Optica Applicata, Vol. XLIV, No. 2, 2014
DOI: 10.5277/oa140208

Visual attention pooling and understanding the structural similarity index in multi-scale analysis

BOBAN P. BONDZULIC[1*], VLADIMIR S. PETROVIC[2], SRDJAN T. MITROVIC[1], BOBAN Z. PAVLOVIC[1], MILENKO S. ANDRIC[1]

[1]Military Academy, University of Defence in Belgrade, Generala Pavla Jurisica Sturma 33, 11000 Belgrade, Serbia

[2]Imaging Science and Biomedical Engineering, University of Manchester, Oxford Rd, Manchester, M13 9PT, UK

[*]Corresponding author: bondzulici@yahoo.com

We present a novel spatial pooling strategy and the results of an extensive multi-scale analysis of the well-known structural similarity index metric (SSIM) for objective image quality evaluation. We show, in contrast with some previous studies, that even relatively simple perceptual importance pooling strategies can significantly improve objective metric performance evaluated as the correlation with subjective quality assessment. In particular, we define an attention and quality driven pooling mechanism that focuses structural comparisons within the SSIM model to only those pixels exhibiting significant structural degradations. We show that optimal objective metric performance is achieved over very sparse spatial domains indeed that ignore most of the signal data. We also investigate an explicit breakdown of the structural models within SSIM and show that in combination with the proposed attention and quality driven pooling some of these models represent well performing metrics in their own right, when applied at appropriate scale for which there may not be a single optimal value. Our experiments demonstrate that the augmented SSIM metric using the proposed pooling model provides performance advantage on an extensive LIVE dataset covering hundreds of degraded images and 5 different distortion types compared to both conventional SSIM and state-of-the-art objective quality metrics.

Keywords: image quality assessment, structural similarity index, visual attention pooling.

1. Introduction

The performance of digital imaging applications for representation of visual information and communication depends greatly on the quality of input and output images, making efficient and robust estimation of image quality a priority. It would therefore be of great advantage if a transmission system is able to quantify quality degradations that occur, so that it can maintain control and possibly enhance the quality of output data [1]. The most reliable way of assessing image quality is subjective trials, by hu-

mans as ultimate users of such images. Subjective (mean opinion) scores however are impractical, slow and expensive to obtain in most applications. An objective image quality metric can predict perceived quality computationally and can be employed to: *i*) benchmark and monitor compression and processing algorithms and *ii*) optimize their performance for a given application (content, bandwidth, packet loss...) [1]. However quantifying perceived quality objectively (computationally) is not easy.

Conventional metrics such as mean squared error (MSE) and peak signal-to-noise ratio (PSNR) are still used even though better understanding of the quality paradigm has produced more robust image/video quality metrics [2]. To be considered useful, metrics generally must demonstrate relevance by being in agreement with subjective opinions of observers. An ideal image quality measure should be able to describe the amount of distortion, the type of distortion, and the distribution of errors in a degraded image with minimal complexity [3].

Objective metrics of image quality can be classified according to the availability of an original (distortion free) image, with which the distorted image is to be compared, into three major categories, namely, no-reference (NR) or blind quality assessment, reduced-reference (RR) quality assessment, and full-reference (FR) quality assessment [3]. Each of the three categories of metrics has its own advantages and disadvantages.

Among the full-reference image quality assessment (IQA) algorithms, the structural similarity index (SSIM) [4] has become a *de facto* standard. Since natural scenes are highly structured, the human visual system (HVS) is highly adapted for extracting structural information from a scene, SSIM compares structural information between the reference and distorted images through comparisons on luminance, contrast and structure [4, 5]. The three comparisons are combined to provide a final quality score (SSIM), in the range [–1, 1].

A number of modifications have been derived from SSIM [6–15] with the aim of improving correlation with subjective ratings. In [6] a multi-scale structural similarity index (MS-SSIM) is proposed which evaluates structural similarity at different resolutions/scales. CHEN *et al*. improved SSIM by using edge information as the most important information in their gradient-based structural similarity (GSSIM) [7]. In [8], images are not compared directly, but their similarity is measured by SSIM between feature maps (corner, edge and symmetry maps). In [9] the structure term is replaced with additional terms that depend on regional statistics.

SSIM and metrics derived from it initially produce a local quality/distortion map over the image space (scene). In the second stage, local quality/distortion scores are combined over the image using a spatial pooling strategy [10]. Numerous methods have been used for spatial pooling of local SSIM scores [11–15]. Strategies used are visual fixation [11], quality-based weighting [11, 12], region-type weighting [13], information content weighting [14], and visual attention information (with eye movements) weighting [15].

Originally designed for comparison of two monochromatic images, SSIM is additionally used for quality assessment of color images, image fusion and video quality

assessments [2]. The application scope of SSIM today goes far beyond its original purpose, and it has been employed in video object tracking, coding, image denoising, image classification, and recognition [1, 2]. The performance of SSIM on different image datasets has been demonstrated in [14, 16–18].

Contributions of three SSIM components: luminance, contrast and structure to quality evaluation of common image artefacts are analyzed in [19] where a publicly subject rated image database (LIVE) was used for performance analysis of the individual components and their pairwise products.

Research presented in this paper represents an extension of SSIM analyses reported in [6, 11, 12, 19]. Regarding [6] in which the results of SSIM multi-scale analysis of two distortions (JPEG and JPEG2000 compression) are presented, this paper addresses the multi-scale analysis of three further types of distortion: white noise addition, blurring and fast fading. Comparing two analyses of the lowest SSIM scores on quality assessment in [11, 12], detail analyses of individual components and their products for five scales are presented in our research, while [19] presents similar analysis in a single scale. Our research confirms some of the findings from [6, 11, 12, 19], but also presents new significant results for image quality evaluation.

The rest of the paper is organized as following. A brief description of the structural similarity index is provided in Section 2. Section 3 presents the integration of significant local quality scores in the final quality measure. The influence of the individual components, and their products, on the image quality assessment was analyzed in detail through multi-scale analyses in Sections 4 and 5.

2. Structural similarity index

Based on an observation that HVS is highly adapted for extraction of structural information from a scene, [4] introduces a new image quality paradigm proposing that quality can be evaluated through structural similarity between two non-negative signals **x** and **y** (say, original and distorted image) measured by comparing their luminance $l(\mathbf{x}, \mathbf{y})$, contrast $c(\mathbf{x}, \mathbf{y})$ and structure $s(\mathbf{x}, \mathbf{y})$:

$$l(\mathbf{x}, \mathbf{y}) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \tag{1}$$

$$c(\mathbf{x}, \mathbf{y}) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \tag{2}$$

$$s(\mathbf{x}, \mathbf{y}) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \tag{3}$$

where μ_x and μ_y are the local sample means, σ_x^2 and σ_y^2 are the sample variances, and σ_{xy} is the correlation of local **x** and **y** samples. Constants C_1, C_2, and C_3 are used

to avoid instability when the denominator is close to zero (typically $C_1 = (K_1 L)^2$, $C_2 = (K_2 L)^2$ and $C_3 = C_2/2$, where L is the dynamic range of the pixel values – 255 for 8-bit grayscale images, $K_1 = 0.01$ and $K_2 = 0.03$).

The multiplicative form of SSIM model for comparison of **x** and **y** signals is,

$$\mathrm{SSIM}(\mathbf{x}, \mathbf{y}) = \left[l(\mathbf{x}, \mathbf{y})\right]^{\alpha} \cdot \left[c(\mathbf{x}, \mathbf{y})\right]^{\beta} \cdot \left[s(\mathbf{x}, \mathbf{y})\right]^{\gamma} \tag{4}$$

where α, β and γ are the parameters used to define the relative importance of the three components. Typically, the importance of luminance, contrast and structure terms are equal, $\alpha = \beta = \gamma = 1$. This results in a specific form of the SSIM model ($C_3 = C_2/2$),

$$\mathrm{SSIM}(\mathbf{x}, \mathbf{y}) = \frac{2\mu_x \mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \frac{2\sigma_{xy} + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \tag{5}$$

The universal image quality index (UQI) proposed in [5] corresponds to the special case of Eq. (5) where $C_1 = C_2 = C_3 = 0$.

SSIM index, Eq. (5), is evaluated at all pixels locations, by using an 11×11 circular symmetric Gaussian weighting function [4]. Overall quality is obtained as a mean of local SSIM values.

Pixel-domain full-reference examples are shown in Figs. 1–4, where the goal is to evaluate the quality of image (**b**) with a given perfect-quality reference image (**a**).

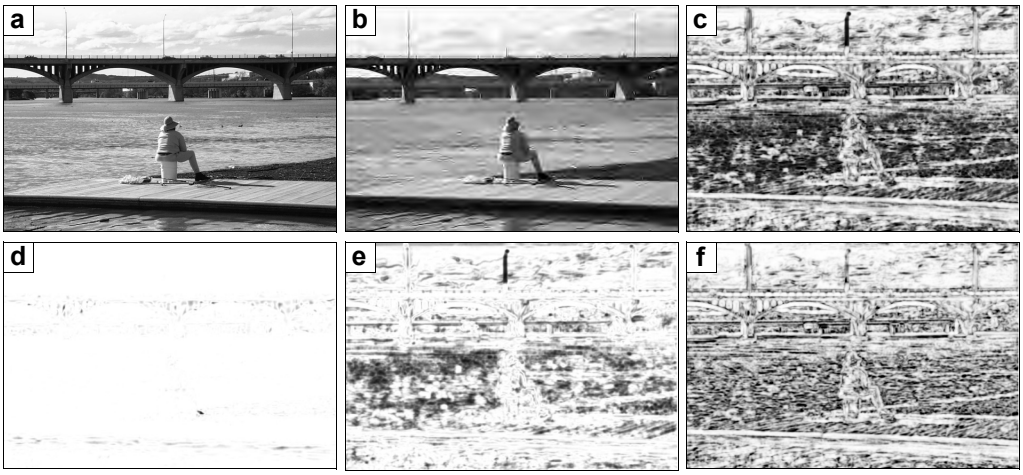

Fig. 1. Original image (**a**); distorted image (created by JPEG2000 compression) (**b**); final SSIM local quality map computed using Eq. (5) – $\mathrm{SSIM}_{\mathrm{mean}} = 0.609$ (**c**); luminance preservation map computed using Eq. (1) – $l_{\mathrm{mean}} = 0.995$ (**d**); contrast preservation map computed using Eq. (2) – $c_{\mathrm{mean}} = 0.842$ (**e**); structure preservation map computed using Eq. (3) – $s_{\mathrm{mean}} = 0.706$; subjective quality score (DMOS (differential mean opinion score)) of this image is 68.4 out of 100 (**f**).

The resulting quality/distortion maps are shown in (**c**)–(**f**) images: brighter indicates better quality (larger local quality value).

SSIM information preservation maps shown in Figs. 1**c**, 2**c**, 3**c** and 4**c** reflect the spatial variations of perceived image quality. Careful inspection shows that lumi-

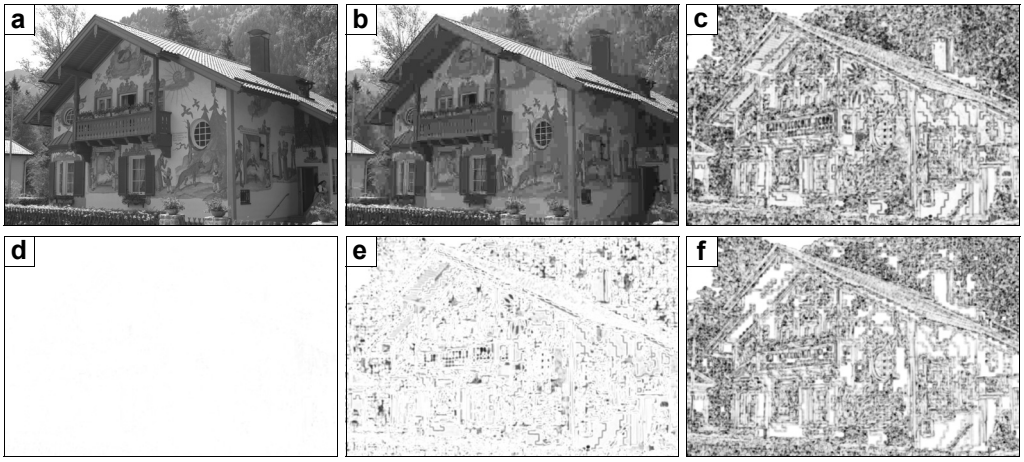

Fig. 2. Original image (**a**); distorted image (created by JPEG compression) (**b**); final SSIM local quality map computed using Eq. (5) – $\mathrm{SSIM}_{\mathrm{mean}} = 0.690$ (**c**); luminance preservation map computed using Eq. (1) – $l_{\mathrm{mean}} = 0.999$ (**d**); contrast preservation map computed using Eq. (2) – $c_{\mathrm{mean}} = 0.929$ (**e**); structure preservation map computed using Eq. (3) – $s_{\mathrm{mean}} = 0.742$; DMOS score of this image is 80.0 out of 100 (**f**).

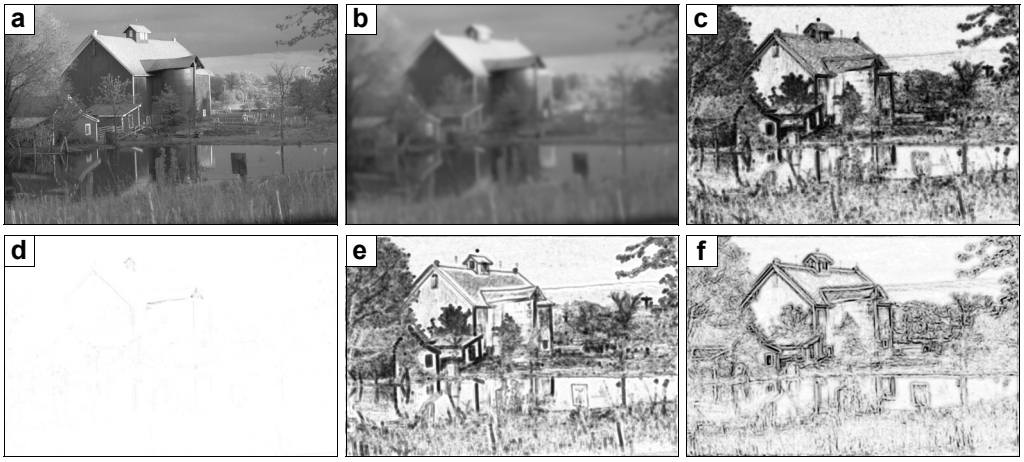

Fig. 3. Original image (**a**); distorted image (created by Gaussian blurring) (**b**); final SSIM local quality map computed using Eq. (5) – $\mathrm{SSIM}_{\mathrm{mean}} = 0.572$ (**c**); luminance preservation map computed using Eq. (1) – $l_{\mathrm{mean}} = 0.997$ (**d**); contrast preservation map computed using Eq. (2) – $c_{\mathrm{mean}} = 0.701$ (**e**); structure preservation map computed using Eq. (3) – $s_{\mathrm{mean}} = 0.785$; DMOS score of this image is 73.0 out of 100 (**f**).

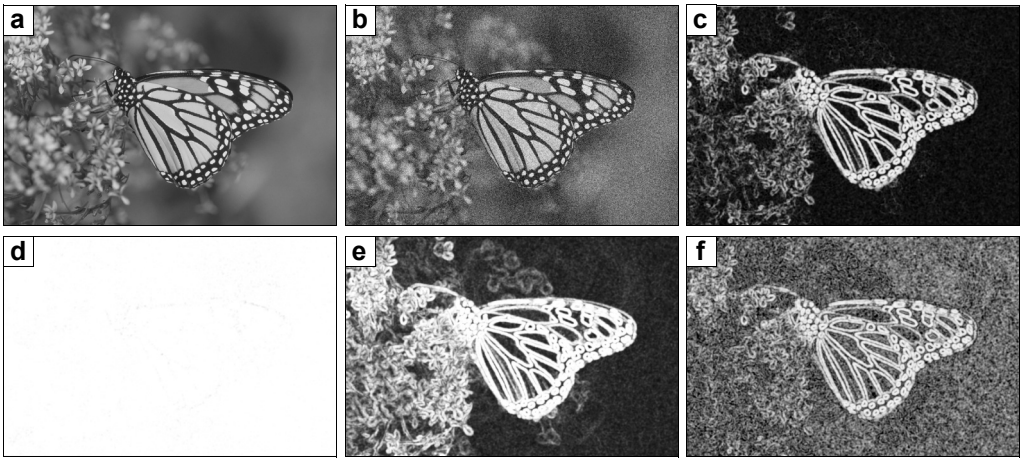

Fig. 4. Original image (**a**); distorted image (created by adding noise) (**b**); final SSIM local quality map computed using Eq. (5) – $\mathrm{SSIM}_{\mathrm{mean}} = 0.232$ (**c**); luminance preservation map computed using Eq. (1) – $l_{\mathrm{mean}} = 0.998$ (**d**); contrast preservation map computed using Eq. (2) – $c_{\mathrm{mean}} = 0.443$ (**e**); structure preservation map computed using Eq. (3) – $s_{\mathrm{mean}} = 0.447$; DMOS score of this image is 65.2 out of 100 (**f**).

nance information is almost perfectly preserved in all examples, so the SSIM map is practically a product of the contrast and structure degradation maps.

3. Visual importance pooling for image quality assessment

In this paper we explore the possibility of improving performance of the SSIM metric, its constituent components and their pairwise combinations, by assigning visual importance weights to local quality values during the pooling process. It is intuitive that different image regions may be of different importance to observers so methods that selectively pool quality scores spatially are an appealing possibility for improving SSIM scores.

There are two hypotheses which may influence human perception of image quality. The first is visual attention and gaze direction – where a human looks, the second that humans tend to perceive regions of poor quality in an image with more severity than good ones. Inspired by these observations, we investigate pooling local quality scores using the concept of perceptual importance by pooling the lowest local quality scores. These lowest scores are pooled as sample percentiles [11, 12].

In our further discussion of percentile scores, we assume that a quality map of the image has been evaluated using one of the quality metrics discussed above, Eqs. (1)–(3) and (5). Given a quality map, we arrange the local quality values obtained using Eqs. (1)–(3) and (5) (or their pairwise products) in ascending order of magnitude and a mean score is calculated at the end from the lowest $p_{\%}$ of these values. The pixels that do not fall within the percentile range are ignored.

Even with this simple approach, many weighting mechanisms are possible. Here, we consider simple percentile weighting, yet the question remains – what percentile

should be used? And how much should we weight the percentile score by? In order to determine the optimal value for $p_{\%}$, we performed an exhaustive optimization from 2% to 100% in 2% increments. Rather than using an arbitrary monotonic function of quality such as the smooth power-law function for example, we use the statistical principle of heavily weighting the extreme values – in this case, lowest percentiles. The lowest $p_{\%}$ of the quality scores are (equally) weighted, the rest of the frame scores are ignored, their weights effectively 0. Non-equal weights of rank-ordered values are also possible, but this investigation was outside the scope of this paper.

Figure 5 shows the quality assessments for four examples in Figs. 1–4 using individual SSIM components, and combined local SSIM values, Eq. (5), as a function of $p_{\%}$ of the lowest scores. Figures 5**a** and 5**b** show that sorted s values are closer to the SSIM values, while sorted c values are closer to the SSIM in Figs. 5**c** and 5**d**. The luminance is almost ideally preserved in all four examples. Significant difference in quality assessment for the individual components in the whole range of the parameter $p_{\%}$ can be observed.

Similar to the analysis performed in [20], the sorted SSIM local quality scores tend to follow a curve shape that contains a steep or nearly steep section (left side of plots on Fig. 5) corresponding to regions of severe quality degradation and a saturation sec-

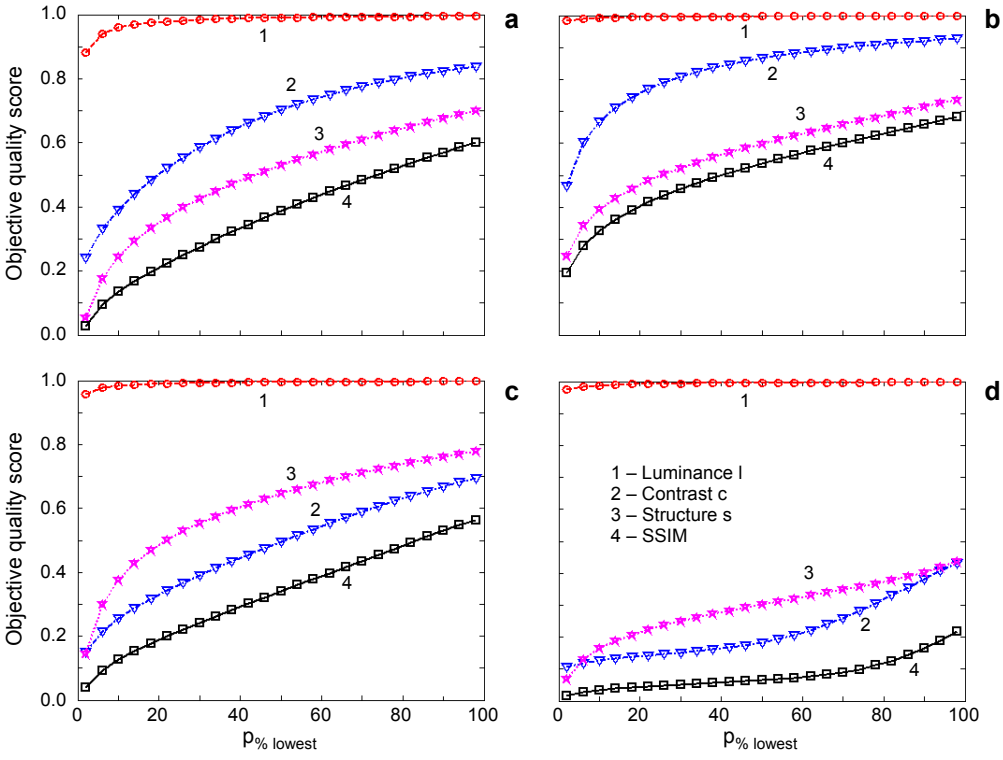

Fig. 5. SSIM and its components values sorted in ascending order for: example shown in Fig. 1 (**a**), example shown in Fig. 2 (**b**), example shown in Fig. 3 (**c**), and example shown in Fig. 4 (**d**).

tion (right side of plots on Fig. 5) that arises from regions suffering from very low degrees of quality degradation.

4. Predicting visual quality with SSIM and its constituent components

In order to demonstrate the performance of SSIM, its components and their pairwise combinations using the proposed pooling approach, we work on the LIVE image quality assessment database [21]. LIVE database contains 29 source images reflecting the diversity of image contents, distortion types and subjective quality scores for each distorted image (differential mean opinion scores, DMOS scores).

Performance of different objective quality models, Eqs. (1)–(3) and (5), and their combinations was evaluated with respect to three aspects of their ability to estimate subjective assessment of image quality [22]:

– Prediction accuracy measured using: linear correlation coefficient (CC), mean absolute error (MAE), and root-mean-square error (RMSE);

– Prediction monotonicity, measured using the Spearman rank-order correlation coefficient (SROCC);

– Prediction consistency quantified using the outlier ratio (OR), defined as percent of predictions outside $\pm 2\sigma_0$.

Logistic functions are used in a fitting procedure to provide nonlinear mapping between the objective/subjective scores as used in a benchmarking study in [22]. The nonlinearity chosen for regression for each of the models tested was a 4-parameter logistic function.

T a b l e 1. Validation scores for quality assessment models on LIVE database – scale 1 (original resolution).

	Model				
	CC	SROCC	MAE	RMSE	OR
c	0.875	0.889	10.782	13.237	10.911
c (the lowest $p_{\%}$)	0.934 (2%)	0.944 (2%)	7.788 (2%)	9.743 (2%)	2.567 (2%)
$l \times c$	0.878	0.892	10.664	13.076	10.783
$l \times c$ (the lowest $p_{\%}$)	0.935 (2%)	0.944 (2%)	7.748 (2%)	9.703 (2%)	2.567 (2%)
s	0.865	0.878	10.894	13.690	11.040
s (the lowest $p_{\%}$)	0.933 (2%)	0.938 (2%)	8.035 (2%)	9.780 (2%)	3.081 (6%)
$l \times s$	0.868	0.881	10.797	13.546	10.141
$l \times s$ (the lowest $p_{\%}$)	0.933 (4%)	0.938 (2%)	8.066 (2%)	9.819 (4%)	2.952 (6%)
$c \times s$	0.900	0.910	9.382	11.888	7.445
$c \times s$ (the lowest $p_{\%}$)	**0.942 (2%)**	**0.949 (2%)**	**7.223 (2%)**	**9.199 (2%)**	**2.054 (4%)**
SSIM (standard)	0.901	0.910	9.334	11.832	7.317
SSIM (the lowest $p_{\%}$)	0.941 (2%)	**0.949 (2%)**	7.248 (2%)	9.230 (2%)	**2.054 (4–6%)**

All available LIVE database images were used (779 images) in the validation and optimization of the similarity metrics. The metric performance was evaluated against DMOS scores using five conventional performance metrics listed above. The results at different image scales: the performance of standard structural similarity metrics as the benchmark and the dependence of this performance on the percentage of relevant $p_{\%}$ lowest quality scores of the proposed method are given in Tables 1–5 (the best metric is marked by bold). Taking the original and distorted image signals as the input, we iteratively down-sample them by a factor of 2. We index the original image as

T a b l e 2. Validation scores for different quality assessment models on LIVE database – scale 2.

	Model				
	CC	SROCC	MAE	RMSE	OR
c	0.856	0.873	11.471	14.131	13.736
c (the lowest $p_{\%}$)	0.908 (2%)	0.915 (2%)	9.170 (2%)	11.438 (2%)	4.365 (2%)
$l \times c$	0.857	0.876	11.339	13.962	13.350
$l \times c$ (the lowest $p_{\%}$)	0.909 (2%)	0.916 (2%)	9.117 (2%)	11.381 (2%)	4.108 (2%)
s	0.897	0.907	9.695	12.068	7.574
s (the lowest $p_{\%}$)	0.922 (10%)	0.925 (12%)	8.736 (10%)	10.606 (10%)	**3.723 (18%)**
$l \times s$	0.898	0.908	9.649	12.020	7.317
$l \times s$ (the lowest $p_{\%}$)	0.922 (10%)	0.925 (12%)	8.723 (10%)	10.585 (10%)	**3.723 (12–22%)**
$c \times s$	0.912	0.920	8.938	11.233	5.520
$c \times s$ (the lowest $p_{\%}$)	**0.923 (4%)**	**0.928 (12%)**	**8.232 (4%)**	10.497 (4%)	4.108 (2%)
SSIM (standard)	0.912	0.920	8.907	11.203	5.263
SSIM (the lowest $p_{\%}$)	**0.923 (6%)**	**0.928 (14%)**	8.244 (4%)	**10.493 (6%)**	3.979 (2%)

T a b l e 3. Validation scores for different quality assessment models on LIVE database – scale 3.

	Model				
	CC	SROCC	MAE	RMSE	OR
c	0.834	0.859	12.399	15.087	15.276
c (the lowest $p_{\%}$)	0.884 (2%)	0.896 (2%)	10.331 (2%)	12.787 (2%)	7.831 (2%)
$l \times c$	0.838	0.863	12.271	14.901	14.891
$l \times c$ (the lowest $p_{\%}$)	0.886 (2%)	0.898 (2%)	10.262 (2%)	12.676 (2%)	7.574 (2%)
s	0.909	0.917	9.229	11.411	5.135
s (the lowest $p_{\%}$)	0.917 (6%)	0.922 (6%)	8.832 (4%)	10.870 (6%)	4.236 (16%)
$l \times s$	0.909	0.917	9.210	11.386	5.006
$l \times s$ (the lowest $p_{\%}$)	**0.918 (6%)**	**0.923 (6%)**	**8.811 (4%)**	**10.837 (6%)**	**4.108 (14–32%)**
$c \times s$	0.908	0.916	9.199	11.461	5.006
$c \times s$ (the lowest $p_{\%}$)	0.912 (2%)	0.917 (8%)	8.882 (2%)	11.184 (2%)	4.878 (2%)
SSIM (standard)	0.908	0.917	9.186	11.441	4.750
SSIM (the lowest $p_{\%}$)	0.913 (2%)	0.918 (8%)	8.876 (2%)	11.162 (2%)	4.750 (84–100%)

T a b l e 4. Validation scores for different quality assessment models on LIVE database – scale 4.

	Model				
	CC	SROCC	MAE	RMSE	OR
c	0.804	0.839	13.164	16.261	18.100
c (the lowest $p_{\%}$)	0.843 (2%)	0.863 (2%)	12.011 (2%)	14.706 (2%)	12.452 (2%)
$l \times c$	0.809	0.843	13.011	16.057	17.458
$l \times c$ (the lowest $p_{\%}$)	0.845 (2%)	0.866 (2%)	11.969 (2%)	14.617 (2%)	12.195 (2%)
s	0.902	0.912	9.492	11.769	5.391
s (the lowest $p_{\%}$)	**0.909 (2%)**	**0.916 (2%)**	9.069 (2%)	11.397 (2%)	**5.135 (2%)**
$l \times s$	0.903	0.912	9.482	11.754	**5.135**
$l \times s$ (the lowest $p_{\%}$)	**0.909 (2%)**	**0.916 (2%)**	**9.053 (2%)**	**11.371 (2%)**	**5.135 (2%, 100%)**
$c \times s$	0.895	0.907	9.777	12.153	6.547
$c \times s$ (the lowest $p_{\%}$)	0.901 (2%)	0.908 (2%)	9.421 (2%)	11.855 (2%)	6.547 (100%)
SSIM (standard)	0.896	0.907	9.768	12.139	6.547
SSIM (the lowest $p_{\%}$)	0.901 (2%)	0.908 (2%)	9.404 (2%)	11.825 (2%)	6.419 (88–92%)

T a b l e 5. Validation scores for different quality assessment models on LIVE database – scale 5.

	Model				
	CC	SROCC	MAE	RMSE	OR
c	0.77	0.80	13.65	17.35	21.95
c (the lowest $p_{\%}$)	0.78 (2%)	0.81 (2%)	13.64 (52%)	17.00 (2%)	19.25 (2%)
$l \times c$	0.78	0.81	13.39	17.02	21.57
$l \times c$ (the lowest $p_{\%}$)	0.78 (2%)	0.82 (2%)	13.39 (100%)	16.91 (2%)	19.38 (2%)
s	**0.90**	**0.91**	**9.75**	**12.02**	**5.52**
s (the lowest $p_{\%}$)	**0.90 (100%)**	**0.91 (100%)**	**9.75 (100%)**	**12.02 (100%)**	**5.52 (100%)**
$l \times s$	**0.90**	**0.91**	**9.75**	**12.02**	**5.52**
$l \times s$ (the lowest $p_{\%}$)	**0.90 (100%)**	**0.91 (100%)**	**9.75 (100%)**	**12.02 (100%)**	**5.52 (100%)**
$c \times s$	0.89	0.90	10.21	12.54	7.06
$c \times s$ (the lowest $p_{\%}$)	0.89 (100%)	0.90 (100%)	10.21 (100%)	12.54 (100%)	6.80 (76–78%)
SSIM (standard)	0.89	0.90	10.21	12.54	7.06
SSIM (the lowest $p_{\%}$)	0.89 (100%)	0.90 (100%)	10.21 (100%)	12.54 (100%)	6.93 (70–74%)

scale 1 (typically 768×512 pixels), and the highest scale as scale 5, which is obtained after four iterations. Values of $p_{\%}$ yielding optimal metric performance are given in brackets in Tables 1–5. The luminance term for CC, SROCC, MAE, RMSE and OR is intentionally left out, because its narrow dynamic range (near perfect scores) shows no real correlation with quality.

The results show that for most scale levels and metric formulations, smaller values of $p_{\%}$ produce best performing metrics. If we leave out the highest scale where only a small amount of data remains – 48×32 pixels, the optimal $p_{\%}$ value is close to 0 indicating that a relatively small number of the worst regions provide optimal quality assessment performance. This means that quality degradations are focused within

the image and possibly present in only some regions. These regions would then determine the overall impression of quality and be vital for its objective assessment. This simple modification exhibits 4% increase in linear correlation with subjective ratings from the standard SSIM evaluation (see Table 1 and values for SSIM).

Figure 6 shows the scale 1 SSIM optimization plot obtained for all 5 different metrics. It is clear from these plots that all the metrics are in agreement on the optimal

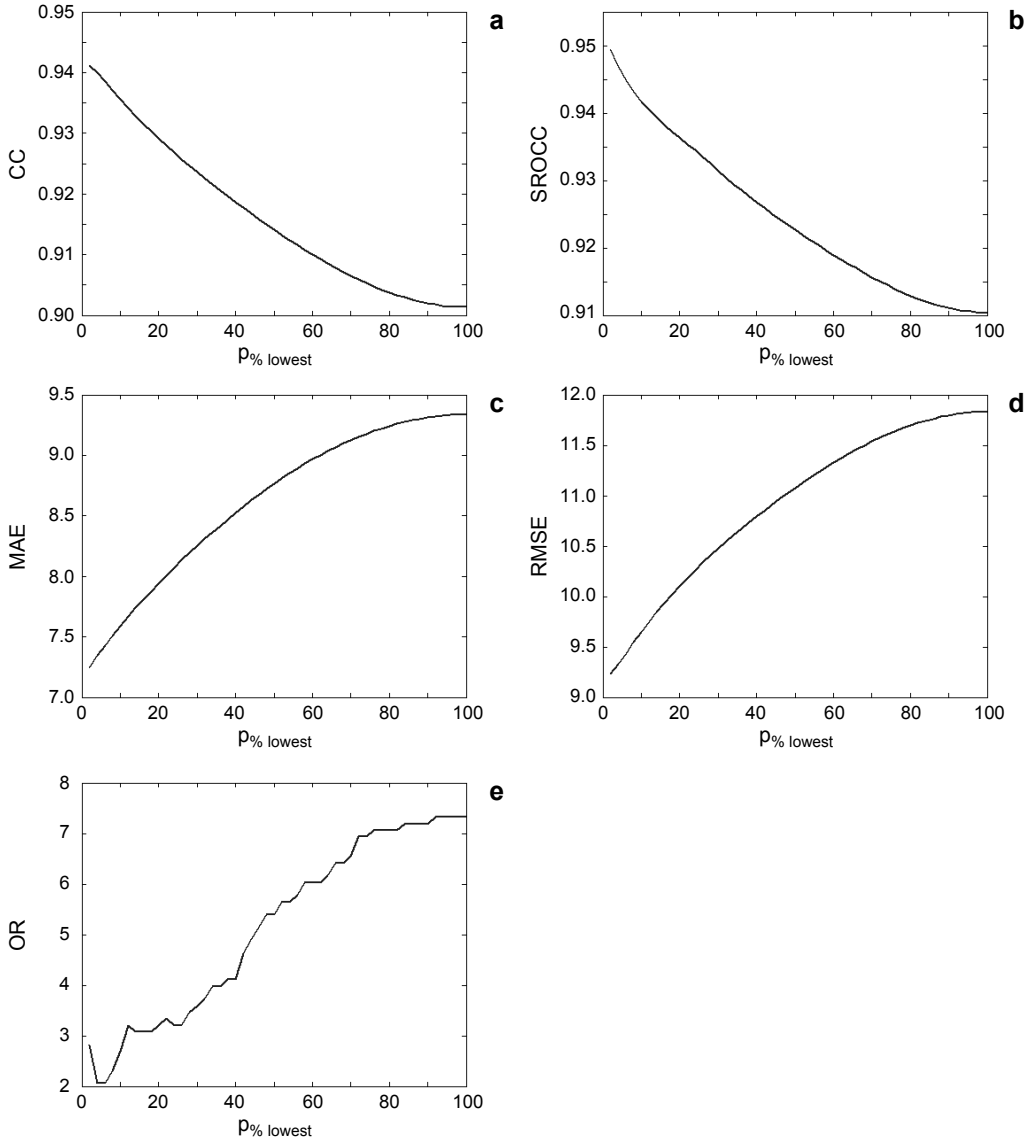

Fig. 6. Performance of the standard SSIM metric at scale 1 for a range of $p_{\%}$ values on LIVE dataset: CC (**a**), SROCC (**b**), MAE (**c**), RMSE (**d**), and OR (**e**). Note that for CC and SROCC higher values are desirable as opposed to the error terms MAE and RMSE and the outlier term OR where lower values are better.

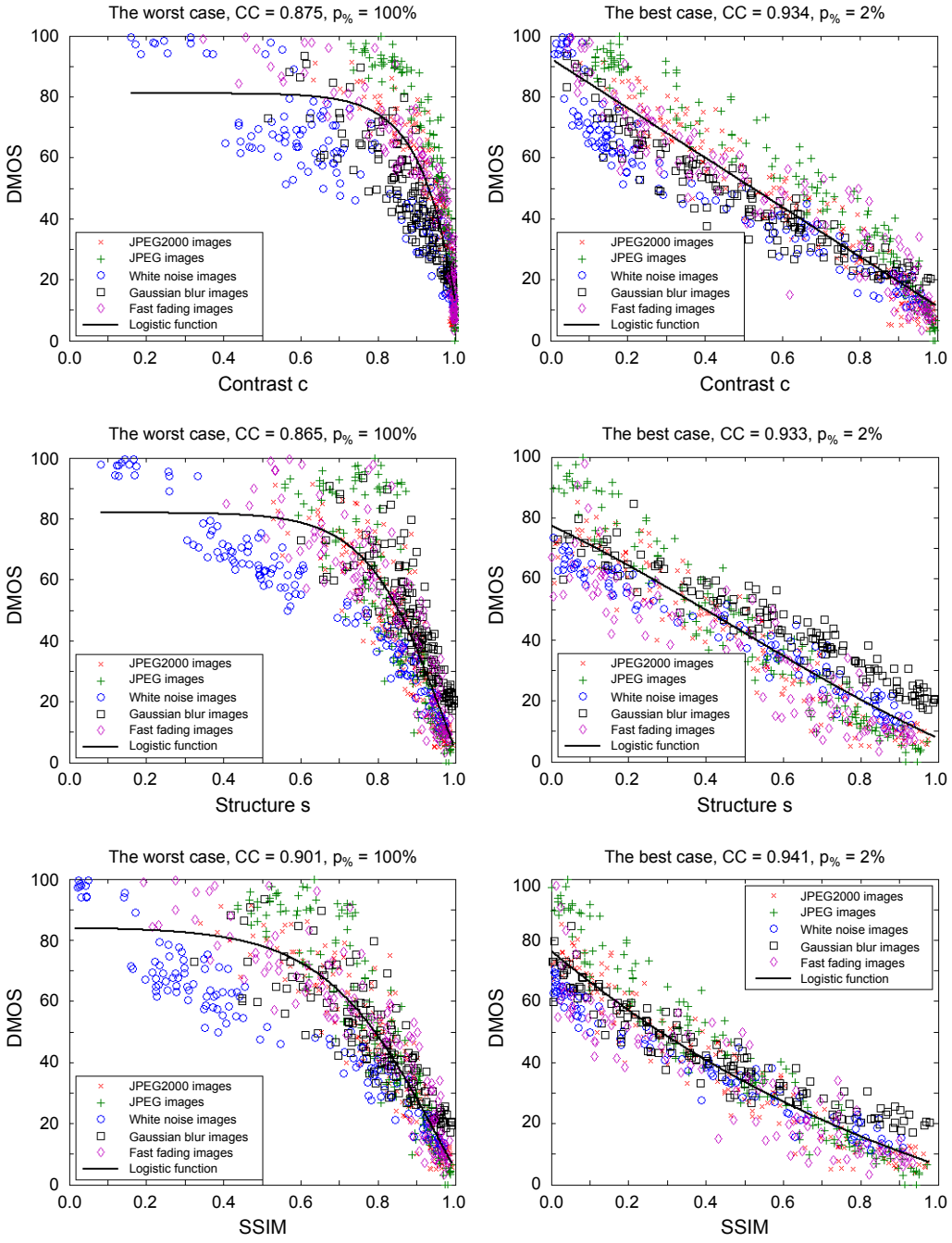

Fig. 7. Scatter plots of DMOS *versus* contrast/structure/SSIM model predictions along with logistic function fit for the best and the worst cases.

range of $p_{\%}$ values. Scale 1 is selected because the best agreement of subjective and objective scores is achieved (above 94%).

Figure 7 shows the logistic function fit for the best and the worst (standard) performing $p_{\%}$ values at scale 1 for contrast, structure and SSIM comparisons. Scatter plots of DMOS *versus* objective predictions are shown, with horizontal axes representing the objective quality score. The plots in Fig. 7 clearly show that the presented $p_{\%}$ lowest quality scores model results in a very well behaved metric response with a linear relationship between objective and subjective scores across the entire range. This is observed for both, individual contrast and structure components, and for the overall SSIM values. Moreover, spreading of the standard SSIM scores is present for low quality images (high DMOS values). Metrics with $p_{\%} = 100\%$ correspond to the conventional SSIM approach and exhibit a considerably worse, less realistic response to quality distortions requiring an additional logistic function to map the scores to a reasonable response, clearly not the case with the proposed model.

5. Analysis of results

Graphical interpretations of results from Tables 1–5 are shown in Fig. 8. Since all the metrics from Tables 1–5 show a high level of agreement, only CC and SROCC are shown for the sake of brevity. Subscript "opt" indicates the results obtained for the best (optimal) choice of the lowest $p_{\%}$ scores.

Figure 8 clearly shows that the correlation of subjective and objective scores increases when using the presented pooling approach (lowest $p_{\%}$ scores) over all scales, both for individual SSIM components and for the combined score. It also shows that performance drops with the increase in scale, reduction in resolution, which is a logical consequence of the removal of fine detail structures from the evaluation.

In terms of individual measurements, structure comparison achieves optimal subjective agreement on scale 3 while contrast comparison is best performed at top resolution. In general, performance of contrast decreases pretty linearly with a rise in scale. This invites a conclusion that a separate scale contrast and structure evaluation, *i.e.*, contrast at lower scales with structure information at higher scales, can produce more appropriate assessment in general.

According to [6], optimal agreement between subjective and SSIM scores is at scale 2, confirmed by our analysis (black lines in Fig. 8). Lower correlation coefficient obtained here, 0.91 *versus* 0.96 in [6], is a result of different database releases used, LIVE release 1 in [6] is a subset of the release 2 used here.

The finding from [19] that SSIM luminance comparison can be ignored (see Table 2 and values for $c \times s$ and SSIM) is also confirmed here, with an extension that it is valid for all five scales (see Tables 1–5 and values for $c \times s$ and SSIM).

Structure comparison is more significant than the comparison of contrast, according to [19], where SSIM components were computed at half resolution, equivalent to

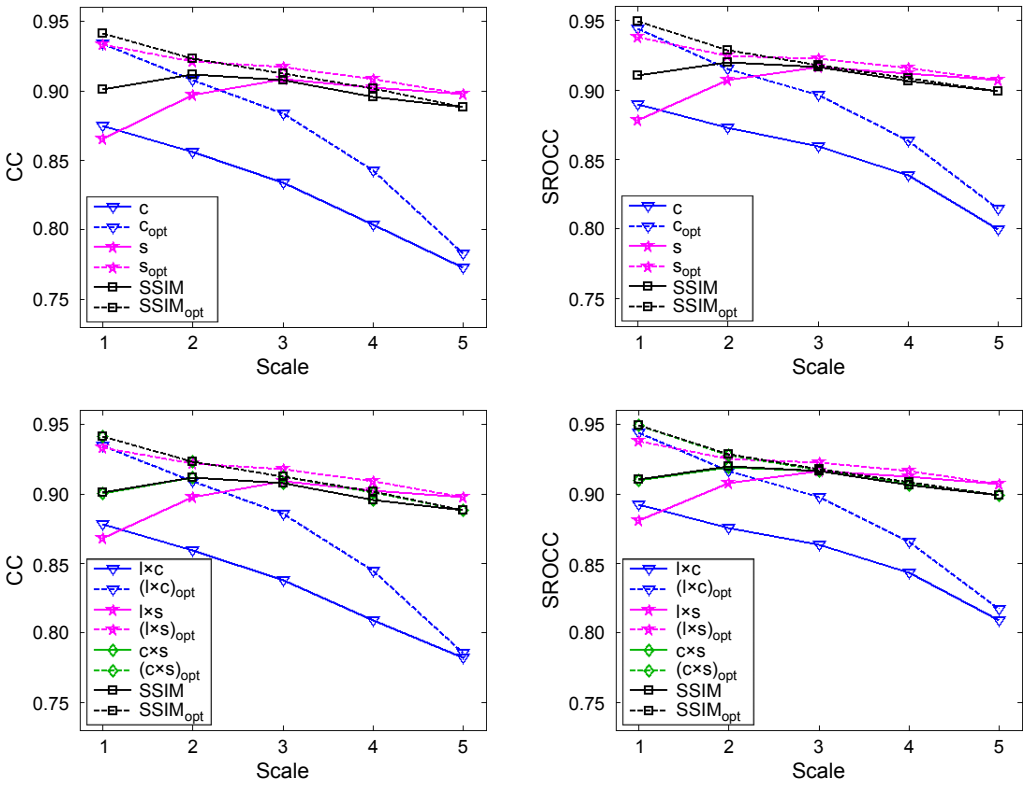

Fig. 8. Graphical illustrations of the results from Tables 1–5 obtained for different SSIM model formulations and using optimal choice of the lowest $p_{\%}$ scores.

scale 2 in our experiments. Therefore, contrast has an advantage in original resolution, leading to conclusion that components' importance depends on scale at which it is compared. However, making a binary conclusion about the usefulness of contrast over structure and *vice versa* probably does not do the SSIM approach justice. Both comparisons are important, which is illustrated in Table 6 using three additional publicly available databases (CSIQ [23], IVC [24] and Toyama [25]) for contrast, structure and final SSIM index comparisons (the best metric is marked by bold). These results show that individual SSIM component – contrast *c*, in all cases produces better results than the structure, *s* model. This leads to assumption that a novel fusion method of contrast and structure information can perhaps improve performance further, for example using a weighed additive model, with degradation dependent relative contribution of each channel (different importance), as in [12].

Multiplicative integration of individual similarity models defined in Eq. (4) has been a constant feature of all SSIM derivative metrics so far. Additive, single-scale, SSIM integration model, introduced in [12], achieved optimal agreement with subjective quality scores when only a small fraction of the worst objective scores are considered in estimation global quality. Furthermore, in [12] it is shown that within

T a b l e 6. Validation scores for different quality assessment models on CSIQ, IVC and Toyama databases.

			Model				
			CC	SROCC	MAE	RMSE	OR
Database	CSIQ	Contrast *c*	0.814	0.835	0.122	0.153	35.797
		Structure *s*	0.732	0.696	0.147	0.179	44.457
		SSIM (standard)	**0.815**	**0.837**	**0.116**	**0.152**	**33.487**
	IVC	Contrast *c*	**0.797**	**0.796**	0.573	**0.736**	18.378
		Structure *s*	0.747	0.723	0.620	0.810	22.162
		SSIM (standard)	0.792	0.779	**0.555**	0.743	**17.838**
	Toyama	Contrast *c*	**0.898**	**0.890**	**0.425**	**0.550**	**6.548**
		Structure *s*	0.760	0.747	0.643	0.813	16.071
		SSIM (standard)	0.798	0.787	0.589	0.754	14.286

T a b l e 7. Validation scores for different quality assessment models on LIVE database.

	Model				
	CC	SROCC	MAE	RMSE	OR
PSNR	0.870	0.876	10.535	13.467	10.270
VSNR	0.923	0.927	8.075	10.521	4.108
MS-SSIM	0.938	**0.953**	7.573	9.456	**2.054**
SSIM, scale 1, $p = 2\%$	**0.941**	0.949	**7.248**	**9.230**	**2.054**

the context of structural similarity, preservation of contrast and the local structure influence subjective quality roughly evenly. Our comprehensive evaluation demonstrated that the multi-scale analysis and careful selection of operating scale can further improve metric performance. So, an obvious extension of the research presented here would be to extend the additive integration model to multiple-scales.

To provide a global perspective on the performance of presented SSIM lowest scores model, we compared it to conventional (PSNR), state-of-the-art metrics, visual signal-to-noise ratio (VSNR) [26] and the multi-scale SSIM [6]. The results for the LIVE database are given in Table 7 (the best metric is marked by bold).

SSIM model offers better assessment than PSNR and VSNR. The performance of this model is comparable to MS-SSIM algorithm, such as the appropriate choice of scale and use of the lowest $p_{\%}$ scores could provide the suitable alternative to more complex multi-scale SSIM index.

6. Conclusions

We presented a novel spatial pooling strategy and a multi-scale analysis of the well-known structural similarity index for objective image quality evaluation. We have found that, in contrast with some previous studies, the perceptual importance pooling strategy can significantly improve metric performance evaluated as the correlation with

subjective quality assessment. Individual comparison models defined within SSIM, in particular contrast and local structure correlation, were shown to capable objective quality metrics on their own, provided an optimal scale is used. Proposed pooling model was demonstrated to provide performance advantage with the full SSIM model on an extensive LIVE dataset covering 779 degraded images and 5 different distortion types compared to both SSIM with conventional pooling approach and state-of-the-art objective quality metrics.

Although comprehensive evaluation demonstrated that the metric achieves high levels of subjective agreement for a wide range of data, an obvious extension of the research presented here would be to extend the evaluation model to multiple-scales rather than just selecting one optimal scale as well as allowing different structural comparison methods, contrast and local structure in particular to be combined across scales in a collaborative manner. Regional aggregation of local scores, in a two stage pooling process rather than a single global model, also promises potential improvements in metric performance.

References

[1] WANG Z., *Applications of objective image quality assessment methods*, IEEE Signal Processing Magazine **28**(6), 2011, pp. 137–142.

[2] WANG Z., BOVIK A.C., *Mean squared error: love it or leave it? A new look at signal fidelity measures*, IEEE Signal Processing Magazine **26**(1), 2009, pp. 98–117.

[3] SHNAYDERMAN A., GUSEV A., ESKICIOGLU A.M., *An SVD-based gray-scale image quality measure for local and global assessment*, IEEE Transactions on Image Processing **15**(2), 2006, pp. 422–429.

[4] WANG Z., BOVIK A.C., SHEIKH H.R., SIMONCELLI E.P., *Image quality assessment: from error visibility to structural similarity*, IEEE Transactions on Image Processing **13**(4), 2004, pp. 600–612.

[5] WANG Z., BOVIK A.C., *A universal image quality index*, IEEE Signal Processing Letters **9**(3), 2002, pp. 81–84.

[6] WANG Z., SIMONCELI E.P., BOVIK A.C., *Multi-scale structural similarity for image quality assessment*, Conference Record of the 37th IEEE Asilomar Conference on Signals, Systems and Computers, 2003, pp. 1398–1402.

[7] CHEN G.-H., YANG C.-L., XIE S.-L., *Gradient-based structural similarity for image quality assessment*, Proceedings of the IEEE International Conference on Image Processing, 2006, pp. 2929–2932.

[8] CUI L., ALLEN A.R., *An image quality metric based on corner, edge and symmetry maps*, Proceedings of the 19th British Machine Vision Conference, 2008, pp. 1–10.

[9] ZHAO X., REYES M.G., PAPPAS T.N., NEUHOFF D.L., *Structural texture similarity metrics for retrieval applications*, Proceedings of the 15th IEEE International Conference on Image Processing, 2008, pp. 1196–1199.

[10] WANG Z., SHANG X., *Spatial pooling strategies for perceptual image quality assessment*, Proceedings of the IEEE International Conference on Image Processing, 2006, pp. 2945–2948.

[11] MOORTHY A.K., BOVIK A.C., *Visual importance pooling for image quality assessment*, IEEE Journal of Selected Topics in Signal Processing **3**(2), 2009, pp. 193–201.

[12] BONDZULIC B., PETROVIC V., *Additive models and separable pooling, a new look at structural similarity*, Signal Processing **97**, 2014, pp. 110–116.

[13] LI C., BOVIK A.C., *Content-partitioned structural similarity index for image quality assessment*, Signal Processing: Image Communication **25**(7), 2010, pp. 517–526.

[14] WANG Z., LI Q., *Information content weighting for perceptual image quality assessment*, IEEE Transactions on Image Processing **20**(5), 2011, pp. 1185–1198.

[15] NINASSI A., MEUR O.L., CALLET P.L., BARBA D., *Does where you gaze on an image affect your perception of quality? Applying visual attention to image quality metric*, Proceedings of the IEEE International Conference on Image Processing, 2007, pp. II-169–II-172.
[16] SHEIKH H.R., SABIR M.F., BOVIK A.C., *A statistical evaluation of recent full reference image quality assessment algorithms*, IEEE Transactions on Image Processing **15**(11), 2006, pp. 3441–3452.
[17] TOURANCHEAU S., AUTRUSSEAU F., SAZZAD Z.M., HORITA Y., *Impact of subjective dataset on the performance of image quality metrics*, Proceedings of the 15th IEEE International Conference on Image Processing, 2008, pp. 365–368.
[18] LARSON E.C., CHANDLER D.M., *Most apparent distortion: full-reference image quality assessment and the role of strategy*, Journal of Electronic Imaging **19**(1), 2010, article 011006.
[19] ROUSE D.M., HEMAMI S.S., *Understanding and simplifying the structural similarity metric*, Proceedings of the 15th IEEE International Conference on Image Processing, 2008, pp. 1188–1191.
[20] PARK J., SESHADRINATHAN K., LEE S., BOVIK A.C., *Video quality pooling adaptive to perceptual distortion severity*, IEEE Transaction on Image Processing **22**(2), 2013, pp. 610–620.
[21] SHEIKH H.R., WANG Z., CORMACK L., BOVIK A.C., *LIVE image quality assessment database*, http://live.ece.utexas.edu/research/quality/subjective.htm, 08.03.2013.
[22] Tutorial, I.T.U.T., *Objective perceptual assessment of video quality – full reference television*, ITU-T Telecommunication Standardization Bureau ITU-T, 2004.
[23] LARSON E.C., CHANDLER D.M., *The CSIQ image database*, http://vision.okstate.edu/?loc=csiq, 08.03.2013.
[24] LE CALLET P., AUTRUSSEAU F., *Subjective quality assessment IRCCyN/IVC database*, http://www2.irccyn.ec-nantes.fr/ivcdb/, 08.03.2013.
[25] PARVEZ SAZZAD Z.M., KAWAYOKE Y., HORITA Y., *MICT image quality evaluation database*, http://mict.eng.u-toyama.ac.jp/mictdb.html, 08.03.2013.
[26] CHANDLER D.M., HEMAMI S.S., *VSNR: A wavelet-based visual signal-to-noise ratio for natural images*, IEEE Transactions on Image Processing **16**(9), 2007, pp. 2284–2298.

Received September 13, 2013
in revised form January 19, 2014