Journal of Intelligent & Fuzzy Systems 28 (2015) 1039–1046
DOI:10.3233/IFS-141387
IOS Press

Image quality assessment metrics combining structural similarity and image fidelity with visual attention

Engin Mendi*
Department of Computer Engineering, KTO Karatay University, Konya, Turkey

Abstract. Image quality assessment has a great importance in several image processing applications. Recently, various objective image quality metrics have been proposed in order to predict human visual perception. In this paper, novel image quality metrics, S-SSIM (saliency-based structural similarity index) and S-VIF (saliency-based visual information fidelity), are proposed based on a visual attention model extracting frequency-tuned salient region. Saliency maps are produced from the color and luminance features of the image. SSIM and VIF in pixel domain are modified by the weighting factors of the saliency maps. We validated our approach using 2 image databases as test bed: These databases contain subjective scores for each image. Our results showed that our technique is more correlated with human subjective perception.

Keywords: Image quality assessment, visual attention, saliency maps, structural similarity, visual information fidelity

1. Introduction

Image quality assessment (IQA) has a great importance in several image and video processing applications such as filter design, image compression, restoration, denoising, reconstruction, and classification. The aim of image quality assessment is predicting image quality of display output perceived by the final user. Multimedia contents are subjected to a variety of artifacts during acquisition, processing, storage and delivery, which may lead to reductions in the quality. IQA can be used to dynamically monitor and adjust the image quality, so that the output quality of the image or video presented to the user can be maximized for available resources such as network conditions and bandwidth requirements.

IQA methods fall into two categories: 1) subjective assessment by humans, and 2) objective assessment by algorithms. Subjective image quality experiments are classical statistical measurements of how humans perceive the image quality. Subjective measures are determined by a Mean Opinion Score (MOS) which relies on human perception.

The mathematical tools for subjective assessment of image quality are well defined, although certain practical aspects in designing efficient experiment remain to be defined. While subjective assessment is often used to judge the image quality, it is time consuming and cannot be implemented in the real time. This is the main reason behind the development of new algorithms that predict subjective image quality measure accurately. In [1], how well an algorithm performs is determined by how well it correlates with the human perception of the image quality. Objective quality metrics are algorithms designed to characterize the quality of an image and to predict the viewer's opinion.

*Corresponding author. Engin Mendi, Department of Computer Engineering, KTOKaratay University, Konya 42040, Turkey. Tel.: +90 332 2217207; Fax: +90 332 2020044; E-mails: esmendi@ualr.edu, engin.mendi@karatay.edu.tr.

Different types of objective metrics exist, as illustrated in [2]. These are based on mathematical measurements that are practical and are applied without the need of human observers. Objective quality metrics can be classified into 3 categories: 1) Full Reference (FR), 2) Reduced Reference (RR) and 3) No Reference (NR). These metrics are based on the availability of an original non-distorted reference image to be compared with a corresponding distorted image. In a FR case, reference image information is available; in a RR case, partial information of reference image is known and no information about the reference image is available in a NR case.

In the area of image processing, more than 50 years, mean squared error (MSE) is being used as quasi –standard fidelity metrics. The MSE still continues to be widely used as a signal fidelity measure. At the same time, there are recent studies that have developed more advanced signal fidelity measures especially in applications where perceptual criteria might be relevant. It is interesting to demonstrate how the image quality is measured for different regions in an image, as they may not have the same importance. Visual importance has been explored in the context of visual saliency [3], fixation calculation [1], and moving object tracking [4]. In [5], an experiment proposes to record gaze coordinates corresponding to human eye movements and the Gaze – Attentive Fixation Finding Engine (GAFFE). In [1], GAFFE is used to find points of potential visual importance and have developed an algorithm for fixation-based and quality – based weighting. The region-of interest based image quality assessment still remains unexplored.

In this study, we develop novel image quality metrics, S-SSIM (saliency-based structural similarity index) and S-VIF (saliency-based visual information fidelity), based on frequency-tuned salient region detection. Saliency maps are produced from the color and luminance features of the image. The structural similarity index (SSIM) and the visual information fidelity (VIF) in pixel domain are modified by the weighting factors of the saliency maps. Our results show that our technique is more correlated with human subjective perception.

The rest of this paper is organized as follows: Section 2 provides a brief overview of SSIM and VIF in pixel domain. Proposed image quality metrics based on frequency-tuned salient region are presented in Section 3. The results of our approach are presented in Section 4. Finally, in Section 5 the conclusions of this paper are summarized.

2. Background

2.1. SSIM

Considering two images $x = \{x_i | i = 1, 2, \ldots, N\}$ and $y = \{y_i | i = 1, 2, \ldots, N\}$ where N is the number of pixels and x_i and y_i are the ith pixels of the images of x and y, respectively, SSIM(x, y) combines three comparison components, namely luminance-$l(x,\ y)$, contrast-$c(x,\ y)$ and structure-$s(x,\ y)$ [6]:

$$\mathrm{SSIM}(x,\ y) = f(l(x,\ y),\ c(x,\ y),\ s(x,\ y)) \tag{1}$$

Luminance, contrast and structure comparisons are defined as follows:

$$\begin{aligned} l(x, y) &= \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}, \quad C_1 = (K_1 L)^2 \\ c(x, y) &= \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}, \quad C_2 = (K_2 L)^2 \\ s(x, y) &= \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}, \quad C_3 = \frac{C_2}{2} \end{aligned} \tag{2}$$

where μ_x, μ_y, σ_x, σ_y and σ_{xy} are means of x and y, variances of x and y and correlation coefficient between x and y. K_1 and K_2 are scalar constants that $K_1, K_2 << 1$ and L is the dynamic range of the pixel values. The overall SSIM is obtained by the product of luminance, contrast and structure components:

$$\begin{aligned} \mathrm{SSIM}(x, y) = \left[\frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}\right] \cdot \left[\frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}\right] \\ \cdot \left[\frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}\right] \end{aligned} \tag{3}$$

Finally, SSIM index yields to:

$$\mathrm{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{4}$$

The fundamental principle of the structural approach is that the human visual system is highly adapted to extract structural information from the visual scene and therefore measurement of structural similarity (or distortion) could provide good approximation of subjective perceptual image quality. The main drawback of SSIM algorithm in spatial domain is that it is highly sensitive to translation, rotation and scaling of images [19].

2.2. VIF in pixel domain

VIF index relates image fidelity to the mutual information between the test and the reference images using source and distortion models and as well as human visual system model. It is given as [7]:

$$\mathrm{VIF} = \frac{\sum_{j=1}^{S}\sum_{i=1}^{M_j} I(C_{i,j}; F_{i,j})}{\sum_{j=1}^{S}\sum_{i=1}^{M_j} I(C_{i,j}; E_{i,j})} \tag{5}$$

where $I(C_{i,j};\ E_{i,j})$ and $I(C_{i,j};\ F_{i,j})$ represent the information perceived by the human observer from a particular sub band in the reference and the test images respectively. C is a block vector from a given location in the reference image, E is the perception of block C by a human observer from reference image, which can be represented as $E = C + N$, where N is additive noise. F is the perception of block C by a human observer from test image, which can be represented as $F = D + N$. D is the block vector from the test image given as $D = GC + V$ where G and V are the blur and noise distortions, respectively. S denotes the number of all sub-bands and M_j is the number of blocks at j th sub-band.

3. Image quality assessment with visual attention

Image quality assessment (IQA) has a great importance in several image and video processing applications such as filter design, image compression, restoration, denoising, reconstruction, and classification. The aim of image quality

In recent years, it has become clear that many problems in perception organization are difficult to solve without introducing the contextual information of a visual scene. Subjects often search for the component feature of a target rather than searching for the target itself. Even if the target is a simple geometric form most computational models of attention ignore contextual information provided by the correlation between objects and the scene. Schyns and Oliva [8] showed that a coarse representation of the scene initiates semantic recognition before the identification of objects is processed. Many studies support the idea that scene semantics can be available early in the chain of information processing and suggest that scene recognition may not require object recognition as a first step [9, 20]. Human vision can recognize the scene even using low-spacial frequency image.

Another reason for features–driven attention is that this reflects the attempt of the eye to maximize the information it can gather at each fixation [10]. The purpose of early visual processing is to transform the highly redundant sensory input into more efficient factorial code. At the same time the human visual system has evolved multiple mechanisms for controlling gaze. Tracking can be formulated in a probabilistic framework in both the feature- and intensity-driven settings. The principal component analysis (PCA) and the independent component analysis (ICA) are two common techniques that allow for probabilistic treatment. The PCA assumes the data distribution has a Gaussian structure and model data with an appropriate orthogonal basis functions. The ICA generalizes PCA by permitting non-Gaussian distributions and non-orthogonal bases. However, these techniques do not allow noise to be modeled separately from the signal structure, and they do not permit overcomplete codes in which there are more basis functions than input dimensions. Bell and Sejnowski [11] applied their Infomax-based ICA algorithm to image coding and reported that the independent components of the natural scenes resemble edge filters. Such Gabor-like filters are believed to be a good model of the spatiotemporal receptive fields of simples' cells in primary visual cortex (V1). In [12], Olshausen and Field argued for maximizing the sparseness of the distribution of output activities, or "minimum entropy" coding as a good feature detector. In this study, we propose to model conjunction search (a search for a unique combination of two features – e.g, orientation and spatial frequency – among distractions that share only one of these features), which examines how the system combines features into perceptual wholes. We propose to improve the effectiveness of the decomposition algorithm by providing the algorithm with classification awareness. Attentional guidance does not depend solely on local visual features, but must also include the effects of interactions among features. The idea is to group filters (basis components) which become responsible for extracting similar features. A certain feature will be shared by the nearest neighbors of fixations [21].

In this study, we propose a visual attention model based on the extended frequency-tuned saliency model [13] and incorporating conjunction search [10]. A flowchart of the image quality assessment metrics is depicted in Fig. 1. The proposed model finds low-level bottom–up saliency. It is inspired by the biological

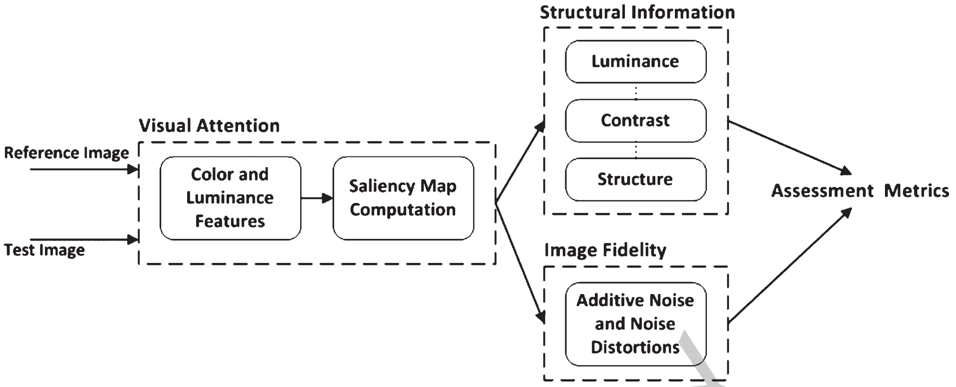


Fig. 1. The flowchart of the proposed image quality assessment metrics.

concept of center-surround contracts sensitivity of human visual system.

The proposed approach offers three advantages over existing methods: uniformly highlighted salient regions with well defined boundaries, full resolution and computational efficiency.

The proposed model finds low-level bottom–up saliency. It is inspired by the biological concept of center-surround contracts sensitivity of human visual system. The proposed approach offers three advantages over existing methods: uniformly highlighted salient regions with well defined boundaries, full resolution and computational efficiency.

Saliency maps are produced from the color and luminance features of the image. Saliency map S is formulated for the image I as follows:

$$S(x, y) = \left\| I_\mu - I_w(x, y) \right\| \qquad (6)$$

where I_μ is the mean pixel value of the image, $I_w(x,\ y)$ is the corresponding pixel vector value of the Gaussian blurred image from the original image and $\|.\|$ is the Euclidean distance. Each pixel location is the Lab color space vector, i.e. $[L,\ a,\ b]^T$.

Blurred image is a Gaussian blurred version (using 5 × 5 separable binominal kernel) of the original image. The method finds the Euclidean distance between the Lab pixel vector in a Gaussian filtered image with the average Lab vector for the input image.

3.1. S-SSIM and S-VIF in pixel domain

VIF index relates image fidelity to the mutual information between One of the common shortcomings of existing image quality metrics is the fact that they analyze the entire image uniformly. In human visual system, the importance of a visual event should increase with the information content, and decrease with the perceptual uncertainty [14]: we incorporated saliency map as weighting function into the SSIM and VIF indexes, so saliency factors can be instated into the quality metrics. The weighting function is:

$$w(x, y) = \left\| I_\mu - I_{w_{hc}}(x, y) \right\| \qquad (7)$$

We define saliency-based SSIM as S-SSIM and saliency-based VIF as S-VIF as follows:

$$\text{S-SSIM} = \frac{\sum_x \sum_y w(x, y)\text{SSIM}(x, y)}{\sum_x \sum_y w(x, y)}$$
$$\text{S-VIF} = \frac{\sum_S \sum_M w(C, F)\text{VIF}(C, F)}{\sum_S \sum_M w(C, F)} \qquad (8)$$

SSIM and VIF in pixel domain mainly focus on local information and do not take global saliency features into consideration [15]. Figure 2 shows an example case that SSIM and VIF in pixel domain fail. Figure 2(a) and (b) show a reference image and its frequency tuned saliency map, respectively. In Fig. 2(c) and (e), the images are distorted at visually attended and less-attended locations by higher amount of Gaussian noise and blurring effect, respectively. Less amount of same distortions are applied to the images at only less-attended locations in Fig. 2(d) and (f). It is easy to see that the quality of images in Fig. 2(d) and (f) are better than of Fig. 2(c) and (e).

However, as shown in Table 1, SSIM and VIF in pixel domain give incorrect results; S-SSIM and S-VIF in pixel domain scores are more realistic.

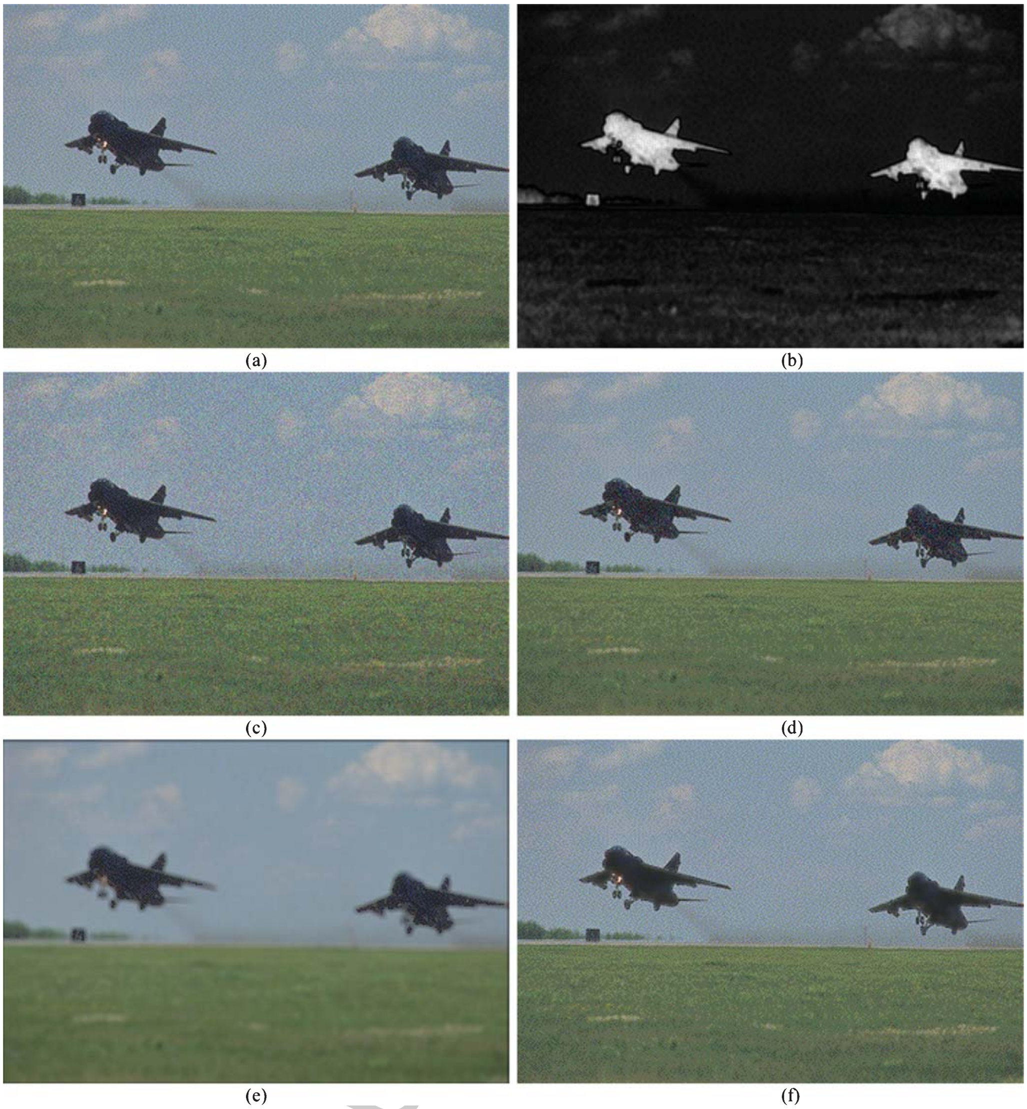


Fig. 2. a) reference image, b) saliency map of the reference image), c) Distorted image with higher amount of Gaussian noise applied to attended and less-attended locations, d) Distorted image with less amount of Gaussian noise applied to only less-attended locations, e) Distorted image with higher amount of blurring effect applied to attended and less-attended locations, f) Distorted image with less amount of blurring effect applied to only less-attended locations.

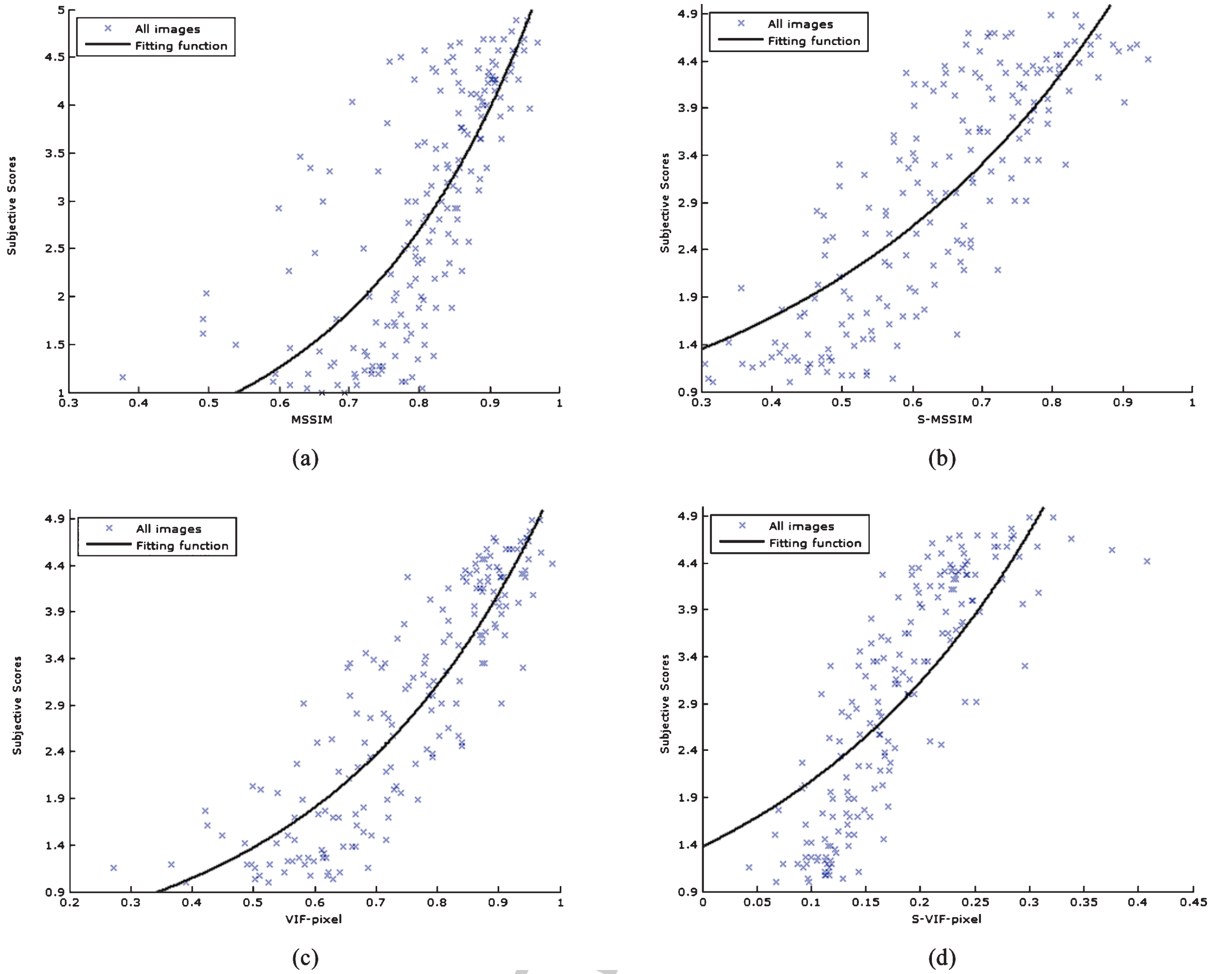
Table 1
Scores of SSIM, S-SSIM, VIF and S-VIF in pixel domains for images in Fig. 2

	SSIM	S-SSIM	VIF in pixel	S-VIF in pixel
Fig. 2(c)	0.9846	0.9698	0.9842	0.8766
Fig. 2(d)	0.9739	0.9705	0.9556	0.8965
Fig. 2(e)	0.9830	0.9656	0.9287	0.8483
Fig. 2(f)	0.9690	0.9672	0.8255	0.8525

4. Experimental results

We validated our approach using 2 image databases as test bed. These databases contain subjective scores for each image. First is the IVC Image database [16] consisting of 10 reference images with 235 distorted images (JPEG, JPEG2000, LAR coded and blurred). Second is the LIVE Image Database [17] consisting of 29 original images and 460 distorted images (227 JPEG2000 images and 233 JPEG images.). Non-linear regression analysis has been performed to fit the data. To measure the association between subjective and objective scores Pearson correlation coefficient is used.

Figures 3 and 4 show the results for IVC and LIVE databases, respectively. Each sample point denotes the subjective/objective scores of one test image. The y axis



(a)

(b)

(c)

(d)

Fig. 3. Scatter plots of subjective/objective scores on IVC Database. (a) SSIM; (b) S-SSIM, c) VIF in pixel domain, d) S-VIF in pixel domain.

Table 2
Pearson correlation coefficients

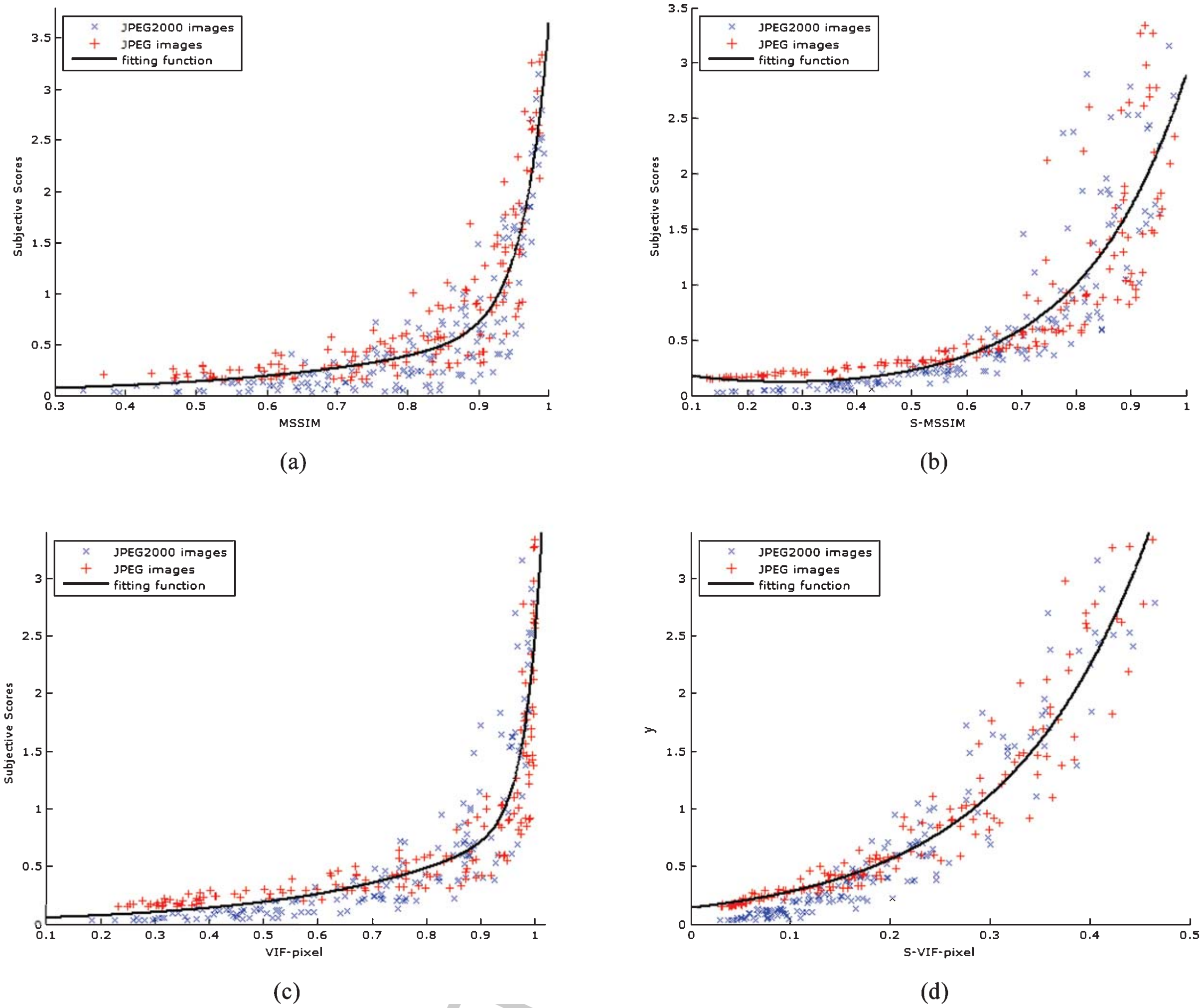
	IVC - all images	LIVE - JPEG&JPEG2000 images
SSIM	0.7047	0.6823
S-SSIM	0.8261	0.7475
VIF-pixel	0.8435	0.7126
S-VIF-pixel	0.8715	0.9083

in the figures represents the subjective scores in the databases. The x axis represents the predicted quality of images after a nonlinear regression toward 4 objective scores, which are SSIM, S-SSIM, VIF and S-VIF in pixel domains, respectively. The Pearson validation scores between assessment metrics are given in Table 2.

The Pearson correlation coefficient varies from −1 to 1 and widely used to measure the relation between two variables. High absolute values indicate that the two variables being evaluated have high correlation. As shown in Table 2, our technique is more correlated with human subjective perception.

5. Conclusions

This paper presents two novel image quality metrics, S-SSIM and S-VIF in pixel domain. The metrics are based on frequency-tuned salient region detection and computationally inexpensive. Salient region detection captures full resolution saliency maps exploiting the color and luminance features of the images. Saliency maps are then set as weighting functions and incorporated into SSIM and VIF in pixel domain. The approach has been validated using two image databases: 1) IVC Image database consisting of 10 reference images with



(a)

(b)

(c)

(d)

Fig. 4. Scatter plots of subjective/objective scores on LIVE Database. Red points (+) and blue points (x) denote JPEG and JPEG2000 images, respectively, (a) SSIM; (b) S-SSIM, c) VIF in pixel domain, d) S-VIF in pixel domain [18].

235 distorted images (JPEG, JPEG2000, LAR coded and blurred) and LIVE Image Database consisting of 29 original images and 460 distorted images (227 JPEG2000 images and 233 JPEG images.). Experiments show that the proposed metrics match with Human Visual System better than SSIM and VIF in pixel domain.

References

[1] A. Moorthy and A. Bovik, Visual importance pooling for image quality assessment, *IEEE Journal of Selected Topics in Signal Processing* **3**(2) (2009), 193–201.

[2] H.R. Wu and K.R. Rao, Digital Video Image Quality and Perceptual Coding, CRC Press, 2006.

[3] L. Itti, C. Koch and E. Niebur, A model of saliency based visual attention fro rapid scene analysis, *IEEE Trans Pattern Anal Mach Intell* **20**(11) (1998), 1254–1259.

[4] E. Mendi, M. Milanova, Y. Zhou and J. Talburt, Objective Video Quality Assessment for Tracking Moving Objects from Video Sequences, *9th WSEAS International Conference on Signal Processing, Robotics and Automation (ISPRA '10)*, Cambridge, UK, 2010, pp. 121–126.

[5] U. Rajashekar, I. Linde, A. Bovik and L.K. Cormack, GAFFE: A gaze attentive fixation finding engine, *IEEE Trans Image, Processing* **17**(4) (2008), 564–573.

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

[7] H.R. Sheikh and A. Bovik, Image information and visual quality, *IEEE Transactions on Image Processing* **15**(2) (2006), 430–444.

[8] P.G. Schyns and A. Oliva, From blobs to boundary edges: Evidence for time and spatial scale dependent scene recognition, *Psychol Sci* **5** (1994), 195–200.

[9] A. Oliva and A. Torralba, Modeling the shape of the scene: A holistic representation of the spatial envelope, *International Journal of Computer Vision* **42** (2001), 145–175.

[10] M. Milanova, S. Rubin, R. Kountchev, V. Todorov and R. Kountcheva, Combined visual attention model for video sequences, *IEEE ICPR* (2008), 1–4.

[11] A. Bell and T. Sejnowski, The 'independent components' of natural scenes are edge filters, *Visual Research* **37**(3) (1997), 3327–3338.

[12] B. Olshausen, "Sparse Codes and Spikes" in Probabilistic Models of Perception and Brain Function, R.P.N.B. Rao, A. Olshausen, M. Lewicki, Eds, MIT Press, 2001.

[13] R. Achanta, S. Hemami, F. Estrada and S. Süsstrunk, Frequency-tuned salient region detection, *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[14] Q. Li and Z. Wang, Video quality assessment by incorporating a motion perception model, *in Image Processing 2007, ICIP 2007, IEEE International Conference on*, **2** (2007), 173–176.

[15] Q. Ma and L. Zhang, Image quality assessment with visual attention, *19th International Conference on Pattern Recognition (ICPR)*, Tampa, FL, USA, 2008, pp. 1–4.

[16] P.L. Callet and F. Autrusseau, Subjective quality assessment IRCCyN/IVC database, http://www.irccyn.ec-nantes.fr/ivcdb/

[17] H.R. Sheikh, Z. Wang, L. Cormack and A. Bovik, LIVE Image Quality Assessment Database Release 2, http://live.ece.utexas.edu/research/quality

[18] E. Mendi, S. Cecen, E. Ermisoglu and C. Bayrak, Automated neurosurgical video segmentation and retrieval system, *Journal of Biomedical Science and Engineering (JBiSE)* **3**(6) (2010), 618–624.

[19] S. Kaya, T. Bennett, M. Milanova, J. Talburt, B. Tsou, M. Altynova and H. Xu, "Perception-based Image/Video Quality Metric Using CIELAB Color Space", Proc. SPIE 8019, Sensors, and Command, Control, Communications, and Intelligence (C3I) *Technologies for Homeland Security and Homeland Defense X*, 801908, June 2, 2011.

[20] M. Milanova, R. Kountchev, S. Rubin, V. Todorov and R. Kountcheva, "Content Based Image Retrieval Using Adaptive Inverse Pyramid Representation", In: G. Salvendy, M. J. Smith, (eds.), *HCI International 2009*. LNCS, vol. 5618, pp. 304–314, Springer, Heidelberg, 2009.

[21] M. Milanova, S. Rubin, R. Kountchev, V. Todorov and R. Kountcheva, "Combined Visual Attention Model for Video Sequences," *IEEE International Conference on Pattern Recognition (ICPR)*, Tampa, 2008, pp. 1-4.