) and discuss (Section 4) the results based on the correlation between the outcome of the subjective study and corresponding predictions of our metric, PDM, HDRVDP and DRIVDP, and conclude with final remarks and future directions (Section 5).

2 Experimental Methods

The set of 9 reference-test video pairs (1 LDR-LDR, 2 HDR-LDR, and 6 HDR-HDR) used in the experiment are listed in Table 1. The video stimuli were generated by imposing temporally varying visual artifacts to HDR scenes (Figure 1), such as HDR video compression artifacts and temporal random noise along with temporal luminance modulation and tone mapping. The magnitudes of the visual artifacts were carefully selected so that there were *sub-*, *near-* and *supra-threshold* distortions present in the experimental videos.

The temporal random noise was generated by filtering a three dimensional array of random values between -0.5 and 0.5 by a Gaussian with standard deviations 20 (high) and 5 (low) pixels along each dimension. The magnitude of noise was adjusted by multiplying with two constants separately, such that the artifacts are barely visible in one setting (low), and clearly visible in the other (high). HDR compression [Mantiuk et al. 2004] was similarly applied at two levels to the HDR scenes, where the luminance was globally modulated over time by 0.5% of the maximum scene luminance to vary the visibility of image details over time. Videos generated by applying tone mapping operators [Fattal et al. 2002; Pattanaik et al. 2000] to each input HDR video frame were used in the dynamic range independent comparisons.

All test videos consisted of 60 frames, and were presented at 24 fps. In order to faithfully reproduce the luminance values on the HDR display, the response function of the display was measured using a Minolta LS-100 luminance meter. The measurements consisted of 32 samples taken from the displayable luminance range with equal logarithmic spacing. The sample points were then fitted to a 3^{rd} degree polynomial function, from which 100 points were resampled and stored as a lookup table. Finally, the pixel values for the HDR videos were determined by cubic spline interpolation between nearest two luminance levels. Furthermore, the displayed luminance of the HDR videos were measured again at various regions, and whenever necessary, the scenes were slightly recalibrated to ensure that the displayed luminance values match the actual scene luminance.

#	Source	Ref. DR	Test DR	Artifact Type of Test Video
1	Cars	HDR	HDR	Noise - high magnitude, low stddev
2	Lamp	HDR	HDR	Noise - high magnitude, low stddev
3	Desk	HDR	HDR	Noise - low magnitude, low stddev
4	Tree	HDR	HDR	Noise - high magnitude, high stddev
5	Cafe	HDR	HDR	HDR compression - high quality, luminance mod.
6	Tower	HDR	HDR	HDR compression - low quality, luminance mod.
7	Cafe	HDR	LDR	Luminance modulation, Pattanaik's tone mapping
8	Lamp	HDR	LDR	Luminance modulation, Fattal's tone mapping
9	Lamp	LDR	LDR	Noise

Table 1: *List of the experimental stimuli. Refer to text for details.*

The participants of the study were 16 subjects between ages of 23 and 50. They all had near-perfect or corrected to normal vision, and were naïve for the purposes of the experiment. Each subject evaluated the quality of the whole test set through a graphical user interface displayed on a BrightSide DR37-P HDR display (Figure 2). In the HDR-HDR, and LDR-LDR comparisons, the task was to mark the regions in the test video where visible differences were present with respect to the reference video. In the HDR-LDR comparisons, on the other hand, the subjects were asked to assess the contrast loss and amplification. In the instruction phase before the experiment, the subjects were asked to mark a grid tile even if visible differences were present only in a portion of that grid's area. They were also encouraged to mark a grid tile in the case they cannot decide whether it contains a visible difference or not. The subjects were placed 0.75 meters away from the display so that a 512×512 image spanned 16 visual degrees and the grid cell size was approxi-

*e-mail: {tunc, mcadik, karol, hpseidel}@mpi-inf.mpg.de

Figure 1: *The video test set is generated from 6 calibrated HDR scenes (here tone mapped for presentation purpose [Reinhard et al. 2002]). The scene luminance was clipped where it exceeded the maximum display luminance. The displayed luminance of the videos resulting from the scenes were between 0.1 and 3000* cd/m^2.

mately 1 visual degree. The environment illumination was dimmed and controlled, and all subjects were given time to adapt to the room illumination. There were no time limitations set for the experiment, but the majority of the subjects took 15-30 minutes for the entire test set.

Figure 2: *The experiment was performed through a graphical user interface shown on the HDR display. Subjects were shown reference and test videos side by side in a randomized order (right), and were asked to mark the relevant image locations on a 16 × 16 grid according to the instructions (left). The interface and messages were disabled while the videos were being shown. The interface allowed the subjects to watch the videos for an unlimited amount of iterations.*

3 Results

The marked regions for each trial were stored as distortion maps with 16 × 16 resolution, which were then averaged over all subjects to find the mean subjective response. Next, the metric prediction for the corresponding stimulus was computed, averaged over the whole 60 frames, and downsampled to the same resolution as the mean subjective response. For each video pair, we computed the 2D correlation between the mean subjective response and the metric prediction (Table 3) and used the results to evaluate the performance of our metric.

The resulting correlations for our metric vary from 0.733 to 0.883. The first two columns of Figure 3 show the mean subjective distortion maps along with the corresponding metric predictions for visual inspection. Furthermore, the descriptive statistics of these maps are summarized in Table 2. While not optimal, we believe that the presented correlations, along with the fact that the maps obtained by the metric's predictions and the subjective experiment look visually similar, clearly show that our metric's predictions are accurate for practical purposes. Highest correlations were obtained for the #2 HDR-HDR Lamp stimulus with high magnitude, low standard deviation noise, and the #7 HDR-LDR Cafe stimulus with luminance modulation and Pattanaik's tone mapping (0.883 and 0.879, respectively). For these two cases, the magnitude of the probability of detection predicted by the metric, and the average of the binary maps over subjects obtained experimentally are also very similar. In other cases, either the magnitudes of the mean subjective maps were lower than the corresponding detection probability magnitude predictions (such as #4 Tree HDR-HDR stimulus with high magnitude, high standard deviation noise, and #9 Lamp LDR-LDR stimulus with noise), or a certain region with visible distortions was missed out (#1 Cars HDR-HDR stimulus with high magnitude, low standard deviation noise). For the remaining stimuli, a combination of both deviations can be observed in the metric predictions and subjective responses. However, even in the worst case (#8, 0.733), the correlation was at an acceptable level.

Figure 4 shows the standard deviations for each stimulus over the test subjects, separately for each grid tile. Over all images, the minimum and maximum values are obtained as 0 and 0.51, the former indicating the tiles on which all subjects gave the same response, and the latter indicating the tiles where approximately half of the subjects have marked.

4 Discussion

A problem we experienced during the experiment was the extreme brightness of the sky region of the *Tower* scene, reaching the maximum displayable luminance level ($\approx 3000 cd/m^2$). We observed that subjects were disturbed by the high luminance level and rushed to the next scene. We also found that the subjects had difficulties understanding the concept of contrast amplification. We believe the reason for that might be that contrast amplification often improves quality, unlike other distortions that were employed in the experiment. As a result, the correlation results in these two cases are slightly worse compared to the others.

We also computed the predictions of PDM [Winkler 2005], HDRVDP [Mantiuk et al. 2005], and DRIVDP [Aydın et al. 2008]. The latter two metrics are designed for image quality evaluation, thus, as in the main publication, the video stimuli was evaluated for each frame separately. HDRVDP, while capable of evaluating the quality of HDR images, lacks any temporal processing and is geared towards comparing images with the same dynamic range. The DRIVDP addresses the latter limitation, but still suffers from the former. Consequently, DRIVDP's predictions for the HDR-LDR stimuli (numbers 7 and 8) is slightly better than HDRVDP. PDM, on the other hand, is designed for the video stimuli, but lacks the HDR and dynamic range independent mechanisms of HDRVDP and DRIVDP, producing the least average correlation with the subjective responses. As shown in Table 3, our metric significantly outperforms others in most cases. The significant difference in average correlations over the entire test set (last row of Table 3) shows that overall our metric's predictions are clearly more accurate than others. The corresponding distortion maps predicted by PDM, HDRVDP and DRIVDP are shown in Figure 3 columns 3 - 5 (averaged and downsampled to 16 × 16 after the computation).

While the relation between the correlation values and distortion maps is obvious in most cases, the high correlation of PDM for

Stimulus #	Subjective Response	Our Metric	PDM	HDRVDP	DRIVDP
	[min, max]; avg; std	[min, max]; avg; std	[min, max]; avg; std	[min, max]; avg; std	[min, max]; avg; std
1	[0.000, 1.000]; 0.177; 0.276	[0.000, 0.850]; 0.128; 0.230	[0.000, 0.301]; 0.082; 0.079	[0.000, 0.019]; 0.001; 0.002	[0.075, 0.417]; 0.194; 0.058
2	[0.000, 1.000]; 0.201; 0.347	[0.000, 0.954]; 0.185; 0.282	[0.000, 0.813]; 0.061; 0.138	[0.000, 0.893]; 0.050; 0.157	[0.072, 0.799]; 0.218; 0.155
3	[0.000, 1.000]; 0.082; 0.242	[0.000, 0.307]; 0.015; 0.045	[0.000, 0.052]; 0.003; 0.008	[0.000, 0.889]; 0.163; 0.247	[0.006, 0.440]; 0.090; 0.078
4	[0.000, 1.000]; 0.124; 0.250	[0.001, 0.457]; 0.094; 0.115	[0.000, 0.024]; 0.007; 0.006	[0.000, 0.000]; 0.000; 0.000	[0.067, 0.240]; 0.137; 0.039
5	[0.000, 1.000]; 0.066; 0.186	[0.000, 0.420]; 0.026; 0.063	[0.000, 0.952]; 0.146; 0.207	[0.000, 0.866]; 0.074; 0.166	[0.040, 0.873]; 0.241; 0.199
6	[0.000, 1.000]; 0.399; 0.389	[0.072, 0.468]; 0.232; 0.103	[0.810, 0.984]; 0.965; 0.026	[0.180, 0.942]; 0.657; 0.202	[0.626, 0.928]; 0.789; 0.058
7	[0.000, 1.000]; 0.312; 0.392	[0.037, 0.984]; 0.451; 0.342	[0.838, 0.984]; 0.980; 0.018	[0.002, 0.953]; 0.448; 0.327	[0.031, 0.953]; 0.374; 0.288
8	[0.000, 0.812]; 0.108; 0.180	[0.041, 0.942]; 0.225; 0.146	[0.606, 0.984]; 0.971; 0.043	[0.005, 0.953]; 0.509; 0.274	[0.148, 0.884]; 0.406; 0.172
9	[0.000, 1.000]; 0.105; 0.238	[0.000, 0.502]; 0.054; 0.104	[0.000, 0.396]; 0.032; 0.066	[0.000, 0.211]; 0.006; 0.025	[0.067, 0.577]; 0.176; 0.097

Table 2: *Descriptive statistics of distortion maps (depicted in Figure 3) for each input stimulus. Abbreaviations used: min=minimal value, max=maximal value, avg=average value, std=standard deviation.*

stimulus #3 deserves further explanation. While PDM correctly detects the distorted regions in that stimulus in a spatial sense, the magnitude of detection probabilities are very low (refer to Table 2), to the point that they are quantized by the visualization. Thus the map appears to be blank, but since the relation with the subjective data is linear, the correlation is high.

For the purposes of generating the maps in Figure 3, in cases of PDM and HDRVDP we simply used the distortion maps produced by those metrics. In the DRIVDP case however, the output of the metric is three separate maps for contrast loss, amplification and reversal. Thus, it is not clear how to produce a single distortion map for HDR-HDR and LDR-LDR stimuli. After experimenting with various methods for combining the distortion maps predicted by DRIVDP, we found that the combined map defined as:

$$P^{k,l,m}_{combined} = 1 - (1 - P^{k,l,m}_{loss}) \cdot (1 - P^{k,l,m}_{ampl}), \qquad (1)$$

gives the best correlation with subjective data. Here, $P^{k,l,m}_{loss|ampl}$ refer to the detection probability of contrast loss and amplification at scale k, orientation l, and temporal channel m. The resulting map $P^{k,l,m}_{combined}$ corresponds to the probability of detecting either contrast loss or amplification at a visual channel. Leaving contrast reversal resulted in slightly improved correlations.

5 Conclusion

The high correlations between the metric predictions and subjective responses over a diverse test set, including HDR and LDR stimuli with distortions of various types and magnitudes indicate that the proposed metric provides a reliable estimate of the video quality as a function of spatial location.

We believe the establishment of a public, standardized test set containing video pairs with diverse dynamic ranges and types of artifacts, coupled with corresponding spatially varying subjective responses, is essential for this line of research. As future work, we would like to extend our data set and make it publicly available as a first step in that direction.

Stimulus #	Our Metric	PDM	HDRVDP	DRIVDP
1	**0.765**	-0.0147	0.591	0.488
2	**0.883**	0.686	0.673	0.859
3	0.843	**0.886**	0.0769	0.865
4	**0.815**	0.0205	0.211	-0.0654
5	**0.844**	0.565	0.803	0.689
6	**0.761**	-0.462	0.709	0.299
7	0.879	0.155	0.882	**0.924**
8	**0.733**	0.109	0.339	0.393
9	**0.753**	0.368	0.473	0.617
Average	**0.809**	0.257	0.528	0.563

Table 3: *Correlations of subjective responses with predictions of our metric, PDM, HDRVDP, and DRIVDP. The last row shows the average correlations over the test test, the best correlations for each stimulus are printed in bold text.*

References

AYDIN, T. O., MANTIUK, R., MYSZKOWSKI, K., AND SEIDEL, H.-P. 2008. Dynamic range independent image quality assessment. In *Proc. of ACM SIGGRAPH*, vol. 27(3). Article 69.

FATTAL, R., LISCHINSKI, D., AND WERMAN, M. 2002. Gradient domain high dynamic range compression. In *SIGGRAPH '02*, ACM Press, 249–256.

MANTIUK, R., KRAWCZYK, G., MYSZKOWSKI, K., AND SEIDEL, H.-P. 2004. Perception-motivated high dynamic range video encoding. *ACM Trans. Graph. 23*, 3, 733–741.

MANTIUK, R., DALY, S., MYSZKOWSKI, K., AND SEIDEL, H.-P. 2005. Predicting visible differences in high dynamic range images - model and its calibration. In *Human Vision and Electronic Imaging X*, vol. 5666 of *SPIE Proceedings Series*, 204–214.

PATTANAIK, S. N., TUMBLIN, J. E., YEE, H., AND GREENBERG, D. P. 2000. Time-dependent visual adaptation for fast realistic image display. In *Proc. of ACM SIGGRAPH 2000*, 47–54.

REINHARD, E., STARK, M., SHIRLEY, P., AND FERWERDA, J. 2002. Photographic tone reproduction for digital images. In *SIGGRAPH '02*, ACM Press, 267–276.

WINKLER, S. 2005. *Digital Video Quality: Vision Models and Metrics*. Wiley.

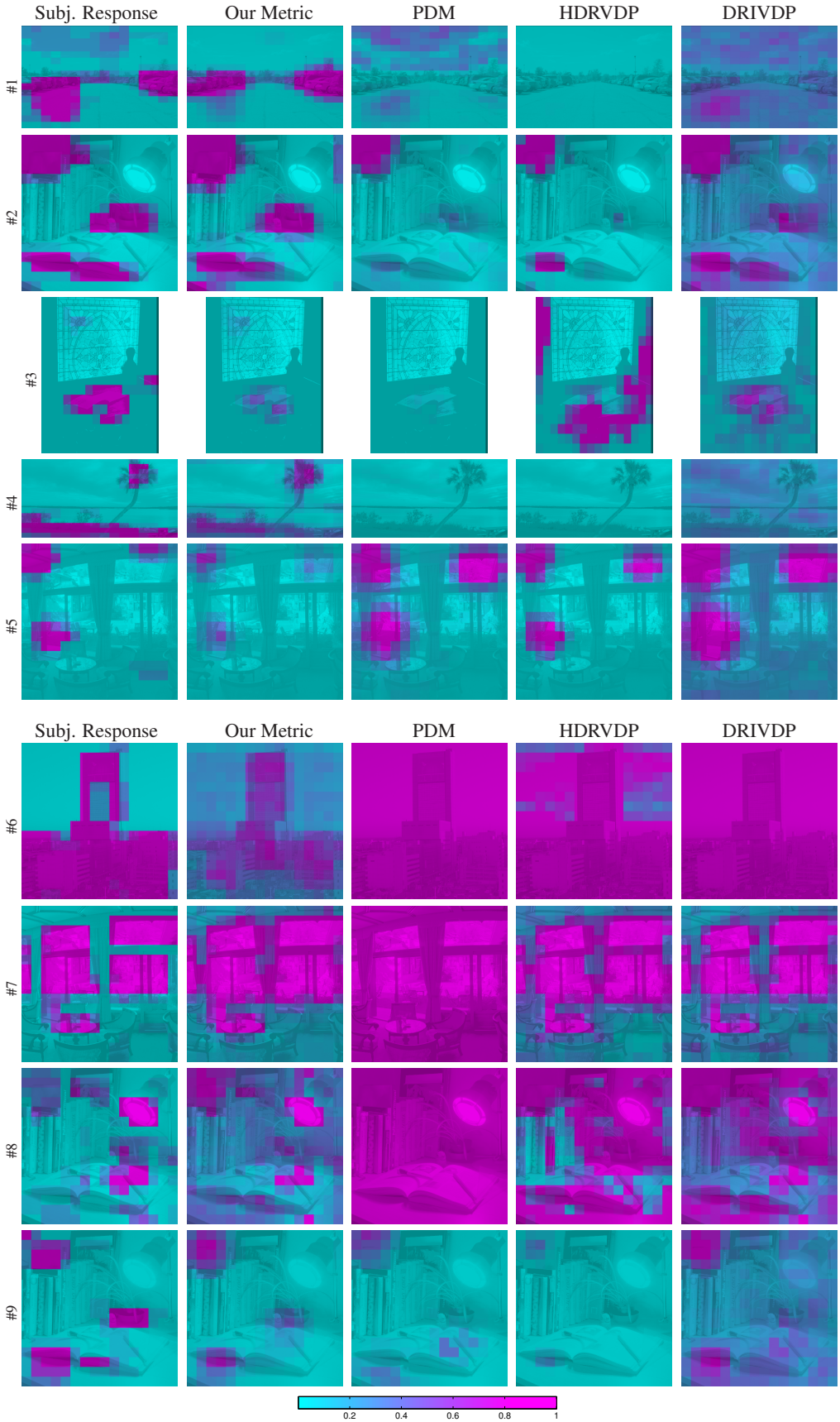

Figure 3: *Mean subjective response maps and corresponding metric predictions pairs.*

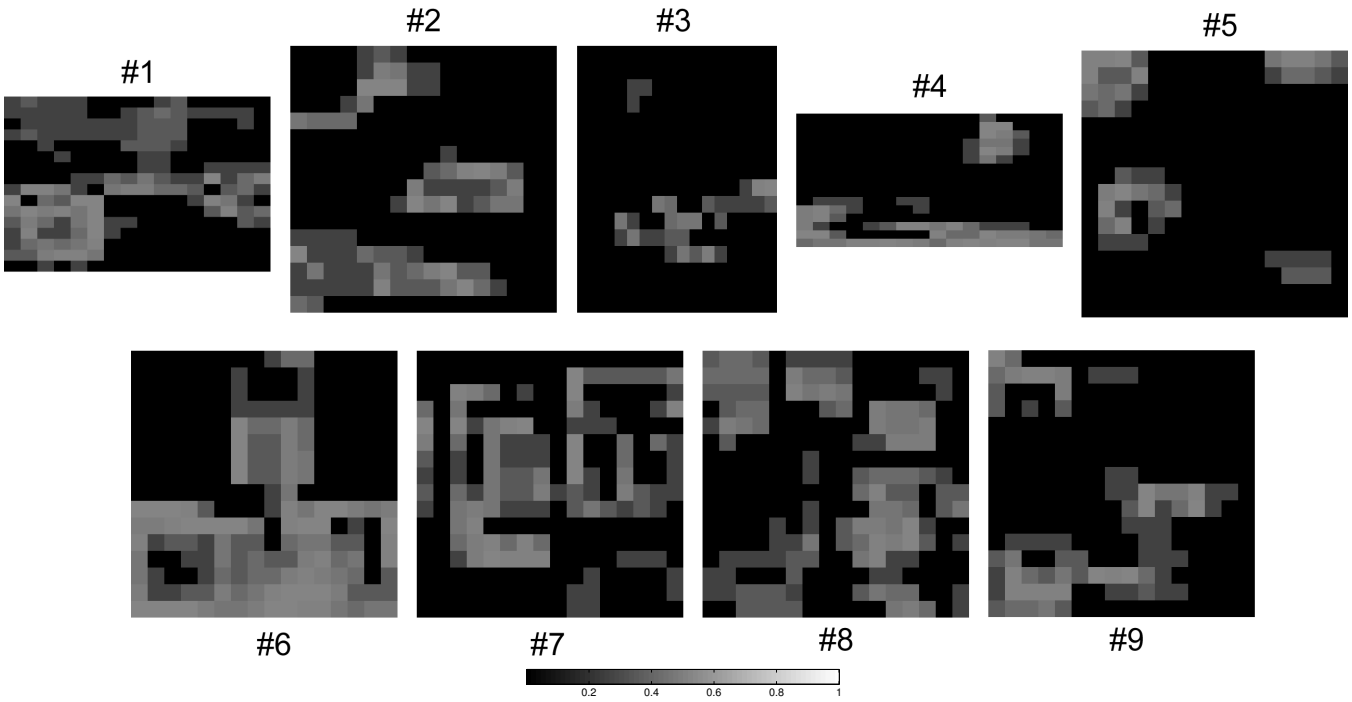

Figure 4: *Maps showing the standard deviations over subjects for each stimulus. The numbers refer to the first column of Table 1.*