Pacific Graphics 2013
B. Levy, X. Tong, and K. Yin
(Guest Editors)

Volume 32 (2013), Number 7

Learning to Predict Localized Distortions in Rendered Images

Martin Čadík*◇ Robert Herzog* Rafał Mantiuk° Radosław Mantiuk▷ Karol Myszkowski* Hans-Peter Seidel*
*MPI Informatik Saarbrücken, Germany
°Bangor University, United Kingdom
◇Brno University of Technology, Czech Republic
▷West Pomeranian University of Technology, Poland

Abstract

In this work, we present an analysis of feature descriptors for objective image quality assessment. We explore a large space of possible features including components of existing image quality metrics as well as many traditional computer vision and statistical features. Additionally, we propose new features motivated by human perception and we analyze visual saliency maps acquired using an eye tracker in our user experiments. The discriminative power of the features is assessed by means of a machine learning framework revealing the importance of each feature for image quality assessment task. Furthermore, we propose a new data-driven full-reference image quality metric which outperforms current state-of-the-art metrics. The metric was trained on subjective ground truth data combining two publicly available datasets. For the sake of completeness we create a new testing synthetic dataset including experimentally measured subjective distortion maps. Finally, using the same machine-learning framework we optimize the parameters of popular existing metrics.

Categories and Subject Descriptors (according to ACM CCS): I.3.3 [Computer Graphics]: Picture/Image Generation—Image Quality Assessment

1. Introduction

Image quality evaluation [WB06, PH11] is one of the fundamental tasks in imaging pipelines, in which the role of synthesized images continuously increases. Modern rendering tools differ significantly in terms of the employed algorithms, e.g., global illumination techniques, which are prone to a great variability of visual artifacts [MKRH11]. Typically such artifacts are of local nature, and their visual appearance differs from more uniformly distributed image blockiness, noise, or blur that arise in compression and broadcasting applications. Existing objective image quality metrics (IQM) are specialized in predicting the level of annoyance caused by such globally present artifacts, and conform well with a single quality value, which is derived in mean opinion score (MOS) experiments with human observers [SSB06]. While some of the objective IQMs such as structural similarity index (SSIM) [WB06, Ch. 3], Sarnoff visual discrimination model (VDM) [Lub95], or the high-dynamic range visual difference predictor (HDR-VDP) [MKRH11] can locally predict perceived differences, they are not always reliable in rendering [ČHM*12]. Clearly, a need arises for novel metrics that can locally predict the visibility of numerous rendering artifacts, which are simultaneously present in a single image.

Many traditional IQMs can be modeled with a generic two-stage processing: (1) extraction of carefully designed features from the image, and (2) pooling of those features to correlate the aggregated value with subjective experiment data. At the feature extraction stage typically multi-resolution filtering with optional perceptual scaling is performed (VDM, HDR-VDP), or alternatively local pixel statistics are computed (SSIM). At the pooling stage the Minkowski summation of feature differences with respect to the reference solution (VDM, HDR-VDP), or the product of feature differences with optionally controlled non-linearity of each component (SSIM) are considered. However, such a limited feature set might not be sufficient to correctly predict the multitude of rendering-specific distortion types, especially given the variety of image content and nonuniformly distributed, mixed distortion types in a single image. Another limiting factor is the rigid form of the pooling models, which prevents the adaptation to local scene configurations and artifact constellations.

In this work, we propose a novel *data-driven full-reference metric*, which outperforms existing metrics in the

[*] e-mail: mcadik@mpi-inf.mpg.de, project webpage: http://www.mpii.de/resources/hdr/metric/

prediction of visible rendering artifacts. First, we systematically analyze the features used in IQMs, and then introduce a great variety of other features originating from the fields of computer vision [TM08] and natural scene statistics (NSS) [SBC05]. Additionally, we propose a few custom features including saliency data captured with an eye tracker (Section 3). We select the best suited features based on their discriminative power with respect to the rendering artifacts (Section 4). Our feature selection ensures that any distortion type we investigate is covered by a sufficiently large subset of supporting features. Instead of the feature pooling used in IQMs, we refer to machine learning solutions (Section 5), which learn an optimal mapping from the selected feature descriptors to a local quality map with respect to the perceptually measured ground-truth data [HČA*12] and [ČHM*12] (jointly referred in this paper as the LOCCG dataset for LOCalized Computer Graphics artefacts). This way our metric implicitly encapsulates highly non-linear behavior of the human visual system (HVS) that was learned from the perceptual data. To evaluate its generalization performance we also test our metric on an independent synthetic dataset, which we designed as a comprehensible tool that is suitable for evaluating other local quality metrics as well. At last, we use the same methodology to improve the performance of SSIM and HDR-VDP in rendering applications, by carefully tuning the weights associated with the features at the pooling stage (Section 6).

2. Previous Work

In this section we focus on quality metrics, which employ machine learning tools. While the metric proposed in this paper belongs to the category of *full-reference* (FR) metrics as it requires a non-distorted copy of the test image, in our discussion we refer also to *non-reference* (NR) and *reduced-reference* (RR) metrics, where data-driven approach is more common. For a more general discussion and applications of quality metrics we refer the reader to [WB06, PH11], more graphics oriented insights concerning FR metrics can be found in [MKRH11, ČHM*12].

The utility of machine learning methods in image quality evaluation has mostly been investigated for NR metrics. Typically it is assumed that the distortion type is known in advance, and then based on the correlation of its amount with human perception the image quality prediction is reported. The blind image quality index (BIQI) [MB10] introduces a distortion-type classifier to estimate the probability of distortions that are supported by the metric, and then a distortion-specific IQM is deployed to measure its amount. NSS features are employed, whose correlation with subjective quality measure for each distortion is known, and an SVM classifier is used for the quality prediction. NSS features expressed as statistics of local DCT coefficients are used in BLIINDS [SBC10], which can handle multiple distortions as well. Overall, the performance similar to the FR PSNR metric (peak signal-to-noise ratio) is reported for the LIVE dataset [SWCB06], but both BIQI and BLIINDS have trouble for JPEG and Fast Fading (FF) noise distortions. Better results have been reported in [LBW11] when instead of NSS-based features, the more perceptually relevant features: phase congruency, local information (entropy), and gradients are used. Better performance than BIQI and BLIINDS is also reported for the learning-based blind image quality measure (LBIQ) [TJK11] where complementary properties of features stem from NSS, texture and blur/noise statistics.

In RR IQM that are used in digital broadcasting a challenge is to select a representative set of features, which are extracted from an undistorted signal and transmitted along with the possibly distorted image. Redi et al. [RGHZ10] identify the *color correlograms* as suitable feature descriptors for this purpose, which enable the analysis of alterations in the color distribution as a result of distortions.

Machine learning in FR IQM remains mostly an uninvestigated area. Narwaria and Lin [NL10] propose an FR metric based on support vector regression (SVR), which uses singular vectors computed by a singular value decomposition (SVD) as features that are sensitive for structural changes in the image. Remarkably, the proposed metric shows good robustness to untrained distortions and overall outperforms SSIM.

All discussed IQMs have successfully been tested with the LIVE database [SWCB06] (and two other similar databases [NL10]), where a single value with the quality score (MOS) is available for each image. Such testing strategy precludes any conclusions concerning the accuracy of artifact localization and its visibility in the distortion map, which is the goal of this work. While the number of images available in LIVE approaches one thousand, the diversity of distortions is limited to five major classes with the emphasis on compression distortions, noise, and blur, which structurally differ significantly from rendering artifacts. For each stimulus only one distortion is present, which makes the metric performance evaluation for distortion superposition less reliable.

Machine learning solutions have been used in the context of rendered image quality assessment. Ramanarayanan et al. [RFWB07] employed an SVM classifier to predict *visual equivalence* between a pair of images with blurred or warped environment maps that are used to illuminate the scene, but problematic regions in the image cannot be identified. Herzog et al. [HČA*12] proposed a NR metric (NoRM), which is trained independently for three different rendering distortion types. The metric can produce a distortion map, and the lack of reference image is partially compensated by exploiting internal rendering data such as per pixel texture and depth. In this work we focus on solutions that are based merely on images, and can simultaneously handle more than one artifact. We utilize perceptual data derived in [HČA*12] (a part of the LOCCG dataset) to train our FR metric and we compare its performance with respect to NoRM.

3. Features for Image Quality Assessment

Many FR (i.e. the undistorted reference image needs to be available) IQMs have been developed that claim to predict localized image distortions as observed by a human [WB06]. It was shown that there exists no clear winner and each metric has its pros and cons for image distortions measured on synthetic ground-truth datasets [ČHM*12]. For better understanding of which parts of the varying metrics are important for predicting image distortions, we decompose those metrics into their individual features and analyze their strength by means of a data-driven learning framework. Moreover, we introduce new complementary features commonly used in information theory and computer vision [TM08]. Finally, we acquire saliency maps using an eye tracker and include these as a feature into our framework for the analysis of the importance of visual attention. We implemented 32 features of various kinds and origins spanning 233 dimensions which, in our opinion, is an exhaustive set (see Table 1).

3.1. Features of Traditional Image Quality Metrics

In our analysis we include features inspired by popular IQMs, including absolute difference (*ad*), SSIM [WB06, Ch. 3], HDR-VDP-2 [MKRH11], and sCIE-Lab [ZW97]. For those metric features that are only computed at a single scale (e.g., SSIM, *ad*), we additionally include their multi-scale variants. This is achieved by decomposing the feature maps into Gaussian or Laplacian pyramids (without subsampling). Despite its simplicity, *ad* (or PSNR and MSE) are still frequently used quality predictors. In contrast, SSIM measures differences using texture statistics (mean and variance) rather than pixel values. It is computed as a product of three terms:

$$SSIM(\mathbf{x},\mathbf{y}) = [lum(\mathbf{x},\mathbf{y})]^{\alpha} \cdot [con(\mathbf{x},\mathbf{y})]^{\beta} \cdot [struc(\mathbf{x},\mathbf{y})]^{\gamma}, \quad (1)$$

which are a luminance term lum, a contrast term con, and a structure term $struc$ (see Fig. 8) computed for a block of pixels denoted by $\mathbf{x}$ and $\mathbf{y}$. We include each SSIM term as a separate feature: *ssim lum*, *ssim con* and *ssim struct*. Similarly, we include all frequency bands of HDR-VDP-2 differences and their logarithms (more details in Section 6.2), and denote them as *hdrvdp band* and *hdrvdp band log*.

We also introduce a few variations of the SSIM contrast components, which we found to be well correlated with subjective data. The standard contrast component is expressed as: $con(\mathbf{x},\mathbf{y}) = \frac{2\,\sigma_x\,\sigma_y + C_2}{\sigma_x^2+\sigma_y^2+C_2}$, where σ_x and σ_y are the per-block variances in the test and reference images, and C_2 is a positive constant preventing division by zero. The product in the nominator introduces a strong non-linear behavior; the increase of contrast (variance) and decrease have different effect on the value of the component. Marginally better results can be achieved if the contrast difference is expressed as: $con_{bal}(\mathbf{x},\mathbf{y}) = \frac{(\sigma_x - \sigma_y)^2}{\sqrt{\sigma_x^2+\sigma_y^2+\varepsilon}}$, where ε is a small constant (0.0001). We denote this feature as *ssim con bal*. The denominators in these expressions are effectively responsible for contrast masking, which reduces sensitivity to contrast changes with increasing magnitude of the contrast. Such masking can be determined by the image of higher contrast (test or reference): $con_{balmax}(\mathbf{x},\mathbf{y}) = \frac{(\sigma_x - \sigma_y)}{\max(\sigma_x,\sigma_y)+\varepsilon}$. We denote this feature as *ssim con bal max*. Finally, we observed that individual distortions are more noticeable when isolated, rather than uniformly distributed over an image. This effect can be captured by the inhibited contrast feature (*ssim con inhibit*): $con_{inhibit}(\mathbf{x},\mathbf{y}) = \frac{con(\mathbf{x},\mathbf{y})}{\overline{con}(\mathbf{x},\mathbf{y})}$, where $\overline{con}(\mathbf{x},\mathbf{y})$ is the mean value of the contrast term in the image.

	Feature Name	**Dim.**	**Multi scale**	**Import.** multi-dim. (greedy)	**Import.** multi-dim. (stacking)	**Import.** scalar (dec. trees)	**Import.** scalar (AUC)
1	ad [Sec.3.1]	11	✓				
2	bow [Sec.3.2]	32			1.0	1.0	
3	dense-sift diff [BZM07]	1		0.72047			0.86216
4	diff [Sec.3.3]	11	✓		0.48596	0.66906	
5	diff mask [Sec.3.3]	1		0.19609			0.85772
6	global stats [Sec.3.3]	5					
7	grad dist [Sec.3.3]	1					
8	grad dist 2 [Sec.3.3]	1			0.32785	0.66382	0.85919
9	Harris corners [HS88]	12	✓			0.76699	
10	hdrvdp band [MKRH11]	6	✓			0.68933	0.85035
11	hdrvdp band log	6	✓				
12	hog9 [DT05]	62			0.46443		
13	hog9 diff [Sec.3.2]	1			0.32178	0.67821	
14	hog4 diff [Sec.3.2]	1					
15	location prior [Sec.3.4]	2					
16	lum ref [Sec.3.3]	11	✓	0.58963			
17	lum test [Sec.3.3]	11	✓	0.21429			
18	mask entropy I [Sec.3.3]	1		0.40419	0.52820	0.99389	0.86358
19	mask entropy II [Sec.3.3]	5	✓	1.0		0.67035	0.86676
20	patch frequency [Sec.3.4]	1			0.41590		
21	phase congruency [Kov99]	10	✓	0.19712			
22	phow diff [BZM07]	1					
23	plausibility [Sec.3.4]	1			0.32051		
24	sCorrel [Sec.3.3]	1		0.18956			0.8496
25	spyr dist [Sec.3.3]	1				0.85793	
26	ssim con [WBSS04]	11	✓				0.8496
27	ssim con inhibit [Sec.3.1]	1			0.44840		0.84517
28	ssim con bal [Sec.3.1]	1					
29	ssim con bal max [Sec.3.1]	1					
30	ssim lum [WBSS04]	11	✓	0.58791			
31	ssim struc [WBSS04]	11	✓	0.18681	0.53080	0.65608	0.86484
32	vis attention [Sec.3.5]	1					
Metric performance (AUC)				0.880	0.897	0.916	0.892

Table 1: *Left to right: implemented features, their dimensionality, scale selection, estimated normalized importance for best joint features, and one-dimensional sub-features (only the best sub-feature importance is reported for scalar selection methods), see Section 4. The importance of the selected features is color-coded (from blue to green to red). For each set we show the performance (area under the ROC curve for the LOCCG dataset) of a data-driven metric utilizing only the selected features (Section 5). Notice that only* ten best *features in each column are reported for clarity.*

3.2. Computer Vision Features

Much research on features comes from the field of computer vision. Therefore, we analyze popular features from computer vision in the mutual spirit *"what's good for computer vision may also help human vision"* and vice versa. In particular, we consider the following features for image quality assessment: bag-of-visual-words (*bow*) [FFP05], histogram-of-oriented-gradients with 9 orientation bins (*hog9*) [DT05], the Euclidean distance between *hog9* (coarse version *hog4*), dense-SIFT [BZM07], pyramid-histogram-of-visual-words [BZM07] computed for test and reference images denoted as *hog9 diff* (*hog4 diff*), *dense-sift diff*, *phow diff*, respectively, *Harris corners* [HS88], and *phase congruency* [Kov99].

Bag-of-visual-words (*bow*) is perhaps the most commonly used feature in computer vision with a whole field of research devoted to it. Briefly, the typical *bow* feature extraction pipeline consists of two steps: first, the computation of a dictionary of visual words and second, encoding an image with a histogram by pooling the individual dictionary responses on the image. The strength (and weakness) of *bow* is that it ignores the location of sub-image parts making it invariant to global image constellation and thus requiring less training data in supervised learning.

We compute the *bow* feature on the error-residual image, i.e., difference between test and reference image. To generate the dictionary we use a set of artifact-reference image-pairs and randomly extract normalized pixel patches of size $n_p \times n_p$ pixels ($n_p = 8$) from all residual images. Then, we run *k-means clustering* on the patches using the L2-distance metric to generate $k = 200$ clusters from which we then extract a smaller dictionary ($k_d = 32$) by iteratively removing the cluster with the highest linear correlation. The remaining clusters form the visual words of the dictionary. To encode a new image-pair using our dictionary, we first compute the correlation of the error-residual image with each visual word and for each pixel we store the index of the visual word with the maximum response, which is pooled to build a histogram of k_d bins. In contrast to the traditional *bow* we do not compute one histogram for the entire image but a histogram for each pixel by pooling the responses in a local window ($4 \times n_p$ pixels) weighted by a Gaussian with $\sigma = n_p$.

3.3. Statistical Features

As shown in [ČHM*12] and [WBSS04] simple statistics may be powerful features for visual perception. We include both local and global statistics for an image. As local statistics we compute non-parametric *Spearman correlation* per 6×6 pixel block (*sCorrel*), parametric correlation is captured by the SSIM structure term (*ssim struc*), the gradient magnitude distance (*grad dist*) between test and reference image, the sum of squared distances between test and reference image decomposed in a steerable pyramid (*spyr dist*), and visual masking computed by a measure of entropy (*mask entropy I*), which is computed per 3×3 pixel block as the ratio of the entropy in the residual-image block $\mathbf{x} - \mathbf{y}$ to the entropy in the reference-image block $\mathbf{y}$:

$$H_{mask}(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i,j} p(x_{ij} - y_{ij}) \log_2 p(x_{ij} - y_{ij})}{\sum_{i,j} p(y_{ij}) \log_2 p(y_{ij})}, \quad (2)$$

where $p(x_{ij} - y_{ij})$, $p(y_{ij})$ is the probability of the value of pixel (i, j) in the normalized residual-, reference-image block, respectively. We also include a multi-scale version of this feature with larger window size (5×5) denoted as *mask entropy II*. For completeness we also add the luminance of the pixel in the test image (*lum test*) and reference image (*lum ref*), as well as the signed difference (*diff*) at varying image scales as individual scalar features.

In order to see whether global image distortions influence the perception of local artifacts, we add global distortion statistics to our analysis that is computed over the entire image. Specifically, we compute the mean, variance, kurtosis, skewness, and entropy of the distortions in the entire image, which are grouped into one feature class denoted as *global stats* in Table 1.

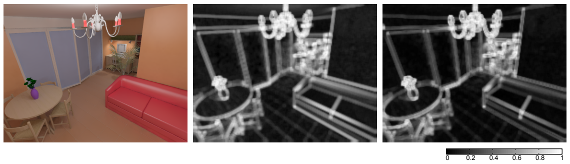

Figure 1: *Plausibility (middle) and the patch frequency feature (right) for the apartment image in LOCCG dataset. Note how repeating structures in the image (e.g., edges and texture on the floor) receive high values.*

3.4. High-level Visual Features

The features described so far are "memory-less" and only of local nature meaning that the information content is restricted to a small image region around the sample point. However, the perception of image distortions is largely dependent on the higher-level human vision following Gestalt laws and learned scene understanding. While a simulation of higher-level human vision is computationally intractable, we added a few features that mimic the global impact of local distortions on the perception of artifacts that is beyond local pixel statistics.

In the LOCCG dataset we observed that some artifacts are subjectively less severe than others depending on the likelihood that such an artifact pattern could also occur in reference images (e.g., darkening in corners). We denote such a phenomenon as *artifact plausibility*. In order to approximately model artifact plausibility we make use of a larger independent dataset of reference photos (the LIVE and Labelme datasets [SWCB06,RTMF07]) from which we sample random sub-images referred to as *patches* of size 16×16 pixels in a pre-process. Since we are mainly interested in the structural similarity of patches, we make patches contrast and brightness invariant by subtracting the mean luminance and dividing by the standard deviation. Moreover, to make later searching efficient, we map these contrast-normalized patches to a truncated DCT basis (12 out of 255 AC-coefficients). For this pool of random patches we build an index data structure for efficiently searching the nearest neighbors. Then, for each sample point in a distorted image we extract a patch following the same steps as in the pre-process and query the k-nearest neighbor patches ($k = 16$) in this database using the L1-distance. Given the distance to the k-th nearest neighbor, we compute an estimate of the probability density for the query patch in the world of all images, which becomes a new feature denoted as *plausibility*.

Inspired by non-local means filtering, we additionally estimate the occurrence frequency of a local image patch by searching for the most similar patches contained in the same image rather than in an independent database as for the *plausibility* feature. This way, patches with common structure (e.g., edges, repeating texture) receive higher values than patches with rare patterns in the same image, see Fig. 1. This feature is denoted as *patch frequency*.

We are also interested in analyzing whether the distribution of the locations of artifacts within an image has an effect on the visibility of local artifacts. Therefore, we compute the first central moments of the artifact distribution in the image, i.e., we compute the mean, variance, kurtosis, and skewness of the artifact distance to the center of the image, which is summarized as *location prior* in Table 1.

3.5. Visual Saliency (Eye Tracking)

Another potentially important cue for the perception of local artifacts may be saliency. To estimate its importance for image quality assessment, we explicitly modeled saliency by employing an eye tracker in a user experiment. Low-resolution saliency maps were generated from the recorded gaze points per image that represent the mean visual exploration, which is stored as a feature denoted by *vis attention*. In the experiment, we were showing images from the LOCCG dataset to observers. The observers were asked to remember the details of the image without any top-level task. The eye tracker collected the gaze data for each image presented for 12 seconds. The answers to these questions were not analyzed and did not affect the results. We calibrated the eye tracker before each set of 5 images to increase the accuracy of the gaze estimation. The observers were asked to use the chin rest to stabilize the head position relative to the display. The experiment was conducted for 13 observers of age 20 to 43 years (12 males and 1 female).

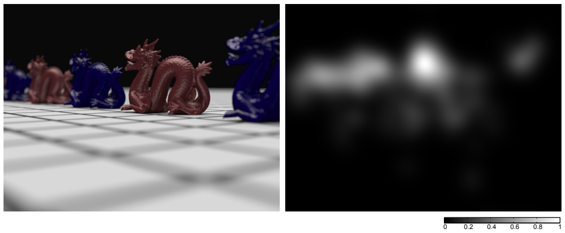

Figure 2: *The new visual attention dataset (examples for scene* dragons*). For each image from the original LOCCG dataset (left), we measure the average saliency map (right).*

The gaze data represents the positions of the gaze points in screen coordinates. For an individual observer we computed the fixation points based on the I-DT technique [Wid84] (with dispersion and duration equal to 100 pixels and 250 ms respectively). The fixation maps were blurred using a low-pass Gaussian filter (σ=20 pixels) to create the saliency maps called heat maps. These maps were averaged and normalized for all observers to prepare one heat map per stimulus image, see Fig. 2.

Our experimental setup consisted of a P-CR RED250 eye tracker controlled by the proprietary SMI iViewX software (version 2.5) running on a dedicated PC. The RED250 eye tracker was mounted under a 22" Dell E2210 LCD display with screen dimensions 47.5 × 30 cm and a native resolution of 1680 × 1050 pixels (60Hz). The results shown in Table 1 indicate that the measured visual saliency maps do not improve the prediction results for the LOCCG dataset. The dataset of the visual attention maps for computer graphics images, however, is interesting for future research and we make it publicly available at the project webpage.

4. Feature Selection

As the number of features we implemented is high (see Table 1), the natural questions we should answer are: first, how significant are particular features to the task of visual distortions prediction, and second, what features should be combined in a joint feature descriptor to give best generalization performance of the new IQM. Optimal feature-subset selection by exhaustive searching is computationally intractable and we experimented with different methods for feature selection where each method provides new information about the strength of individual features.

ROC Analysis One of the easiest ways to rank features is according to area-under-the-curve (AUC) values of their ROC curves [ČHM*12]. Such AUC values are shown in the last column of Table 1. The values show that the dense-sift, masking entropy, and the structural component of SSIM (*ssim struct*) provide the largest predictive power when used alone, though the differences between the best features are moderate. Although ROC analysis identifies strong features, it neither accounts for the correlation of features nor can it detect complementary features that when combined yield the best performance. For that purpose, we attempt three different feature selection strategies.

Greedy Feature Selection This procedure follows in principle the approach proposed in [LSAR10]: among the set of all possible features, we iteratively select the one that gives the smallest cross-validation error when adding it to the pool of selected features and training a classifier on it. The process is continued until adding new features to the pool does not improve the cross-validation error. Here, for classification we use a non-linear support vector machine [CL11] with radial basis function (RBF) kernel with hyper-parameters optimized by a grid-search.

Decision Forests Another common approach for feature selection is to analyze decision trees [Bre01], which we also use for our metric described in Section 5. Ensembles of decision trees are natural candidates for feature selection [Bre01, TBRT09] since they intrinsically perform feature selection at each node of the tree. The expected frequency that a single feature is chosen for a split in a random tree and the trees impurity reduction due to the node split indicates the relative importance of that feature to the tree model [TBRT09]. This type of feature selection differs from the others in the sense that it only provides an importance weight of the scalar components of individual features.

Stacked Classifiers To this end, we also analyzed the importance of individual features by an embedded SVM classifier with L1-regularization [BM98]. To analyze the non-linear discriminative power of individual features, we build a 2-level stack of classifiers [Bre96b] where the first level consists of k non-linear classifiers (SVM) [CL11], one for each feature, that compute the artifact probability based on a single feature. These probability values are fed forward as k independent input features to the second level, which is a single linear classifier $\mathbf{w}_2 \in \mathbb{R}^k$. The classifier $\mathbf{w}_2$ is then trained on a disjoint training set using a SVM with L1-regularization, which results in a sparse vector $\mathbf{w}_2$ that can be interpreted as a joint feature importance – the higher the absolute weight $w_i = |\mathbf{w}_2(i)|, \;\; i \in \{1,..,k\}$ the more discriminative the i^{th} feature. Using this procedure the average weights computed on LOCCG dataset with leave-one-out cross-validation are shown in Table 1.

4.1. Feature Selection Results

All feature selection strategies produce a reasonable feature sub-set that generalizes well when tested with leave-one-out cross-validation on trained decision tree ensembles as shown in the last row of Table 1. Although, the ROC analysis does not exploit correlation of features and selects only the best 1-dimensional features the resulting combined feature subset is still performing well. However, when comparing the feature scores (last 4 columns in Table 1) one can observe some discrepancies in the selected feature sets, which result from slightly different objectives of the methods and correlation among the features. For example decision trees can be considered as ensembles of many weak classifiers based on scalar features, whereas the greedy and the stacking approach operate on multi-dimensional features, and the ROC analysis ignores feature combinations altogether. Further, correlation among individual features can produce different sets that, when carefully observed, may actually be similar. An example are the features *hog9 diff* and *dense-sift diff*, which are highly correlated and chosen mutually exclusively by either method. Also, the signed difference (*diff*) is highly correlated with *ad* and is also a linear combination of *lum test* and *lum ref* and therefore not selected in the greedy approach but for the stacking and decision forest. Nevertheless, in agreement with the majority of the methods, the SSIM structure component (*ssim struc*), the bag-of-words (*bow*), the masking entropy (*mask. entropy I/II*), and the signed difference at multiple image scales (*diff*) can be considered as important features for our task of classifying distortions. Further, we can also rule out certain features that either do not improve performance or are simply redundant. These include absolute difference (*ad*), global image statistics (*global stats*), location of artifacts (*location prior*), and visual attention (*vis attention*). In particular, all high-level and global visual features (Section 3.4) perform rather weak in our analysis. However, this does not necessarily conclude their ineffectiveness but rather our too simplistic modeling of the complex high-level human vision.

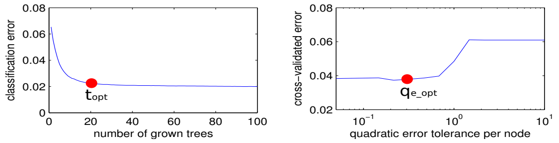

Figure 3: *Optimal parameters of the decision forest. Left: classification error versus number of decision trees t. Right: optimal splitting threshold q_e based on cross-validation.*

5. Data-Driven Metric

We experimented with different classification methods including Naive Bayes classifiers, linear and non-linear support vector machines [CL11], and decision trees [Bre01]. For our data-driven metric we obtained the best results (in terms of ROC area-under-curve) with ensembles of bagged decision trees [Bre01], which we refer to as *decision forest*. Decision forest is a powerful classification and regression tool that is scalable and known for its robustness to noise. Having constructed several random trees by bootstrapping [Bre96a], an observation is classified by traversing each tree from root node to a leaf, which contains the predicted label (artifact/no-artifact) that is averaged across all trees. The path through the tree is determined by comparing single sub-features against learned thresholds in each node. The pruned tree depth and the number of trees controls the accuracy of the classification. Using a cross-validation protocol we empirically set the number of trees to $t = 20$ and the average tree depth to 10 (implicitly controlled by a quadratic error tolerance threshold $q_e = 0.25$ for the node-splitting), which yields good generalization performance (see Fig. 3). We train our metric using the 10 best features as derived in Section 4 (shown in the last but one column of Table 1).

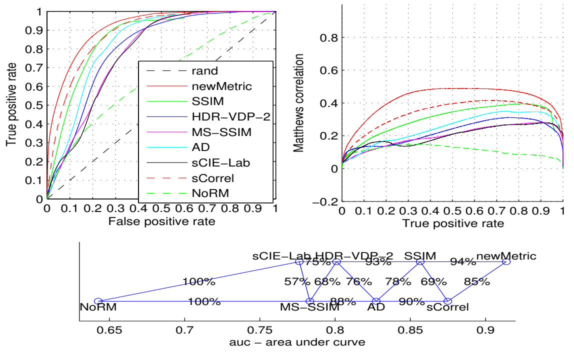

Figure 4: *Quantitative results for quality metrics on LOCCG dataset shown as ROC (top left) and Matthews correlation (top right). The bigger the area under the curve (AUC), the better.* $AUC_{newMetric}$*=0.916,* AUC_{SSIM}*=0.858,* $AUC_{HDRVDP2}$*=0.802,* AUC_{MSSSIM}*=0.786,* AUC_{AD}*=0.832,* $AUC_{sCIELab}$*= 0.783,* $AUC_{sCorrel}$*=0.880,* AUC_{NoRM}*=0.644. Bottom: ranking according to AUC (the percentages indicate how often the metric on the right results in higher AUC when the image set is randomized using a bootstrapping procedure similar to [ČHM*12]).*

5.1. Results

We train our new data-driven metric described above on the LOCCG dataset, which consists of 35 annotated image-pairs that exhibit a variety of computer graphics distortions that are difficult to predict by existing FR IQMs [ČHM*12]. Since the size of the LOCCG dataset is rather small and the images are very diverse showing (combination of) different artifacts and scenes, we do not split it into a train and test set. We instead evaluate our method in a leave-one-out cross validation fashion; i.e., we train it on $n-1$ images and test on the n-th image repeating this process n times. In addition, we validate our metric on a new uncorrelated dataset that is described in Section 5.1.1. We compare the trained metric to 7 state-of-the-art and baseline methods as shown in the quantitative analysis in Fig. 4. Our new metric outperforms all existing FR IQMs on the LOCCG dataset in terms of AUC in Fig. 4 (the higher the AUC the better). Also, the visual results agree with the ground-truth annotation as shown in the color-coded distortion maps for three images of the LOCCG dataset in Fig. 5. Please refer to the supplementary material for all results and a more detailed analysis.

For completeness, we include results of the NR metric NoRM. However, this method was not originally intended to be used for detecting general, mixed image distortions and is tuned for only specific artifacts assuming the depth maps and other cues of the scenes to be available for feature computation. Unfortunately, depth and texture maps are not available in many cases in the LOCCG dataset, and we run NoRM only with color features rendering its performance poor.

We implemented our new metric and feature computation in MATLAB for which the code is available at the project webpage. Reporting the overall computation time of the unoptimized MATLAB code, the data preprocessing and feature computation time per image (800 × 600) is in the order of a few minutes, the time for training the decision forest on our selected feature set based on 100.000 samples takes less than 1 minute, whereas the distortion prediction using our trained decision forest requires only ≈ 0.5 sec.

5.1.1. Results for New Synthetic Dataset

Even though we report the result for cross validation to avoid over-training, we may expect that some distortions appearing in different images are correlated and the metric just learns the distortions that are specific for that data set. To test against this possibility, we measured another dataset.

The new Contrast-Luminance-Frequency-Masking (CLFM) dataset was measured using a similar procedure as in [ČHM*12]. 13 observers provided localized markings for the visible differences in 14 image pairs. The dataset was designed to cover a wide range of problematic cases for image quality assessment in possibly few images. Such problematic cases included increments of different size and contrast, edges shown at different luminance levels, random noise patterns of different frequency and contrast, several cases of contrast masking, image pairs with pixel misalignment and noise patterns generated with a different seed for the test and reference image (see example stimuli in Fig. 6). The CLFM dataset is available at the project webpage.

Fig. 7 shows the result of the tested metrics for the new dataset. Note that our new metric was trained on the LOCCG dataset and none of the new dataset images was used for training. From the shape of the ROC curves, it is clear that the CLFM dataset is extremely challenging and the metrics mispredict in many cases. But it is interesting to notice that, on average, the proposed metric has the highest AUC value.

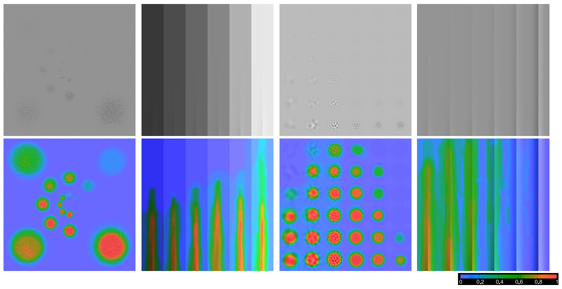

Figure 6: *CLFM: our synthetic validation dataset for testing of IQMs perceptual-masking prediction. Top row: test images containing (from left to right) increments of different size, edges at different luminance levels, and band-limited noise patterns organized in a CSF-like chart. Bottom: subjective data for the corresponding images. (Best viewed in electronic version.)*

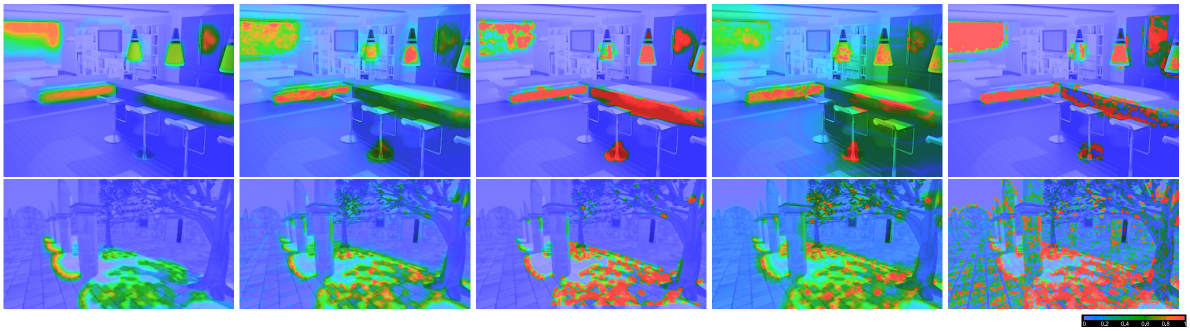

Figure 5: *Comparison of distortion maps predicted by the proposed method with the state-of-the-art metrics for the* red kitchen, *and* sponza tree shadows *scenes. From left: subjective ground-truth, prediction of the new metric, SSIM, HDR-VDP-2, and sCorrel. Please see the complete set of results in the supplementary material.*

The performance expressed as Matthew's correlation coefficient is very steady throughout the range of true positive rates, while many other metrics exhibit significant "dips". This means that the new metric is less prone to loss of performance in the worst-case scenario.

It is encouraging to observe that learning the "real-world" distortions (e.g. based on the LOCCG dataset) may enable decent prediction performance even for the synthetic dataset like CLFM. This is different from the "traditional" approach to modeling quality metrics, where the synthetic cases are used to train the metric and the assumption is made that these will generalize for complex "real-world" cases. Interestingly, this correlates with our experience – when we used synthethic CLFM data for training, it did not lead to better predictions of LOCCG than traditional metrics.

6. Optimizing Existing Metrics

The stack of classifiers described in the last paragraph of Section 4 can be used to optimize the parameters of traditional metrics for the testing datasets. We show the results for two metrics (SSIM and HDR-VDP-2) on the LOCCG dataset as an illustration of this approach.

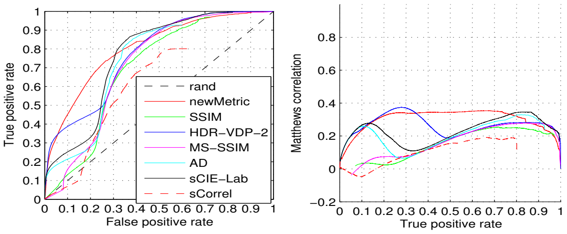

Figure 7: *Quantitative results for the new synthetic dataset (CLFM) for our metric trained on the LOCCG dataset.* $AUC_{newMetric}$*=0.805,* AUC_{SSIM}*=0.695,* $AUC_{HDRVDP2}$*=0.772,* AUC_{MSSSIM}*=0.714,* AUC_{AD}*=0.733,* $AUC_{sCIELab}$*=0.763,* $AUC_{sCorrel}$*=0.624.*

6.1. Training SSIM

The stucture similarity metric (SSIM) consists of 3 terms that were introduced in Eq. (1). The sensitivity or importance of the individual terms is controlled by the parameters α, β, and γ, which are set to 1 by default.

We optimize those 3 parameters on the LOCCG dataset with cross-validation to give the best possible prediction by employing a linear support vector machine [CL11] that computes the optimal 3D weight vector $\mathbf{w} = [\alpha^*, \beta^*, \gamma^*]^T$ for the 3 SSIM terms in the log domain $log(SSIM^*) = \alpha^* \cdot \log(l) + \beta^* \cdot \log(c) + \gamma^* \cdot \log(s) = \mathbf{w}^T \cdot \mathbf{d}^{lcs}$ by minimizing the convex objective function:

$$\arg\min_{\mathbf{w}} \sum_i \max(0, 1 - y_i \cdot \mathbf{w}^T \cdot \mathbf{d}_i^{lcs})^2 + \lambda \|\mathbf{w}\|_2^2, \qquad (3)$$

where y_i are the ground-truth labels in the dataset that are set to -1 or 1 if the distortion for the i-th training sample is visible or not, respectively, and $\mathbf{d}_i^{lcs} \in \mathbb{R}^3$ is the corresponding precomputed vector of the SSIM terms. The regularization is controlled with $\lambda = 1$.

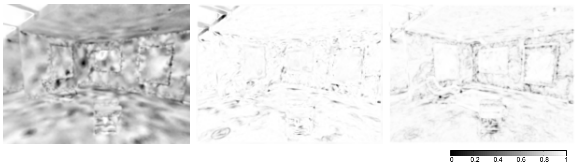

Figure 8: *An illustration of the features of SSIM for the* sala *scene where darker pixels represent more visible distortions. From left: luminance, contrast, and structure term.*

We run this optimization 35 times with randomized set of input images to assess the stability and quality of the coefficients obtained. Interestingly, the results (Fig. 9, left) show a clear tendency towards higher weighting of the structure and contrast components than the luminance component ($\alpha = 0.2$, $\beta = 2.8$, $\gamma = 3.5$). This implies that the structural and contrast components are more important than the luminance for computer graphics artifacts, which agrees with the results presented in Section 4. Please notice that the performance improvement of the new weighted metric (SSIM$_{\text{learned}}$) compared to the original SSIM in Fig. 10. An illustration of improvement of the distortion maps is shown in Fig. 11.

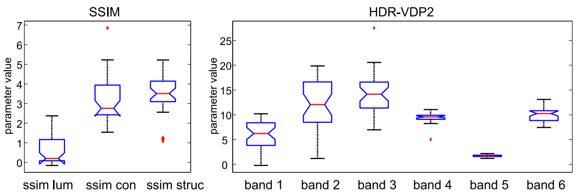

Figure 9: *The results of the optimization of SSIM (left) and HDR-VDP-2 (right) metric parameters. The red mark is the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to extreme data points not considered outliers, and outliers are plotted individually. The notches show 5% level intervals of the median significance.*

6.2. Training HDR-VDP-2

Visible differences predictor for high-dynamic-range images (HDR-VDP-2) [MKRH11] is a perceptual metric that models low-level human vision mechanisms, such as light adaptation, spatial contrast sensitivity and contrast masking. The predicted probability of detecting differences between test and reference images is modeled as psychophysical detection task separately for each spatial frequency band. The cumulative probability is computed as probability summation, which corresponds to summing logarithms of probability values from all bands. To introduce learning component to the HDR-VDP-2, we weighted the logarithmic probabilities before summation. After learning, which used the identical method as for the SSIM (Section 6.1), we found the optimum band weights to be (in decreasing frequency): w_1=6.2, w_2=12.1, w_3=14.2, w_4=9.6, w_5=1.7, w_6=10.2 (Fig. 9, right). Please notice the significant performance gain of the new weighted metric (HDR-VDP-2$_{\text{learned}}$) compared to the original HDR-VDP-2 in Fig. 10. The improved distortion maps can be found in Fig. 11.

7. Conclusions and Future Work

In this work we proposed a novel data-driven full-reference image quality metric, which outperforms existing IQMs in detecting perceivable rendering artifacts and reporting their location in a distortion map. The key element of our metric is a carefully designed set of features, which generalize over distortion types, image content, and superposition of multiple distortions in a single image. We also propose easy to use customizations of existing metrics SSIM and HDR-VDP-2 that improve their performance in predicting rendering artifacts. Finally, as the outcome of this work two new datasets have been created, which are potentially useful for the imaging and computer graphics communities. The Contrast-Luminance-Frequency-Masking (CLFM) dataset contains a continuous range of basic distortions encapsulated in a few images, with the distortion visibility annotated in a perceptual experiment. The distortion saliency maps captured in the eye tracking experiment could be used for further studies on visual attention, for example as a function of rendering distortion type and its magnitude.

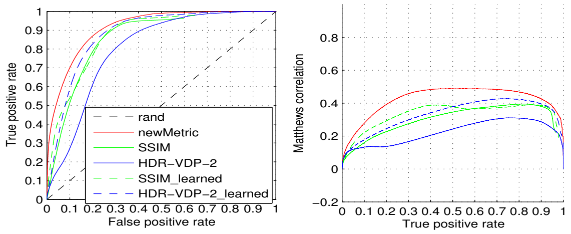

Figure 10: *Comparison of the overall results of optimized and original SSIM and HDR-VDP-2 metrics. Left: ROC, right: Matthews correlations. The bigger the area under the ROC curve (AUC), the better. AUC_{SSIM}=0.858, $AUC_{SSIM_{learned}}$=0.872, $AUC_{HDRVDP2}$=0.802, $AUC_{HDRVDP2_{learned}}$=0.883. The result of newMetric (red) is shown here for comparison.*

The main limitation of our work is the size of the training dataset, and we expect that the performance of our metric can be still improved when a larger dataset is available. Furthermore, it would be interesting to explore other supervised learning techniques, e.g. [GRHS04], both for feature selection and for FR metric prediction. The eye-tracking features deserve further exploration too: for example the combination of eye-tracking data with other features like absolute difference could indicate where people gaze due to severe artifact.

Acknowledgements

This work was partially supported by the Polish Ministry of Science and Higher Education through the grant no. N N516 508539, and by the COST Action IC1005 on “HDRi: The digital capture, storage, transmission and display of real-world lighting”.

References

[BM98] Bradley P. S., Mangasarian O. L.: Feature selection via concave minimization and support vector machines. In *Proc. of 13th International Conference on Machine Learning* (1998), pp. 82–90. 6

[Bre96a] Breiman L.: Bagging Predictors. *Machine Learning 24*, 2 (Aug. 1996), 123–140. 6

[Bre96b] Breiman L.: Stacked regressions. *Machine Learning 24* (1996), 49–64. 6

[Bre01] Breiman L.: Random forests. *Machine Learning 45*, 1 (2001), 5–32. 6

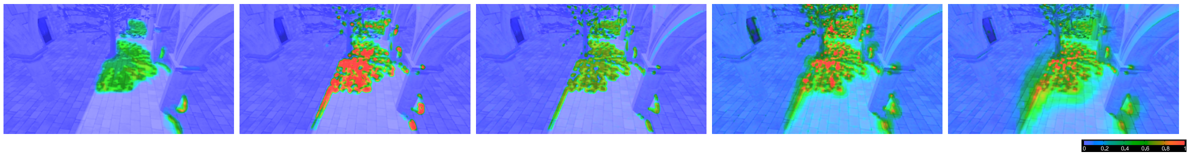

Figure 11: *An example of the improved predictions of SSIM and HDR-VDP-2 for the* sponza above tree *scene after the parameter optimization. From left: subjective ground-truth, prediction of SSIM, SSIM*$_{\text{learned}}$*, HDR-VDP-2, HDR-VDP-2*$_{\text{learned}}$.

[BZM07] Bosch A., Zisserman A., Munoz X.: Image classifcation using random forests and ferns. *In Proc. of ICCV* (2007), 1–8. 3, 4

[ČHM*12] Čadík M., Herzog R., Mantiuk R., Myszkowski K., Seidel H.-P.: New measurements reveal weaknesses of image quality metrics in evaluating graphics artifacts. *ACM TOG (Proc. of SIGGRAPH 2012)* (2012). Article 147. 1, 2, 3, 4, 5, 7

[CL11] Chang C.-C., Lin C.-J.: LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology 2* (2011), 27:1–27:27. 6, 8

[DT05] Dalal N., Triggs B.: Histograms of oriented gradients for human detection. *Proc. of IEEE Computer Vision and Pattern Recognition* (2005), 886–893. 3, 4

[FFP05] Fei-Fei L., Perona P.: A bayesian hierarchical model for learning natural scene categories. *Proc. of IEEE Computer Vision and Pattern Recognition* (2005), 524–531. 4

[GRHS04] Goldberger J., Roweis S., Hinton G., Salakhutdinov R.: Neighbourhood components analysis. In *Advances in Neural Information Processing Systems 17* (2004), MIT Press, pp. 513–520. 9

[HČA*12] Herzog R., Čadík M., Aydin T. O., Kim K. I., Myszkowski K., Seidel H.-P.: NoRM: no-reference image quality metric for realistic image synthesis. *Computer Graphics Forum 31*, 2 (2012), 545–554. 2

[HS88] Harris C., Stephens M.: A combined corner and edge detector. *Proc. of the 4th Alvey Vision Conference* (1988), 147–151. 3, 4

[Kov99] Kovesi P.: Image features from phase congruency. *Videre: A Journal of Computer Vision Research 1*, 3 (1999). 3, 4

[LBW11] Li C., Bovik A. C., Wu X.: Blind image quality assessment using a general regression neural network. *IEEE Transactions on Neural Networks 22*, 5 (2011), 793–9. 2

[LSAR10] Liu C., Sharan L., Adelson E., Rosenholtz R.: Exploring features in a bayesian framework for material recognition. *Proc. of IEEE Computer Vision and Pattern Recognition* (2010), 239–246. 5

[Lub95] Lubin J.: *Vision Models for Target Detection and Recognition. ed. E. Peli.* World Scientific, 1995, ch. A Visual Discrimination Model for Imaging System Design and Evaluation, pp. 245–283. 1

[MB10] Moorthy A., Bovik A.: A two-step framework for constructing blind image quality indices. *IEEE Signal Processing Letters 17*, 5 (2010), 513 –516. 2

[MKRH11] Mantiuk R., Kim K. J., Rempel A. G., Heidrich W.: HDR-VDP-2: a calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM TOG (Proc. of SIGGRAPH 2011)* (2011). Article 40. 1, 2, 3, 9

[NL10] Narwaria M., Lin W.: Objective image quality assessment based on support vector regression. *IEEE Transactions on Neural Networks 21*, 3 (2010), 515–9. 2

[PH11] Pedersen M., Hardeberg J.: Full-reference image quality metrics: Classification and evaluation. *Found. Trends. Comput. Graph. Vis. 7*, 1 (2011), 1–80. 1, 2

[RFWB07] Ramanarayanan G., Ferwerda J., Walter B., Bala K.: Visual equivalence: towards a new standard for image fidelity. *ACM TOG (Proc. of SIGGRAPH 2007)* (2007). 2

[RGHZ10] Redi J. a., Gastaldo P., Heynderickx I., Zunino R.: Color distribution information for the reduced-reference assessment of perceived image quality. *IEEE Trans. on Circuits and Systems for Video Techn. 20*, 12 (2010), 1757–69. 2

[RTMF07] Russell B., Torralba A., Murphy K., Freeman W. T.: Labelme: a database and web-based tool for image annotation. *International Journal of Computer Vision* (2007). 4

[SBC05] Sheikh H., Bovik A., Cormack L.: No-reference quality assessment using natural scene statistics: JPEG2000. *IEEE Trans. on Image Processing 14*, 11 (2005), 1918–1927. 2

[SBC10] Saad M., Bovik A., Charrier C.: A DCT statistics-based blind image quality index. *IEEE Signal Processing Letters 17*, 6 (2010), 583 –586. 2

[SSB06] Sheikh H., Sabir M., Bovik A.: A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Trans. on Image Processing 15*, 11 (2006), 3440–3451. 1

[SWCB06] Sheikh H. R., Wang Z., Cormack L., Bovik A. C.: LIVE image quality assessment database RLS 2, 2006. 2, 4

[TBRT09] Tuv E., Borisov A., Runger G., Torkkola K.: Feature selection with ensembles, artificial variables, and redundancy elimination. *Journal of Machine Learning Research 10* (2009), 1341–1366. 6

[TJK11] Tang H., Joshi N., Kapoor A.: Learning a blind measure of perceptual image quality. *Proc. of IEEE Computer Vision and Pattern Recognition* (2011), 305–312. 2

[TM08] Tuytelaars T., Mikolajczyk K.: Local invariant feature detectors: a survey. *Found. Trends. Comput. Graph. Vis. 3*, 3 (2008), 177–280. 2, 3

[WB06] Wang Z., Bovik A. C.: *Modern Image Quality Assessment*. Morgan & Claypool Publishers, 2006. 1, 2, 3

[WBSS04] Wang Z., Bovik A. C., Sheikh H. R., Simoncelli E. P.: Image quality assessment: From error visibility to structural similarity. *IEEE Trans. on Image Processing 13*, 4 (2004), 600–612. 3, 4

[Wid84] Widdel H.: Operational problems in analysing eye movements. *In A.G. Gale & F. Johnson (Eds.), Theoretical and Applied Aspects of Eye Movement Research. 1* (1984), 21–29. 5

[ZW97] Zhang X., Wandell B. A.: A spatial extension of CIELAB for digital color-image reproduction. *Journal of the Society for Information Display 5*, 1 (1997), 61. 3