Component Weight Tuning of SSIM Image Quality Assessment Measure

Przemysław Skurowski[1] and Mateusz Janiak[2]

[1] Institute of Informatics, Silesian University of Technology, Gliwice, Poland, e-mail: przemyslaw.skurowski@polsl.pl, WWW: http://inf.polsl.pl
[2] Polish-Japanese Institute of Information Technology, Bytom, Poland, e-mail: mjaniak@pjwstk.edu.pl, WWW: http://pjwstk.edu.pl

Abstract. The article describes a method for a parameter tuning for a family of image quality assessment methods based on the concept of SSIM. The method employs curve fitting to model SSIM-MOS relationship then uses inverse relationship to calculate intended measure values and finally the log-log regressive model to estimate parameters for components constituting the measure. The regression was implemented by minimizing of least squares (L2) and least absolute deviation (L1). The results were verified against well tested reference databases.

1 Introduction

The structural similarity index [13] (SSIM) was a seminal proposal- of an image quality assessment (IQA) method. By using complex correlations, which are relatively simple mathematical tools it offers reasonably high relevance of objective measure to the subjective human quality perception responses. It was used in numerous applications and it inspired creation of various further methods which improve precision of the original measure and share the general outline of the design. These are: information weighting (IWSSIM) [14], multiscale approach (MS-SSIM) [12], incorporating color information (CID) [5] and others. Moreover, SSIM is also a basis for quality measures of other types of signals like sound [3] or video [15]. The generalized measure formula (m) and basic $SSIM$ for a pair (x, y) of signals are product of a form:

$$m(x,y) = \prod_{i=1}^{N} [f_i(x,y)]^{\alpha_i}, \qquad SSIM(x,y) = [l(x,y)]^{\alpha}[c(x,y)]^{\beta}[s(x,y)]^{\gamma} \quad \text{(1a,b)}$$

where: f_i are [0,1] valued components, describing similarity - in particular l, c, s are: luminance, contrast and structure measures respectively; α_is and in SSIM case α, β, γ are corresponding weights. The SSIM components are computed as correlation based statistical parameters:

$$l(x,y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1},\ c(x,y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2},\ s(x,y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \qquad (2)$$

where: μ_x, μ_y are mean values of compared signals, $\sigma_x, \sigma_y, \sigma_x^2, \sigma_y^2, \sigma_{xy}$ are standard deviations, variances and covariance of respective signals, C_1, C_2, C_3 are small constant values to avoid division by zero problem. The default implementation mean SSIM (MSSIM) is computed as an average of a local SSIM values (see Eq. 10a), which are computed using local, Gaussian weighted statistical parameters.

In the original paper, for the sake of proposal clarity, the weights of SSIM components in Eq. (1b) were equally set up to 1, although, as it has been demonstrated by Rouse [9] these components are not of the equal relevance to the human quality perception (see also Fig. 1). Component weight tuning was considered in dedicated experiments [12, 1] but for multiscale MS-SSIM only and to the authors knowledge that aspect of single scale SSIM have never been deliberated. Although, other tunable aspects of measure such as spatial pooling [11] and content weighting [14, 6] were studied extensively. Therefore, component weighting still remains promising field for improvement for fine tuning of the measure both in terms of accuracy and precision.

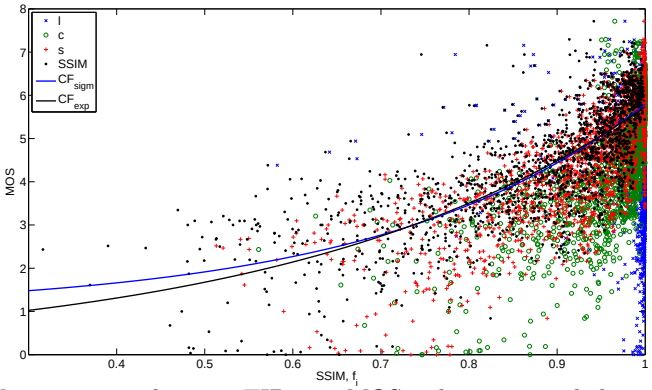

Fig. 1. SSIM components l, c, s vs TID2008 MOS judgements and their nonlinear regressive models (CF - sigmoidal and exponential)

In further parts of this paper, the proposal of regression based method for parameters computation is described. The proposal results in an improvement of both accuracy and precision of SSIM value. As a ground truth, there are used reference databases TID2008, LIVE and CSIQ.

2 The Idea

In order to verify the objective quality measure (m), its results are compared to the subjective judgments provided as MOS values (mean opinion score) which in a numerical scale represent mean values of human quality judgments of an image. The relationship between the IQAs and MOS judgments is modeled by a fitting curve (CF) - which have to be monotonic. Then the prediction (p) of an average human response for a pair of reference distorted images x, y is:

$$p_{x,y} = CF\left(m(x,y)\right) \tag{3}$$

Which has to be compared versus the $MOS(x, y)$ value. Since the MOS values should be treated as a ground truth, so any adjustments, should be done to the SSIM value. Therefore, let's define the intended measure value as:

$$\tilde{m}(x, y) = CF^{-1}\left(MOS(x, y)\right), \tag{4}$$

the value of measure (SSIM) which *should be* according to inverse fit curve (CF^{-1}) for the given x, y pair. This value can be used to formulate a regressive model on the basis of Eq. (1a) and logarithm:

$$\log(\tilde{m}(x, y)) = \sum_{i=1}^{N} \alpha_i \log(f_i(x, y)). \tag{5}$$

Taking the latter for a large enough set of (x, y) pairs, we get an overdetermined system of equations. In matrix notation, where $\tilde{M}_{\log}$ is a column vector of $\log(\tilde{m}(x, y))$s, A is a column vector of α_i-s and $F_{\log}$ contains log-component values $\log(f_i(x, y))$ as rows, the problem gets a form and least squares method solution (L2) as follows:

$$\tilde{M}_{\log} = F_{\log}\ A \tag{6}$$

$$A = (\tilde{M}'_{\log}\tilde{M}_{\log})^{-1}\tilde{M}'_{\log}F_{\log} \tag{7}$$

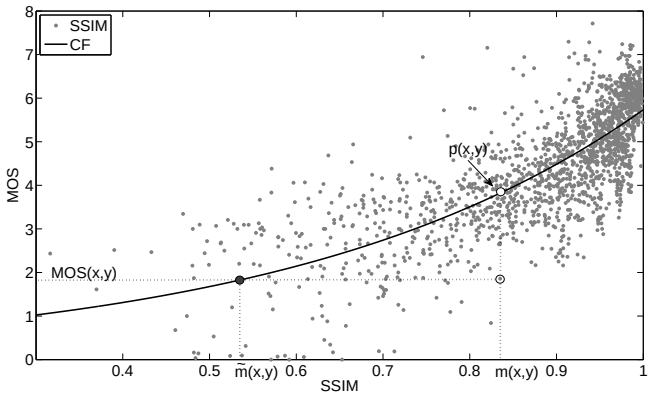

Fig. 2. Exemplary measure value m, prediction p, and intended value $\tilde{m}$

The vector A of results in Eq. (7) consists of weights to be used in Eqs. (1a,b). However, one should be slightly sceptical of the results returned by LSM, since it is sensitive to the outliers. In case of image quality reference databases the data is heavily dispersed, so it is also worth to consider alternative methods of solving the problem (Eq. (6)). Such a solution is least absolute error minimization - L1 regression that can be noted as an optimization problem:

$$A = \min_{A} |\tilde{M}_{\log} - F_{\log}\ A|, \tag{8}$$

which can be expressed as an augmented linear program (using slack variables [2]) to be calculated using an ordinary linear solver.

There is one more potential pitfall to discuss. Since the SSIM is usually (MSSIM) computed by averaging of local SSIM values, therefore l, c, s values are not to be used explicitly in the regression model, due to windowed averaging they are used indirectly. Luckily one can use the approximation - since expected value of product of two random variables X, Y is:

$$\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y) - \mathrm{cov}(X, Y) \tag{9}$$

assuming low spatial covariance of the image data, we can also suppose parameters to be of a low statistical spatial dependence. Therefore one can approximate:

$$MSSIM = \frac{1}{N}\sum_{i=1}^{N} SSIM(x_i, x_i) = \frac{1}{N}\sum_{i=1}^{N} [l_i c_i s_i] \tag{10a}$$

$$\approx \frac{1}{N}\sum_{i=1}^{N} l_i \frac{1}{N}\sum_{i=1}^{N} c_i \frac{1}{N}\sum_{i=1}^{N} s_i = \mu_l \mu_c \mu_s \tag{10b}$$

Please see Fig. 3 demonstrating the empirical verification of the above hypothesis. For the images from the TID2008 database the Pearson CC (PCC) bteween MSSIM and approximation is 0.9984 and the maximal absolute difference (MAD) is about 0.0671 and mean absolute error (MAE) is 0.0045. These values for the LIVE database are PCC: 0.9993, MAD: 0.0538, MAE: 0.004.

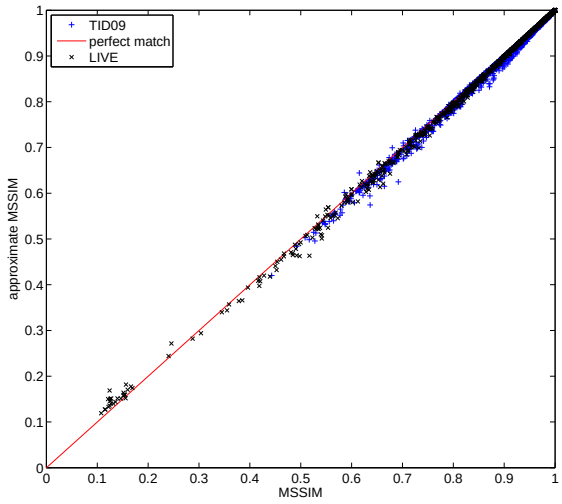

Fig. 3. Approximate MSSIM vs actual MSSIM for TID2008 and LIVE databases

3 Experiments

3.1 The data and evaluation criteria

The experimental verification of the proposal was performed using reference databases: TID2008 [7], LIVE [10] and CSIQ [4] database. All of them are commonly applied for the evaluation of IQA methods, whereas LIVE is a classic one

and both CSIQ and TID are more up-to-date and TID additionally is the most comprehensive of these, containing some odd distortions. The TID2008 database consists of 17 distortions at 4 levels for 25 ref images collected in 256428 MOS scale evaluations. The LIVE contains DMOS scale evaluations of five image distortions at various degrees for 29 reference images collected in 25000 evaluations. The CSIQ database consists of 30 images altered with 6 distoritons at 5 levels, the images were judged 5000 times in DMOS scale.

We used the TID2008 database to find out the weights and LIVE and CSIQ databases to cross-validate the results. LIVE And CSIQ were adopted from DMOS to MOS scale by subtracting ($MOS = max - DMOS$). The evaluation procedure, after [12], involved following criteria:

- Relevance to the human perception - *monotonicity* - measured using Spearman rank order CC (SROCC) - it describes the precision of a measure.
- *Accuracy* of a prediction (dispersion of results) was measured using *Coefficient of Determination* (COD) [8] which is a fraction of variance explained by the model (CF): $COD = 1 - \sum_i (y_i - \hat{y}_i)^2 / \sum_i (y_i - \bar{y})^2$. It outperforms mean absolute error or root means square error (RMS) as it provides the error with respect to the proper prediction.
- Prediction *consistency* measured with *outlier ratio* (OR) as a fraction of MOS measures oustide ± 2 standard deviations form the prediction.
- Scatter plots for visual qualitative examination.

In order to verify the results, the weights obtained for the TID database were cross-validated against LIVE database. Such a choice of the roles for the databases stems from the fact that the TID2008 is more comprehensive (17 classes of distortions) so the weights for this test set should be more general.

The key role of a CF both in the procedure and in evaluation requires careful choice of adequate curve to fit. The common choice in such cases is to use sigmoid, although fitting its parameters can be cumbersome, requiring numerous search evaluations due to start point sensitivity - to overcome that problem, for the assumed data sets, one can use the exponential function $y = ae^{bx}$ as in this case it conforms the sigmoid function very well, it achieves fine fit to the data very fast and gives robust results regardless on the starting point for the curve fitting.

3.2 The results

Optimizing SSIM measure using L1 and L2 regressions gives weights (of Eq. (1b)) provided in Table 1. In the Table 2 there are provided quality evaluation criteria for both exact and approximate values of MSSIM (Eq. 10b) for the tuned (TID) and cross validation (LIVE, CSIQ) datasets. Scatter plots of the results for these datasets are provided in Fig. 4.

3.3 Discussion of Results

The proposed method for fine tuning worked well for the original measure MSSIM and approximation as well. As one could expect L1 regression appeared

Table 1. Weights for MSSIM

Regression	α	β	γ
L2	0.1292	3.7979	1.2862
L1	0.1121	1.1640	0.8345

Table 2. Evaluation results of tuned MSSIM

A.) Approximate form

database	weights	SROCC	COD	OR
TID2008	original	0.7921	0.6019	0.0476
	L2	0.8279	0.6462	0.0476
	L1	0.8315	0.6447	0.0535
LIVE	original	0.9496	0.6743	0.0153
	L2	0.9440	0.6976	0.0224
	L1	0.9486	0.6611	0.0153
CSIQ	original	0.8792	0.7239	0.0520
	L2	0.9025	0.7431	0.0647
	L1	0.9179	0.7795	0.0600

B.) Exact formula

database	weights	SROCC	COD	OR
TID2008	original	0.7749	0.5854	0.0541
	L2	0.7775	0.5936	0.0606
	L1	0.8174	0.6314	0.0541
LIVE	original	0.9496	0.6758	0.0153
	L2	0.9442	0.7111	0.0193
	L1	0.9488	0.6634	0.0173
CSIQ	original	0.8755	0.7123	0.0554
	L2	0.8907	0.7131	0.0658
	L1	0.9124	0.7675	0.0612

to provide results better than L2, therefore they will be considered further. The cross validation of results using LIVE and CSIQ databases provided slightly ambiguous results. For CSIQ we observed improvement meanwhile tests for LIVE database reveal slight degradation. Fortunately, the optional loss in quality criteria for the worst case of cross validation was slight and can be neglected as it was approx. 1-2% of the measure value, whereas improvement for the TID and CSIQ database was 5 times bigger (5-10%).

When we observe first two criteria (SROCC and COD) improved the OR in all these cases grow a little bit. The explanation for that is in increasing of the COD which is a measure of concentration around the regression curve, so the higher it is, the lower variance of the residuals is; which are used to calculate boundaries for the outlier ratio. Apparently, fine tuning of weights is not able to reduce the outliers while improving overall dispersion of results, therefore tightening the boundaries for the OR would result in increasing the number of the measures classified as *outliers*.

Another interesting (and surprising) observation are the fine results of approximation of MSSIM which outperfoms regular MSSIM - one could favor this version of SSIM computing for its efficiency.

Another interesting, side observation is overall resilience of a LIVE database to the tuning. In additional test we tried to tune the SSIM weights to that database but with no success - the improvement of quality criteria was marginal. Apparently the LIVE database seems to be fit to the SSIM measure (or SSIM to the database). This raises a question whether we should use LIVE to evaluate SSIM - luckily the number image quality databases has grown over years so that is not a limitation.

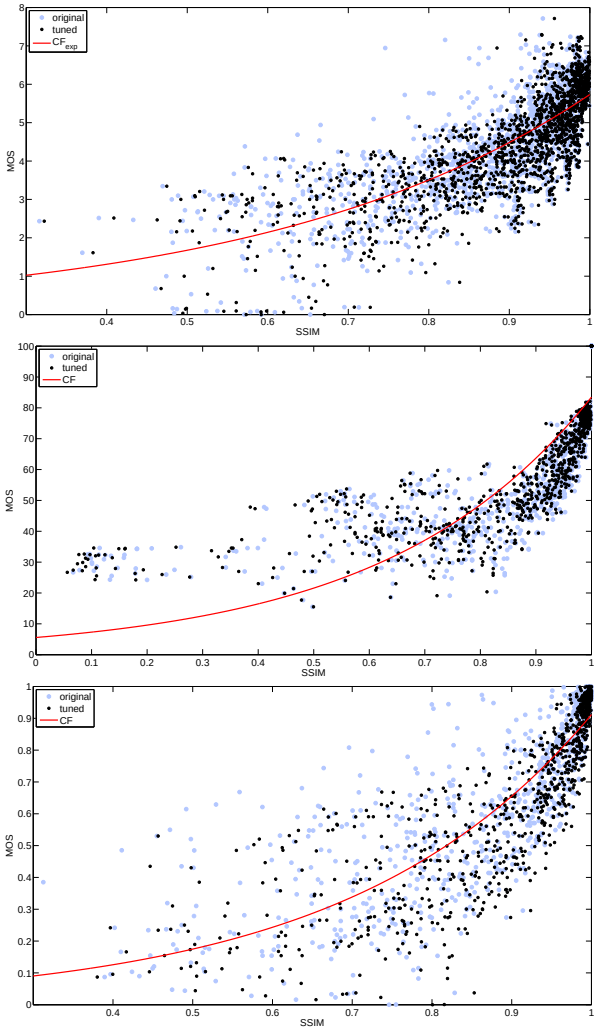

Fig. 4. Scatter plot of original and L1 tuned MSSIM for: TID (a), LIVE (b), CSIQ (c)

4 Summary

The proposed approach demonstrated its ability to improve performance of the SSIM measure. Its serious advantage is that it allows to estimate weights on basis of the massive human judgments without painstaking collecting human responses or need to deliberate measures' specific parts. On the other hand, estimating weights on the basis of the the rough structure of the measure can also a drawback of proposed method. It is relatively easy to overcome that limitation by posing extended requirements regarding the weights. A prior knowledge of the measure designer can be included into the weight estimation by defining minimization program where one can use constraints for ensuring required properties such as non-negativity or summation of certain weights to one.

Further progress would involve tests using further databases and applying the method to IQA measures of a design similar to SSIM. Another interesting idea to check is to verify if the iterative tuning is able to improve the results.

Acknowledgements

Costume for acquisition of human movement based on IMU sensors with collection, visualization and data analysis software. This project has been supported by Applied Research Programme of NCRD. (project ID 178438 path A)

References

1. Charrier, C., Knoblauch, K., Maloney, L.T., Bovik, A.C., Moorthy, A.K.: Optimizing multiscale ssim for compression via mlds. IEEE Transactions on Image Processing 21(12), 4682–4694 (2012)
2. D'Errico, J.: Optimization tips and tricks (May 2011), http://www.mathworks.com/matlabcentral/fileexchange/8553
3. Kandadai, S., Hardin, J., Creusere, C.: Audio quality assessment using the mean structural similarity measure. In: IEEE International Conference on Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. pp. 221–224 (2008), 00024
4. 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)
5. Lissner, I., Preiss, J., Urban, P., Lichtenauer, M.S., Zolliker, P.: Image-difference prediction: From grayscale to color. IEEE Transactions on Image Processing 22(2), 435–446 (Feb 2013)
6. Moorthy, A.K., Bovik, A.C.: Perceptually significant spatial pooling techniques for image quality assessment. In: Proc. SPIE 7240, Human Vision and Electronic Imaging XIV. vol. 7240, pp. 724012–724012–11 (2009), 00029
7. Ponomarenko, N., Lukin, V., Zelensky, A., Egiazarian, K., Carli, M., Battisti, F.: Tid2008 a database for evaluation of full reference visual quality assessment metrics. Advances of Modern Radioelectronics 10(4), 3045 (2009)
8. Rice, J.A.: Mathematical statistics and data analysis. Duxbury advanced series, Thomson/Brooks/Cole, Belmont, CA, 3rd ed edn. (2007)
9. Rouse, D.M., Hemami, S.S.: Understanding and simplifying the structural similarity metric. In: 2008 IEEE Int. Conf. on Image Processing. p. 11881191 (2008)
10. Sheikh, H., Sabir, M., Bovik, A.: A statistical evaluation of recent full reference image quality assessment algorithms. IEEE Transactions on Image Processing 15(11), 3440–3451 (Nov 2006)
11. Wang, Z., Shang, X.: Spatial pooling strategies for perceptual image quality assessment. In: 2006 IEEE Int. Conf. on Image Processing. pp. 2945–2948 (2006)
12. Wang, Z., Simoncelli, E., Bovik, A.: Multiscale structural similarity for image quality assessment. In: Conference Records of the 37th Asilomar Conference on Signals, Systems and Computers. vol. 2, pp. 1398–1402. IEEE (2004)
13. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing 13(4), 600–612 (2004)
14. Wang, Z., Li, Q.: Information content weighting for perceptual image quality assessment. IEEE Transactions on Image Processing 20(5), 1185–1198 (May 2011)
15. Wang, Z., Lu, L., Bovik, A.C.: Video quality assessment based on structural distortion measurement. Signal processing: Image communication 19(2), 121132 (2004)