Learning Data-driven Image Similarity Measure

Takumi Kobayashi
National Institute of Advanced Industrial Science and Technology
1-1-1 Umezono, Tsukuba, Japan
E-mail: takumi.kobayashi@aist.go.jp

***Abstract*—Image quality assessment gains a greater interest due to development of digital imaging and storage. In that field, structural similarity (SSIM) index has been shown to favorably agree with human perceptual assessment, significantly outperforming the method of mean squared error, *i.e.*, L_2 distance. The similarity measure function in SSIM which compares a target (distorted) image with its reference (original) image is hand-crafted in a simple form via a top-down approach based on the human visual system. It, however, might lack optimality without directly considering the relationships between image data and the perceptual assessment (scores). In this paper, we propose a method to construct an image similarity measure based on actual data. The proposed method optimizes a similarity measure function by exploiting annotated data in a bottom-up and data-driven manner, while retaining the favorable property of structural similarity in SSIM. The non-linear similarity function is optimized as the global optimum of high generalization power. In addition, the proposed method is simply formulated and thus applicable to the family of SSIM, especially to FSIM which has been recently proposed exhibiting superior performance to SSIM. The experimental results on image quality assessment demonstrate the effectiveness of the proposed method compared to the other methods.**

I. Introduction

In image processing, it is a fundamental task to assess quality of images which are subject to various types of distortion caused such as through acquisition, compression, storage and transmission. Such quality assessment can be applied not only to images but also to a variety of signals. The quality is essentially defined based on subjective human perception, not solely on physical property of signals, *i.e.*, pixel values, themselves. The subjective assessment, however, requires large amounts of cost and time, and thus we demand automatic quality assessment which is rather *objective*. Given a source (reference) image, the target (test) image would be assessed via comparison with it. Although mean squared error (MSE) has been successfully applied in the other fields to compare two signals, it is unfortunately unsuitable for the quality assessment due to its incompatibility with human perception [1].

In the last decade, structural similarity index (SSIM) [2] has drawn keen attention for image quality assessment since it significantly outperforms MSE [1]. SSIM is composed of three kinds of patch-based similarity between reference and test images; structural, luminance and contrast similarities. The patch-based similarity is pooled over a whole image to provide single similarity measure which can be regarded as quality measure of the test image. Apart from simple MSE, SSIM that exploits structural similarity in patches is closely related to human perception [1], thus producing favorable performance for automatic image quality assessment. There are variants of SSIM, such as multi-scale SSIM (MS-SSIM) [3], complex wavelet SSIM [4] and information content-weighted SSIM (IW-SSIM) [5], and it is noteworthy that feature similarity index (FSIM) [6] is recently proposed and outperforms the other family of SSIM. The mathematical properties of SSIM are also analyzed in detail by [7]. SSIM is so versatile as to be widely applicable in various fields other than image quality assessment, *e.g.*, image fusion, image denoising, watermarking and compression. Note that these SSIM-based methods are called fully-referenced methods due to assessing image quality based on the similarity measure between reference and test images, while there also exist blind approaches without requiring reference images [8], [9]. This study focuses on the framework of fully-referenced image quality assessment.

In this paper, we propose a novel method for learning image similarity measure from annotated data. SSIM is formulated in a fully top-down manner (so we can call this *hand-crafted* measure) without considering actual data distribution nor its relationship to human perceptual assessment (scores). Thus, it might be said that the SSIM and its variants lack optimality from the data-driven (bottom-up) point of view. We first present a general framework of SSIM and then, in that framework, combine top-down and bottom-up approaches to construct *optimized* similarity measure based on data while leveraging top-down structural similarity measure, the key component of SSIM. In the proposed method, we formulate the similarity learning problem in a convex form to globally optimize *non-linear* similarity measure function. The bottom-up data-driven approach is, however, vulnerable to overfitting. In order to avoid it and improve generalization, model complexity of a similarity measure function is reduced and a prior model is also introduced. The proposed method provides similarity measure in such a simple form as SSIM, due to which it requires a low computational cost to assess image quality. In addition, the simple formulation of the proposed method enables us to extend the method to the variants of SSIM beyond the original SSIM; in this work, we apply it to FSIM [6] which produces high performance.

II. Proposed method

We begin with a brief review of SSIM [2] that measures similarity between reference and test images (Sec. II-A), and then propose a method to learn similarity measure from annotated data (Sec. II-B). We also describe the extension of the method in the framework of FSIM [6] (Sec. II-C).

Preprint submitted to 23rd International Conference on Pattern Recognition.
Received April 20, 2016.

A. Structured similarity index (SSIM)

Given two image patches $\boldsymbol{x}$ and $\boldsymbol{y}$ to be compared, SSIM [1], [2] provides similarity measure between them as follows;

$$\mathcal{S}(\boldsymbol{x},\boldsymbol{y}) = \mathtt{l}(\mathtt{u}(\boldsymbol{x}),\mathtt{u}(\boldsymbol{y}))\,\mathtt{c}(\mathtt{q}(\boldsymbol{x}),\mathtt{q}(\boldsymbol{y}))\,\mathtt{s}(\boldsymbol{x},\boldsymbol{y}), \tag{1}$$

$$\mathtt{l}(\mu_x,\mu_y)=\mathtt{k}(\mu_x,\mu_y;c_1),\ \mathtt{c}(\sigma_x,\sigma_y)=\mathtt{k}(\sigma_x,\sigma_y;c_2), \tag{2}$$

$$\mathtt{k}(a,b;c)=\frac{2ab+c}{a^2+b^2+c},\ \mathtt{s}(\boldsymbol{x},\boldsymbol{y})=\frac{\mathtt{r}(\boldsymbol{x},\boldsymbol{y})+c_3}{\mathtt{q}(\boldsymbol{x})\mathtt{q}(\boldsymbol{y})+c_3}, \tag{3}$$

where $\mathtt{u}(\boldsymbol{x})$, $\mathtt{q}(\boldsymbol{x})$ and $\mathtt{r}(\boldsymbol{x},\boldsymbol{y})$ are functions to compute mean, standard deviation and covariance of pixel values in patches $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively. While the function $\mathtt{s}$ simply computes the correlation coefficient, the function $\mathtt{k}$ is applied to measure similarities of scalars, in this case, the mean μ and standard deviation σ; the small constants c_1, c_2 and c_3 are introduced to avoid instable computation. A similarity between two images $\mathcal{I}^x$ and $\mathcal{I}^y$ is then computed by pooling the patch-based SSIM indexes (1) over a whole image;

$$\bar{\mathcal{S}}(\mathcal{I}^x,\mathcal{I}^y) = \frac{1}{m}\sum_{j=1}^{m}\mathcal{S}(\boldsymbol{x}_j,\boldsymbol{y}_j), \tag{4}$$

where $\{\boldsymbol{x}_j,\boldsymbol{y}_j\}_{j=1}^m$ are m patch pairs densely sampled from the pair of $\mathcal{I}^x$ and $\mathcal{I}^y$.

Three functions $\mathtt{l},\mathtt{c}$ and $\mathtt{s}$ in (2,3) measure similarity of *luminances*, *contrasts* and *structures* in the patches, respectively. The structural similarity $\mathtt{s}(\boldsymbol{x},\boldsymbol{y})$ plays a key role in SSIM by extracting pairwise pixel relationship to take into account a patch structure which is closely related to perceptual similarity [1]. It, however, is too robust in pixel value changes to give favorable similarity measure since it always produces maximum similarity score (*i.e.*, 1) for affine relationships between pixel values, $y_k = \alpha x_k + \beta$, $(\alpha > 0)$. To compensate it, the other two types of similarities $\mathtt{l}$ and $\mathtt{c}$ are complementarily introduced to capture changes regarding luminance (pixel value bias, β) and contrast (pixel value scaling, α).

B. Similarity measure learning

The function $\mathtt{k}$ to compute the luminance and contrast similarities in (2) is not necessarily optimal for actual digital image data. Therefore, we aim to optimize the similarity measuring function based on actual data. From the viewpoint that the luminance and contrast similarities work for compensating the structural similarity $\mathtt{s}(\boldsymbol{x},\boldsymbol{y})$, a patch similarity is generally formulated by

$$\mathcal{L}(\boldsymbol{x},\boldsymbol{y}) = \mathtt{w}(\mathtt{u}(\boldsymbol{x}),\mathtt{u}(\boldsymbol{y}),\mathtt{q}(\boldsymbol{x}),\mathtt{q}(\boldsymbol{y}))\,\mathtt{s}(\boldsymbol{x},\boldsymbol{y}), \tag{5}$$

$$\bar{\mathcal{L}}(\mathcal{I}^x,\mathcal{I}^y) = \frac{1}{m}\sum_{j=1}^{m}\mathcal{L}(\boldsymbol{x}_j,\boldsymbol{y}_j)+\rho, \tag{6}$$

where a bias ρ is added for more generalization and $\mathtt{w}$ is a *non-linear* function of four statistics $\mathtt{u}(\boldsymbol{x})$, $\mathtt{u}(\boldsymbol{y})$, $\mathtt{q}(\boldsymbol{x})$ and $\mathtt{q}(\boldsymbol{y})$; in the case of SSIM, it is constructed by using the function $\mathtt{k}$ as $\mathtt{w}(\mu_x,\mu_y,\sigma_x,\sigma_y) = \mathtt{l}(\mu_x,\mu_y)\mathtt{c}(\sigma_x,\sigma_y) = \mathtt{k}(\mu_x,\mu_y;c_1)\mathtt{k}(\sigma_x,\sigma_y;c_2)$. We learn the function $\mathtt{w}$ from annotated image pairs, assuming that a training set is composed of image pairs which are coupled with perceptual similarity scores, such as mean opinion score (MOS) and differential MOS (DMOS).

1) Feature representation: It is generally difficult to optimize the *non-linear* function $\mathtt{w}$ in (5) since we do not impose any constraints on the form of $\mathtt{w}$. However, it is possible to formulate a feasible optimization problem for $\mathtt{w}$ by *linearizing* (6) to

$$\bar{\mathcal{L}}(\mathcal{I}^x,\mathcal{I}^y)=\iiiint_{\substack{\mu_x,\mu_y\\ \sigma_x,\sigma_y}}\mathtt{w}(\mu_x,\mu_y,\sigma_x,\sigma_y)\mathtt{f}(\mu_x,\mu_y,\sigma_x,\sigma_y)+\rho, \tag{7}$$

$$\mathtt{f}(\mu_x,\mu_y,\sigma_x,\sigma_y)=\sum_{j=1}^{m}\frac{\mathtt{s}(\boldsymbol{x}_j,\boldsymbol{y}_j)}{m}\delta(\mu_x-\mathtt{u}(\boldsymbol{x}_j))\delta(\mu_y-\mathtt{u}(\boldsymbol{y}_j)) \times\delta(\sigma_x-\mathtt{q}(\boldsymbol{x}_j))\delta(\sigma_y-\mathtt{q}(\boldsymbol{y}_j)), \tag{8}$$

where δ is a delta function. In this formulation, a similarity of an image pair is measured by the inner product of the weighting function $\mathtt{w}$ and the feature function $\mathtt{f}$ which is extracted from the image pair by using the structural similarity $\mathtt{s}$. In practice, the mean and standard deviation values are discretized; in the case that a pixel value ranges in $\{0,\cdots,255\}$, those statistics values can also be discretized into $\{0,\cdots,255\}$. The weight $\mathtt{w}$ and the feature $\mathtt{f}$ are accordingly discretized from functional forms into vectors of fixed dimensionality, $\boldsymbol{w}$ and $\boldsymbol{f}$, respectively. As a result, we can obtain the tractable problem for optimizing the weight *vector* $\boldsymbol{w}$ as described in Sec. II-B2.

It, however, is vulnerable to over-learning, resulting in less generalization performance, since the wight vector $\boldsymbol{w}$ is of high dimensionality; in the above-mentioned setting, the dimensionality of $\boldsymbol{w}$ is $256^4 \approx 4\times10^9$. Therefore, we reduce complexity of the model (5) by

$$\mathcal{L}'(\boldsymbol{x},\boldsymbol{y})=\{\mathtt{w}_\mu(\mathtt{u}(\boldsymbol{x}),\mathtt{u}(\boldsymbol{y}))+\mathtt{w}_\sigma(\mathtt{q}(\boldsymbol{x}),\mathtt{q}(\boldsymbol{y}))\}\,\mathtt{s}(\boldsymbol{x},\boldsymbol{y}), \tag{9}$$

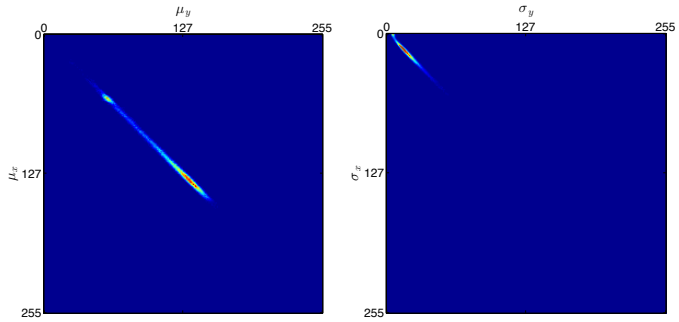
where the joint function $\mathtt{w}(\mu_x,\mu_y,\sigma_x,\sigma_y)$ is replaced with sum of the marginal functions $\mathtt{w}_\mu(\mu_x,\mu_y)$ for mean and $\mathtt{w}_\sigma(\sigma_x,\sigma_y)$ for standard deviation by considering the distinct statistical nature of μ and σ. Similar decomposition of SSIM is also found in [7]. Consequently, in a manner similar to (8), we linearize the similarity computation as

$$\bar{\mathcal{L}}'(\mathcal{I}^x,\mathcal{I}^y) = \iint_{\mu_x,\mu_y}\mathtt{w}_\mu(\mu_x,\mu_y)\mathtt{f}_\mu(\mu_x,\mu_y) + \iint_{\sigma_x,\sigma_y}\mathtt{w}_\sigma(\sigma_x,\sigma_y)\mathtt{f}_\sigma(\sigma_x,\sigma_y)+\rho, \tag{10}$$

$$\mathtt{f}_\mu(\mu_x,\mu_y)=\sum_{j=1}^{m}\frac{\mathtt{s}(\boldsymbol{x}_j,\boldsymbol{y}_j)}{m}\delta(\mu_x-\mathtt{u}(\boldsymbol{x}_j))\delta(\mu_y-\mathtt{u}(\boldsymbol{y}_j)), \tag{11}$$

$$\mathtt{f}_\sigma(\sigma_x,\sigma_y)=\sum_{j=1}^{m}\frac{\mathtt{s}(\boldsymbol{x}_j,\boldsymbol{y}_j)}{m}\delta(\sigma_x-\mathtt{q}(\boldsymbol{x}_j))\delta(\sigma_y-\mathtt{q}(\boldsymbol{y}_j)). \tag{12}$$

Preprint submitted to 23rd International Conference on Pattern Recognition.
Received April 20, 2016.



(a) $\boldsymbol{f}_\mu$ (b) $\boldsymbol{f}_\sigma$

Fig. 1. Example of marginal features $\boldsymbol{f}_\mu$ and $\boldsymbol{f}_\sigma$. For convenience, features are represented in a matrix form of 256×256, which is actually unfolded into a vector form of $256^2 = 65536$ dimensionality. The feature values are depicted in pseudo color. This figure is best viewed in color.

This is discretized into

$$\bar{\mathcal{L}}'(\mathcal{I}^x, \mathcal{I}^y) = \begin{bmatrix} \boldsymbol{w}_\mu \\ \boldsymbol{w}_\sigma \end{bmatrix}^\top \begin{bmatrix} \boldsymbol{f}_\mu \\ \boldsymbol{f}_\sigma \end{bmatrix} + \rho = \boldsymbol{w}^\top \boldsymbol{f} + \rho, \quad (13)$$

where $\boldsymbol{w}_\mu$ and $\boldsymbol{w}_\sigma$ are weight vectors discretized from the functions $\mathtt{w}_\mu$ and $\mathtt{w}_\sigma$, while $\boldsymbol{f}_\mu$ and $\boldsymbol{f}_\sigma$ are discretized feature vectors (Fig. 1), *e.g.*, $\boldsymbol{w}_\mu, \boldsymbol{w}_\sigma, \boldsymbol{f}_\mu, \boldsymbol{f}_\sigma \in \mathbb{R}^{65536=256^2}$. This marginal model is of significantly lower complexity compared to the joint model (8) in which feature dimensionality is 256^4.

2) Optimization formulation: Suppose that we have n annotated image data $\{\mathcal{I}_i^*, \mathcal{I}_i, t_i\}_{i=1}^n$, where the i-th triplet is composed of the reference image $\mathcal{I}_i^*$, the test image $\mathcal{I}_i$ (distorted from $\mathcal{I}_i^*$) and the quality assessment score t_i for $\mathcal{I}_i$. We here assume the score t is given in the form of absolute assessment scores, *e.g.*, MOS, and will discuss the case of DMOS in the later part of this section. For optimizing the weight $\boldsymbol{w}$ (and the bias ρ), we regard an image quality t as a similarity between reference and test images and thereby estimate it according to (13);

$$t \approx \boldsymbol{w}^\top \boldsymbol{f}(\mathcal{I}^*, \mathcal{I}) + \rho, \quad (14)$$

where $\boldsymbol{f}(\mathcal{I}^*, \mathcal{I})$ is the feature vector extracted from the pair of I^* and I as described in the previous section.

(14) is a form of linear regression, and based on the large-margin criterion [10], the optimization problem with respect to $\boldsymbol{w}$ and ρ can be formulated as the support vector regression (SVR) [11];

$$\min_{\boldsymbol{w},\rho,\ \xi_i^+ \geq 0, \xi_i^- \geq 0} \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^n (\xi_i^+ + \xi_i^-), \quad (15)$$

$$s.t.\ \boldsymbol{w}^\top \boldsymbol{f}(\mathcal{I}_i^*, \mathcal{I}_i) + \rho \leq t_i + \epsilon_i + \xi_i^-, \quad (16)$$

$$\boldsymbol{w}^\top \boldsymbol{f}(\mathcal{I}_i^*, \mathcal{I}_i) + \rho \geq t_i - \epsilon_i - \xi_i^+, \quad (17)$$

where ϵ_i indicates insensitivity to t_i; practically, ϵ is set to standard deviation of MOS. This SVR formulation produces the weight vector $\boldsymbol{w}$ which would be overly fit to training data especially on small-scale datasets. Actually, the optimized weight $\boldsymbol{w}$ is represented by the weighted sum of the feature vectors $\{\boldsymbol{f}_i\}_{i=1}^n$ and thus the weights on unseen mean μ and standard deviation σ are obviously zeros, which could cause less generalization.

For improving the generalization performance, we introduce a prior weight vector $\boldsymbol{w}_0$ to give the following optimization formulation;

$$\min_{\boldsymbol{w},\rho,\ \xi_i^+ \geq 0, \xi_i^- \geq 0} \frac{1}{2}\|\boldsymbol{w} - \boldsymbol{w}_0\|_2^2 + C\sum_{i=1}^n (\xi_i^+ + \xi_i^-), \quad (18)$$

$$s.t.\ \boldsymbol{w}^\top \boldsymbol{f}(\mathcal{I}_i^*, \mathcal{I}_i) + \rho \leq t_i + \epsilon_i + \xi_i^-, \quad (19)$$

$$\boldsymbol{w}^\top \boldsymbol{f}(\mathcal{I}_i^*, \mathcal{I}_i) + \rho \geq t_i - \epsilon_i - \xi_i^+, \quad (20)$$

which optimizes the weight $\boldsymbol{w}$ by minimizing the regression error (the second term) while simultaneously making it close to the prior weight $\boldsymbol{w}_0$ (the first term). Since the method stems from SSIM [2], it is natural to employ as the prior model the similarity functions used in SSIM,

$$\mathtt{w}_{\mu 0}(\mu_x, \mu_y) = \frac{\mathtt{l}(\mu_x, \mu_y)}{2},\ \mathtt{w}_{\sigma 0}(\sigma_x, \sigma_y) = \frac{\mathtt{c}(\sigma_x, \sigma_y)}{2}, \quad (21)$$

and these functions are discretized into the prior weight vector $\boldsymbol{w}_0$. By transforming the variable $\boldsymbol{w}$ into $\hat{\boldsymbol{w}} = \boldsymbol{w} - \boldsymbol{w}_0$, the above problem (18) results in the standard SVR formulation;

$$\min_{\hat{\boldsymbol{w}},\rho,\ \xi_i^+ \geq 0, \xi_i^- \geq 0} \frac{1}{2}\|\hat{\boldsymbol{w}}\|_2^2 + C\sum_{i=1}^n (\xi_i^+ + \xi_i^-), \quad (22)$$

$$s.t.\ \hat{\boldsymbol{w}}^\top \boldsymbol{f}(\mathcal{I}_i^*, \mathcal{I}_i) + \rho \leq \hat{t}_i + \epsilon_i + \xi_i^-, \quad (23)$$

$$\hat{\boldsymbol{w}}^\top \boldsymbol{f}(\mathcal{I}_i^*, \mathcal{I}_i) + \rho \geq \hat{t}_i - \epsilon_i - \xi_i^+, \quad (24)$$

where $\hat{t}_i = t_i - \boldsymbol{w}_0^\top \boldsymbol{f}(\mathcal{I}_i^*, \mathcal{I}_i)$.

On the other hand, the quality is often assessed as a relative score, *e.g.*, DMOS, unlike the above-mentioned absolute score of MOS. Based on the regression form (14), the relative score is also estimated by our model as

$$\begin{aligned} \tilde{t} &= \{\boldsymbol{w}^\top \boldsymbol{f}(\mathcal{I}^*, \mathcal{I}^*) + \rho\} - \{\boldsymbol{w}^\top \boldsymbol{f}(\mathcal{I}^*, \mathcal{I}) + \rho\} \\ &= \boldsymbol{w}^\top \{\boldsymbol{f}(\mathcal{I}^*, \mathcal{I}^*) - \boldsymbol{f}(\mathcal{I}^*, \mathcal{I})\}, \end{aligned} \quad (25)$$

where the constant bias ρ is canceled out. Thus, the optimization problem to deal with DMOS is slightly different from (22-24) only in that the feature vectors $\boldsymbol{f}(\mathcal{I}^*, \mathcal{I})$ are replaced with the differential features $\boldsymbol{f}(\mathcal{I}^*, \mathcal{I}^*) - \boldsymbol{f}(\mathcal{I}^*, \mathcal{I})$ and the bias ρ is removed in the constraints (23,24);

$$s.t.\ \hat{\boldsymbol{w}}^\top \{\boldsymbol{f}(\mathcal{I}_i^*, \mathcal{I}_i^*) - \boldsymbol{f}(\mathcal{I}_i^*, \mathcal{I}_i)\} \leq \hat{t}_i + \epsilon_i + \xi_i^-, \quad (26)$$

$$\hat{\boldsymbol{w}}^\top \{\boldsymbol{f}(\mathcal{I}_i^*, \mathcal{I}_i^*) - \boldsymbol{f}(\mathcal{I}_i^*, \mathcal{I}_i)\} \geq \hat{t}_i - \epsilon_i - \xi_i^+. \quad (27)$$

C. Extension to feature similarity index (FSIM)

While Sec. II-B has described the proposed method in the framework of SSIM, the method can be easily extended to the family of SSIM, especially to the feature similarity index (FSIM) [6] which exhibits superior performance even to SSIM. The FSIM is formulated based on both the phase congruency [12] and the gradient magnitude on images as follows;

$$\mathcal{S}_j = \mathtt{k}(G_j^x, G_j^y; c_4)\, \mathtt{k}(P_j^x, P_j^y; c_5)\, \hat{P}_j, \quad (28)$$

$$\hat{P}_j = \max[P_j^x, P_j^y], \quad (29)$$

Preprint submitted to 23rd International Conference on Pattern Recognition.
Received April 20, 2016.

where G_j^x and G_j^y are gradient magnitudes by applying Scharr operator [13] at the j-th pixel position on the images $\mathcal{I}^x$ and $\mathcal{I}^y$, respectively, and P_j^x and P_j^y are the phase congruency computed by means of [14]. As is the case for μ and σ in SSIM, the function $\mathtt{k}$ in (3) is applied to measure similarities for those two types of characteristics with constants c_4 and c_5. In contrast to SSIM, the FSIM leverages the phase congruency for exploiting local structures, which is also a biologically plausible model [12]. The FSIM for the images $\mathcal{I}^x$ and $\mathcal{I}^y$ is defined as

$$\bar{\mathcal{S}}(\mathcal{I}^x, \mathcal{I}^y) = \frac{\sum_{j=1}^m \mathcal{S}_j}{\sum_{j=1}^m \hat{P}_j}, \tag{30}$$

where m is the number of pixels in the image. The pixel-wise FSIM scores (28) are pooled and normalized by the sum of $\hat{P}_j$. In this case of FSIM, we can formulate the features in a manner similar to the procedure presented in Sec. II-B1. That is, by considering the weight $\hat{P}$ instead of $\mathtt{s}$ in SSIM,

$$\mathtt{f}_g(g_x, g_y) = \sum_{j=1}^m \frac{\hat{P}_j}{\sum_{j'} \hat{P}_{j'}} \delta(g_x - G_j^x)\delta(g_y - G_j^y), \tag{31}$$

$$\mathtt{f}_p(p_x, p_y) = \sum_{j=1}^m \frac{\hat{P}_j}{\sum_{j'} \hat{P}_{j'}} \delta(p_x - P_j^x)\delta(p_y - P_j^y), \tag{32}$$

which are actually discretized into the feature vectors $\boldsymbol{f}_g$ and $\boldsymbol{f}_p$ as in (13).

The FSIM is extended to cope with *colored* images, called FSIMc [6], while the SSIM (1) and the above-mentioned FSIM (28) are defined on gray-scaled images. Given the colored images, the RGB color channels are first converted into YIQ color representation [15] in which Y conveys the luminance information while I and Q reflect the chrominic characteristics. The similarities regarding the I and Q channels are measured by means of the function $\mathtt{k}$ in the same manner for P and G, while the FSIM (28) is directly applied to measure similarity on the Y channel. Those three types of similarity are finally integrated into

$$\mathcal{S}_j = \mathtt{k}(G_j^x, G_j^y; c_4)\, \mathtt{k}(P_j^x, P_j^y; c_5) \times \{\mathtt{k}(I_j^x, I_j^y, c_6)\, \mathtt{k}(Q_j^x, Q_j^y, c_7)\}^\lambda \hat{P}_j, \tag{33}$$

where I_j^x, I_j^y, Q_j^x and Q_j^y are the I and Q channel values at the j-th position on the images $\mathcal{I}^x$ and $\mathcal{I}^y$, respectively, and λ is a parameter usually set to $\lambda = 0.03$ [6]. Due to such a small λ, the term of I and Q channels are less contributing to the similarity measure and thus we incorporate it into the weighting with $\hat{P}$ for constructing the feature representation in the proposed method;

$$\mathtt{f}_g(g_x, g_y) = \sum_{j=1}^m \frac{\{\mathtt{k}(I_j^x, I_j^y, c_6)\, \mathtt{k}(Q_j^x, Q_j^y, c_7)\}^\lambda \hat{P}_j}{\sum'_j \hat{P}_{j'}} \delta(g_x - G_j^x)\delta(g_y - G_j^y), \tag{34}$$

$$\mathtt{f}_p(p_x, p_y) = \sum_{j=1}^m \frac{\{\mathtt{k}(I_j^x, I_j^y, c_6)\, \mathtt{k}(Q_j^x, Q_j^y, c_7)\}^\lambda \hat{P}_j}{\sum'_j \hat{P}_{j'}} \delta(p_x - P_j^x)\delta(p_y - P_j^y). \tag{35}$$

TABLE I
DATASETS USED IN THE EXPERIMENT. MOS STANDS FOR MEAN OPINION SCORE, WHILE DMOS IS DIFFERENTIAL MEAN OPINION SCORE AS RELATIVE MEASUREMENT.

Dataset	score	image pair	reference image	distortion type
TID2013 [16]	MOS	3,000	25	24
TID2008 [17]	MOS	1,700	25	17
CISQ [18]	DMOS	866	30	6
LIVE [19]	DMOS	779	29	5

As a result, through the above definitions of features $\mathtt{f}$ in (31,32) for FSIM and (34,35) for FSIMc, the optimization procedure can be applied in the same way as described in Sec. II-B2 to learn FSIM/FSIMc-based similarity measure.

III. EXPERIMENTAL RESULTS

We tested the proposed method on four datasets, TID2013 [16], TID2008 [17], CISQ [18] and LIVE [19]. The details of these datasets are described in Table I; the former two datasets are provided with MOS while the scores in the latter two are measured as DMOS. For more details, refer to the respective papers.

For fair comparison, the proposed method is applied in a leave-"one reference image"-out manner; that is, one reference image and its corresponding test (distorted) images are used only for test while the other pairs are fed into training, and the procedure is repeated for all reference images. As a result, quality scores of all test images are estimated by the proposed method excluding them in learning. Performance of the quality assessment is measured based on two kinds of rank correlation coefficient, Kendall's and Spearman's, between the estimated scores and the human perceptual scores of MOS or DMOS.

We first compare variants of the proposed method, all of which lean the similarity function $\mathtt{w}$ from training data. There are four conceivable variants regarding a feature model and a prior in learning; the features would be either of joint model (8) or marginal model (9), while we can optimize similarity function with or without a prior (18,15). Note that the marginal model is of a lower complexity than the joint model in terms of the (discretized) feature dimensionality in (13). Table II shows the performance results of these variants in the frameworks of SSIM [2], FSIM and FSIMc [6]. One can see that the performance is favorably improved by introducing the prior on all the datasets. It is generally difficult to build a proper similarity function from scratch by using such a limited number of training samples and thus the introduction of the prior facilitates to effectively learn it. In addition, the prior can suppress over-fitting in the learning especially for the joint model of high complexity; the performance of the joint model is more effectively boosted by introducing the prior, compared to the marginal model. In particular, the joint model surpasses the marginal one in the discriminative framework of FSIM/FSIMc on the datasets of larger size, TID2013 and TID2008, though the marginal model is still competitive to the joint one with outperforming the original similarity measures, FSIM and FSIMc.

Preprint submitted to 23rd International Conference on Pattern Recognition.
Received April 20, 2016.

TABLE II
PERFORMANCE COMPARISON ON VARIANTS OF THE PROPOSED METHOD.

(a) Kendall's rank correlation coefficient

		Ours based on SSIM			
		joint		marginal	
Dataset	SSIM	w/o prior	w/ prior	w/o prior	w/ prior
TID2013	0.5588	0.5113	0.5981	0.5931	**0.6359**
TID2008	0.5768	0.4894	0.6139	0.5959	**0.6605**
CISQ	0.6900	0.5848	0.6756	0.7447	**0.7700**
LIVE	0.7963	0.6997	0.7787	0.8164	**0.8188**

		Ours based on FSIM			
		joint		marginal	
Dataset	FSIM	w/o prior	w/ prior	w/o prior	w/ prior
TID2013	0.6289	0.6720	**0.7043**	0.6833	0.6901
TID2008	0.6946	0.6940	**0.7353**	0.7156	0.7292
CISQ	0.7561	0.7709	0.7931	0.8136	**0.8177**
LIVE	0.8337	0.7491	0.8016	0.8285	**0.8288**

		Ours based on FSIMc			
		joint		marginal	
Dataset	FSIMc	w/o prior	w/ prior	w/o prior	w/ prior
TID2013	0.6665	0.6729	**0.7126**	0.6858	0.6999
TID2008	0.6991	0.6942	**0.7372**	0.7153	0.7303
CISQ	0.7684	0.7710	0.7976	0.8137	**0.8189**
LIVE	0.8363	0.7481	0.8043	0.8286	**0.8306**

(b) Spearman's rank correlation coefficient

		Ours based on SSIM			
		joint		marginal	
Dataset	SSIM	w/o prior	w/ prior	w/o prior	w/ prior
TID2013	0.7417	0.6765	0.7853	0.7724	**0.8175**
TID2008	0.7749	0.6523	0.7994	0.7807	**0.8413**
CISQ	0.8755	0.7653	0.8644	0.9173	**0.9332**
LIVE	0.9479	0.8825	0.9376	0.9536	**0.9555**

		Ours based on FSIM			
		joint		marginal	
Dataset	FSIM	w/o prior	w/ prior	w/o prior	w/ prior
TID2013	0.8015	0.8558	**0.8810**	0.8617	0.8677
TID2008	0.8805	0.8757	**0.9075**	0.8924	0.9017
CISQ	0.9242	0.9307	0.9453	0.9570	**0.9588**
LIVE	0.9634	0.9205	0.9481	**0.9590**	0.9589

		Ours based on FSIMc			
		joint		marginal	
Dataset	FSIMc	w/o prior	w/ prior	w/o prior	w/ prior
TID2013	0.8510	0.8567	**0.8889**	0.8645	0.8775
TID2008	0.8840	0.8759	**0.9087**	0.8921	0.9023
CISQ	0.9309	0.9308	0.9477	0.9570	**0.9592**
LIVE	0.9645	0.9199	0.9494	0.9591	**0.9598**

TABLE III
PERFORMANCE COMPARISON ON VARIOUS FORMULATIONS FOR ESTIMATING DMOS IN THE FSIMC-BASED FRAMEWORK.

(a) Kendall's rank correlation coefficient

	Ours based on FSIMc			
	joint w/ prior		marginal w/ prior	
Dataset	w/o bias	w/ bias	w/o bias	w/ bias
CISQ	0.7976	0.7311	**0.8189**	0.7908
LIVE	0.8043	0.7645	**0.8306**	0.8040

(b) Spearman's rank correlation coefficient

	Ours based on FSIMc			
	joint w/ prior		marginal w/ prior	
Dataset	w/o bias	w/ bias	w/o bias	w/ bias
CISQ	0.9477	0.9095	**0.9592**	0.9466
LIVE	0.9494	0.9274	**0.9598**	0.9509

Next, we go into the case of DMOS measurement which is found in LIVE and CISQ datasets. The DMOS is based on a relative score and the proposed method to estimate DMOS is accordingly defined as the difference between the scores excluding the bias in (27). Nonetheless, it is possible to put the bias term by force into the DMOS estimation (27). We compare in Table III the methods with and without the bias term in the framework of the FSIMc with a prior. It is observed that the method without the bias outperforms the one including the bias. This result validates the proposed formulation (27) for estimating DMOS.

The proposed methods are then compared to the other methods, SSIM [2], FSIM [6], FSIMc [6], MS-SSIM [3], VIF [20] and IW-SSIM [5]. The proposed methods are equipped with the prior in the framework of FSIMc. Table IV shows the performance comparison results, demonstrating that the proposed methods are superior to the others except on the LIVE dataset; on that dataset, the performance is saturated and the performance difference between the original FSIMc and the proposed method of marginal model is quite small. Thus, from the viewpoint of the overall generalization performance, we can conclude that the *marginal* model *with* a prior is suitable to the proposed method, which produces favorable performance on all the datasets. Note that the proposed method is formulated in the simple framework as in SSIM/FSIM that computes an image similarity via multiplying the structural similarity $\mathtt{s}$ or the phase congruency $\hat{P}$ by the weight function output $\mathtt{w}$, and thus it requires quite a low computation cost to assess images by the proposed method.

Finally, we show in Fig. 2 the estimated scores in comparison with human perceptual scores, MOS and DMOS, on the datasets of TID2013 and CISQ. One can see that the scores estimated by the proposed method approximate human perceptual scores well, while SSIM/FSIMc exhibits highly nonlinear relationship to it as pointed out in [2], [6]. Therefore, the SSIM/FSIMc scores are often transformed by fitting a logistic function to revise such nonlinearity. On the other hand, the proposed method produces favorable scores that are directly compatible to human perceptual measure; the estimated scores are *linearly* related to MOS and DMOS.

IV. CONCLUSION

In this paper, we have proposed a novel method to construct image similarity based on data for image quality assessment. The proposed method learns a similarity function from data in a bottom-up manner while retaining the favorable property of SSIM/FSIM which is defined in a top-down manner; from this perspective, the proposed method combines bottom-up and top-down approaches. To avoid overfitting caused by the bottom-up approach, we reduce the model complexity as well as introduce a prior to the model. In the experiments using publicly available datasets, the proposed method exhibited favorable performance compared to the other methods.

Preprint submitted to 23rd International Conference on Pattern Recognition.
Received April 20, 2016.

TABLE IV
PERFORMANCE COMPARISON WITH THE OTHER METHODS. OUR METHOD IS EQUIPPED WITH A PRIOR IN THE FRAMEWORK OF FSIMC.

Dataset	Measure	SSIM	FSIM [6]	FSIMc [6]	MS-SSIM [3]	VIF [20]	IW-SSIM [5]	Ours-joint	Ours-marginal
TID2013	Kendall	0.5588	0.6289	0.6665	0.6079	0.4567	0.5977	**0.7126**	0.6999
TID2013	Spearman	0.7417	0.8015	0.8510	0.7872	0.6084	0.7779	**0.8889**	0.8775
TID2008	Kendall	0.5768	0.6946	0.6991	0.6568	0.5863	0.6636	**0.7372**	0.7303
TID2008	Spearman	0.7749	0.8805	0.8840	0.8542	0.7496	0.8559	**0.9087**	0.9023
CISQ	Kendall	0.6900	0.7561	0.7684	0.7393	0.7537	0.7529	0.7976	**0.8189**
CISQ	Spearman	0.8755	0.9242	0.9309	0.9133	0.9195	0.9213	0.9477	**0.9592**
LIVE	Kendall	0.7963	0.8337	**0.8363**	0.8044	0.8270	0.8175	0.8043	0.8306
LIVE	Spearman	0.9479	0.9634	**0.9645**	0.9513	0.9632	0.9567	0.9494	0.9598

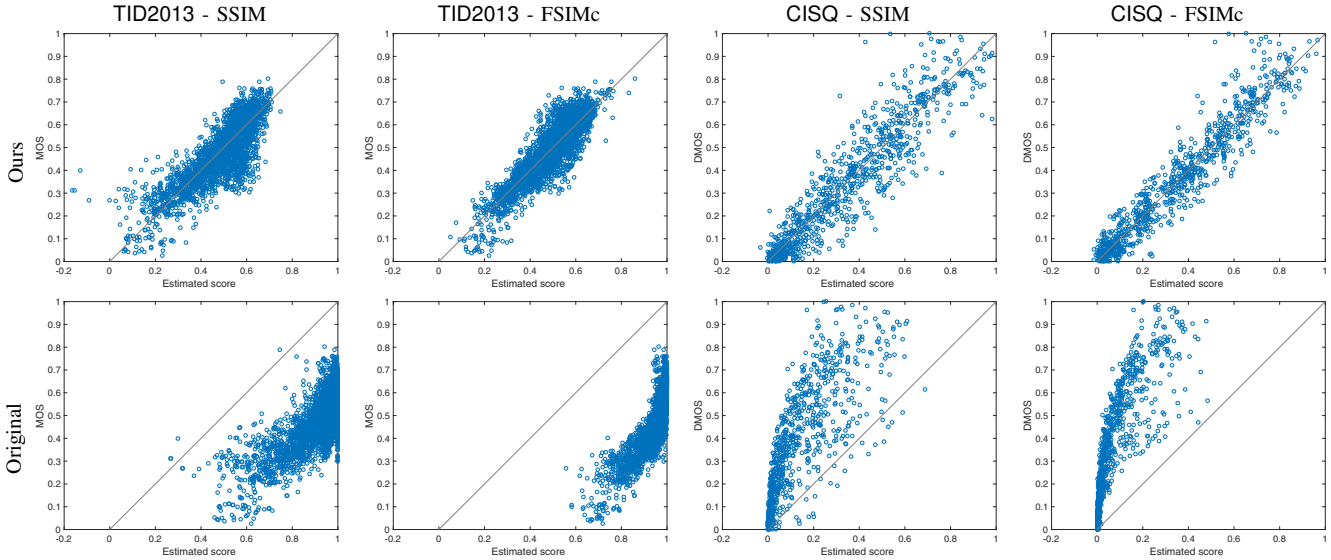


Fig. 2. Estimated scores on TID2013 (MOS) and CISQ (DMOS). The proposed method is compared with the original similarity measure in the frameworks of SSIM and FSIMc.

REFERENCES

[1] Z. Wang and A. C. Bovik, "Mean squared error: love it or leave it? - a new look at signal fidelity measures," *IEEE Signal Processing Magazine*, vol. 26, no. 1, pp. 98–117, January 2009.

[2] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactios on Image Processing*, vol. 13, no. 4, pp. 600–612, April 2004.

[3] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multi-scale structural similarity for image quality assessment," in *IEEE Asilomar Conference on Signals, Systems and Computers*, vol. 2, 2003, pp. 1398–1402.

[4] Z. Wang and E. P. Simoncelli, "Translation insensitive image similarity in complex wavelet domain," in *IEEE International Conference on Acoustics, Speech, Signal Processing*. IEEE, 2005, pp. 573–576.

[5] Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 5, pp. 1185–1198, 2011.

[6] L. Zhang, L.Zhang, X. Mou, and D. Zhang, "Fsim : A feature similarity index for image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, 2011.

[7] D. Brunet, E. R. Vrscay, and Z. Wang, "On the mathematical properties of the structural similarity index," *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 1488–1499, 2012.

[8] A. Mittal, A. Moorthy, and A. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[9] P. Ye, J. Kumar, and D. Doermann, "Beyond human opinion scores: Blind image quality assessment based on synthetic scores," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 4241–4248.

[10] P. J. Bartlett, B. Schölkopf, D. Schuurmans, and A. J. Smola, *Advances in Large-Margin Classifiers*. MIT Press, 2000.

[11] V. Vapnik, *Statistical Learning Theory*. Wiley, 1998.

[12] M. C. Morrone and D. C. Burr, "Feature detection in human vision: A phase-dependent energy model," *Proceedings of the Royal Society of London, B*, vol. 235, no. 1280, pp. 221–245, 1988.

[13] B. Jähne, H. Haußecker, and P. Geißler, *Handbook of Computer Vision and Applications*. Academic Press, 1999.

[14] P. Kovesi, "Image features from phase congruency," *Videre: A Journal Of Computer Vision Research*, vol. 1, no. 3, pp. 1–26, 1999.

[15] C. Yang and S. H. Kwok, "Efficient gamut clipping for color image processing using LHS and YIQ," *Optical Engineering*, vol. 42, no. 3, pp. 701–711, 2003.

[16] N. Ponomarenko, O. Ieremeiev, V. Lukin, K. Egiazarian, L. Jin, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, and C.-C. J. Kuo, "A new color image database TID2013: Innovations and results," in *Advanced Concepts for Intelligent Vision Systems*, 2013, pp. 402–413.

[17] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti, "TID2008 - a database for evaluation of full-reference visual quality assessment metrics," *Advances of Modern Radioelectronics*, vol. 10, pp. 30–45, 2009.

[18] E. C. Larson and D. M. Chandler, "Most apparent distortion: full-reference image quality assessment and the role of strategy," *Journal of Electronic Imaging*, vol. 19, no. 1, March 2010.

[19] H. Sheikh, Z.Wang, L. Cormack, and A. Bovik, "Live image quality assessment database release 2," http://live.ece.utexas.edu/research/quality.

[20] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430–444, 2006.