IEICE TRANS. INF. & SYST., VOL.E102–D, NO.7 JULY 2019

LETTER

Quality Index for Benchmarking Image Inpainting Algorithms with Guided Regional Statistics

Song LIANG[†], ***Member***, **Leida LI**[†], **Bo HU**[†], ***and*** **Jianying ZHANG**[†a)], ***Nonmembers***

SUMMARY This letter presents an objective quality index for benchmarking image inpainting algorithms. Under the guidance of the masks of damaged areas, the boundary region and the inpainting region are first located. Then, the statistical features are extracted from the boundary and inpainting regions respectively. For the boundary region, we utilize Weibull distribution to fit the gradient magnitude histograms of the exterior and interior regions around the boundary, and the Kullback-Leibler Divergence (KLD) is calculated to measure the boundary distortions caused by imperfect inpainting. Meanwhile, the quality of the inpainting region is measured by comparing the naturalness factors between the inpainted image and the reference image. Experimental results demonstrate that the proposed metric outperforms the relevant state-of-the-art quality metrics.

key words: *quality evaluation, image inpainting, GRS, gradient magnitude, naturalness*

1. Introduction

Image inpainting is to automatically restore a damaged image region without leaving visible artifacts. A great number of image inpainting algorithms have been proposed in the literature. However, how to evaluate the performances of inpainting algorithms remains an open problem, and objective quality metrics are highly desired for this purpose. Traditional image quality assessment (IQA) metrics are usually designed for distortions that are evenly distributed in the whole image. However, in image inpainting, the degraded regions are typically localized, which can be determined by user-defined masks. Therefore, traditional IQA metrics are not suitable for benchmarking image inpainting algorithms.

Recently, several works have been done towards the objective quality evaluation of image inpainting. Inspired by the structural similarity (SSIM) [1] metric, Wang et al. [2] proposed the parameter weight for image inpainting quality (PWIIQ) index by combining luminance, definition and gradient similarities. However, PWIIQ is not effective for images with large inpainting areas. Another group of image inpainting quality metrics are based on visual saliency, such as the average squared visual salience (ASVS) [3], degree of noticeability (DN) [3], gaze density inside the hole region (GDin) [4], gaze density outside the hole region (GDout) [4], border saliency (BorSal) [5] and StructBorSal [5]. Moreover, Dang et al. [6] used the visual coherence of the recovered regions and the visual saliency to develop their metric. These metrics are based on the fact that visual saliency indicates the observable artifacts and the change of salient regions due to inpainting is consistent with image quality. Nevertheless, they do not comprehensively consider the overall visual appearance of an image, which may be problematic for relatively high quality images and could only handle a limited number of possible inpainting artifacts. Furthermore, Mariko et al. [7] developed a ranking-by-learning index to estimate the ordering of inpainted images, which is actually a variant of the visually saliency algorithm.

This letter presents a new objective metric for image inpainting quality assessment (IIQA) based on guided regional statistics (GRS), which evaluates the overall quality of image inpainting by simultaneously considering the impacts of both boundary and inpainting regions. The distributions of gradient magnitude (GM) histograms between the exterior and interior boundary region are used to measure the distortion of the inpainting boundary. Moreover, the loss of naturalness in the inpainting region is measured. Finally, an overall quality score is produced by combining the above two aspects. The performance of the proposed metric is evaluated by experiments.

The contributions of this letter are as follows: (1) We re-labeled the public inpainting database TUM-IID [8], increasing its usability and providing a new experimental platform for IIQA. (2) The statistical features, namely, GM statistics and naturalness, are innovatively introduced to evaluate the quality of inpainted images. (3) A new objective IIQA metric is proposed for benchmarking image inpainting algorithms, which outperforms the relevant state-of-the-art traditional IQA and IIQA.

2. Proposed IIQA Index

Figure 1 shows the flowchart of the proposed IIQA metric. With the inpainting mask, an image is divided into the inpainting region Ω and the non-inpainting region Φ. Quality evaluation of the inpainted image is achieved from two aspects, namely bounadry distortion evaluation and inpainting region naturalness evaluation.

2.1 Boundary Feature

An ideal inpainting algorithm is expected to restore the mask region naturally so that the exterior and interior boundary

Manuscript received September 28, 2018.
Manuscript revised February 24, 2019.
Manuscript publicized April 1, 2019.

[†]The authors are with School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221116, China.
a) E-mail: zjycumt@126.com (Corresponding author)

DOI: 10.1587/transinf.2018EDL8206

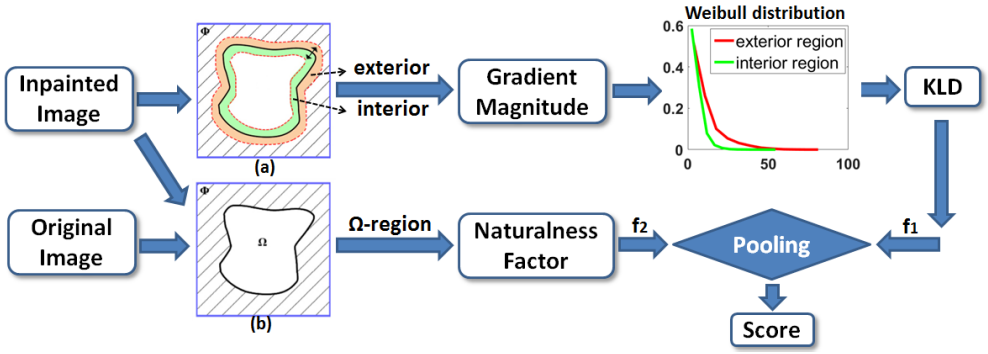

Fig. 1 Flowchart of the proposed IIQA metric.

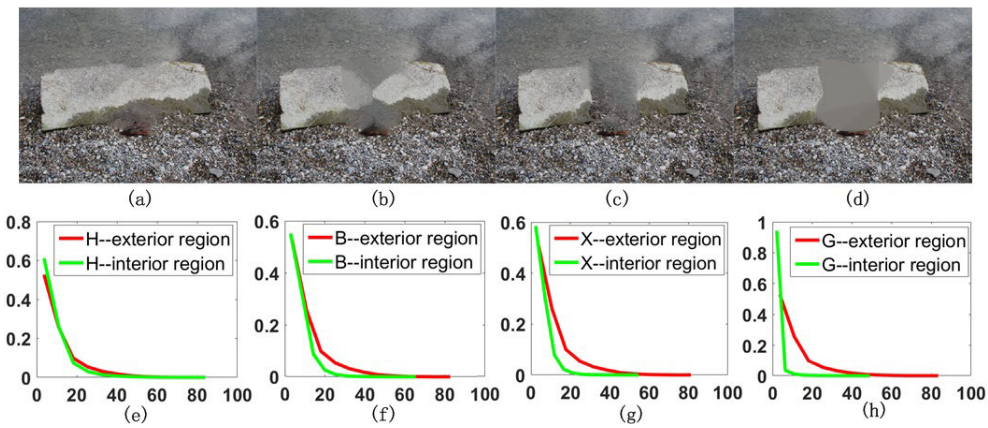

Fig. 2 Illustration of boundary features. Images (a) to (d) are the images inpainted by different algorithms in [11]. Plots (e) to (h) are the corresponding probability distributions of the GM maps in the exterior and interior regions. The horizontal axis shows the range of pixel values in the GM map and the vertical axis is the probability value.

regions look similar. Therefore, the statistical difference between the exterior and interior sides of the mask boundary is closely related to the inpainting quality. The GM feature measures the strength of local luminance change and builds the basic elements (i.e., local contrast) of image semantic structures, which is hence closely related to the perceptual quality of inpainted images. Motivated by this, we measure the quality of the boundary region based on the GM statistics.

For a given inpainted image, the contour of the inpainting mask is first determined, from which the same distances are respectively expanded to both the outer and inner directions. Two pixel strips of the same width are thus formed, namely, the exterior and interior regions, which are respectively marked in salmon pink and palegreen as illustrated in Fig. 1. In this letter, we empirically set the specific width to 16 pixels by experiments. Then, the GM map of the inpainted image **I** can be computed as [9]:

$$G_I = \sqrt{[I \otimes h_x]^2 + [I \otimes h_y]^2}, \tag{1}$$

where $\otimes$ denotes the convolution operation and h_d, $d \in \{x, y\}$, denotes the Gaussian partial derivative in horizontal or vertical direction,

$$\begin{aligned} h_d(x, y|\sigma) &= \frac{\partial}{\partial d} g(x, y|\sigma) \\ &= -\frac{1}{2\pi\sigma^2}\frac{d}{\sigma^2} exp\Big(-\frac{x^2+y^2}{2\sigma^2}\Big), \end{aligned} \tag{2}$$

where $g(x, y|\sigma) = \frac{1}{2\pi\sigma^2} exp(-\frac{x^2+y^2}{2\sigma^2})$ denotes the two-dimensional isotropic Gaussian function and σ is the scale parameter. Then, the Weibull distribution is employed to fit the histogram probability density of the GM map for the exterior and interior boundary regions, respectively. Finally, the Kullback-Leibler Divergence (KLD) [10] is calculated to measure the distance of the two GM distributions, which is defined as follows:

$$KLD(P_{int}, P_{ext}) = \sum_{i=1}^{n} P_{int}(i) ln \frac{P_{in}(i)}{P_{ext}(i)}, \tag{3}$$

where P_{ext} and P_{int} denote the probability distributions of exterior and interior boundary regions, n denotes the variable values in each probability vector. Considering the asymmetry property of KLD and to avoid negative value, the following formula is calculated to generate the final feature:

$$f_1 = \frac{1}{2}[KLD(P_{int}, P_{ext}) + KLD(P_{ext}, P_{int})]. \tag{4}$$

In Fig. 2, an example is given for illustration purpose.

We can easily observe that the shapes of the fitted curves depend on the inpainting algorithms, and different algorithms produce different curves. The order of (a), (b), (c), (d) is ranked in sequence from best to worst according to the quality of inpainted images, which is highly consistent with the deviation degree of the corresponding fitting curves.

2.2 Inpainting Region Naturalness

In addition to the boundary feature, we also evaluate the inpainting quality by measuring the loss of naturalness between the original image and the inpainted image in the mask region. Here, we employ the naturalness factor (NF) [11], whcih is defined as:

$$NF = (1-\theta)\frac{T_1}{T_1^{pr}} + \theta\frac{T_2}{T_2^{pr}}, \tag{5}$$

where T_i and $T_i^{pr}, i \in \{1,2\}$, denote the empirical parameter of the given image and the prior parameter learned from the natural-scene dataset respectively, and $\theta \in [0,1]$ is the weight. For a given image, T_1 and T_2 are obtained by using two parametric distribution models in [11] to severally approximate the cumulative distribution functions (CDF) of the gradient and the Laplace CDF. The corresponding prior values are set as: $T_1^{pr} = 0.38, T_2^{pr} = 0.14$. More details on the calculation of the naturalness factor can be found in [10]. Finally, the naturalness loss of the inpainted regions between the reference image and the inpainted image is computed as:

$$f_2 = |NF_\Omega^{Original} - NF_\Omega^{Inpainted}|. \tag{6}$$

2.3 Pooling

With the boundary feature f_1 around the inpainting region boundary and the naturalness feature f_2 of the inpainting region, an overall quality score Q is generated by:

$$Q = \alpha \cdot f_1 + (1-\alpha)\cdot f_2, \tag{7}$$

where $\alpha \in [0,1]$ is the weight used to balance the relative importance of the boundary feature and the naturalness feature in the quality assessment of inpainted images. In this letter, we set $\alpha = 0.6$ by experiments. This is because in contrast to the relatively uniform artifacts in the inpainting region, the influence of boundary distortions on the human visual system (HVS) is more dominated, especially for large inpainted regions.

3. Experimental Results and Analysis

3.1 Experiment Settings

The experiments are conducted on the TUM-IID database [8], which contains 17 reference images with diverse visual contents, 4 masks with different size and shape, and the corresponding 272 inpainted images based on four state-of-the-art image inpainting methods. Since the original database is only partially labelled, we further conduct a user study to collect the ground truth quality rankings of images generated by different inpainting algorithms. To be specific, 31 observers were invited to rank the quality of a group of four inpainted versions of each original image in the whole database using the interface shown in Fig. 3. The quality of each group of four inpainted images is ranked from "1" to "4", where "1" represents the worst quality and "4" is the best. Further, the original reference image is also displayed to facilitate the rating process.

For performance comparison, Spearman Rank Order Correlation Coefficient (SROCC) and Kendalls Rank Order Correlation Coefficient (KROCC) are adopted to measure

Fig. 3 Interface for user study.

Table 1 Performance comparison of state-of-the-art IQA metrics and our proposed metric on TUM-IID database. The best-performing metric is highlighted in bold for each category. TIQA: trational IQA; FR: full reference; NR: no reference.

Category	Metrics	Type	SROCC	KROCC
TIQA	SSIM [1]	FR	−0.2256	−0.1845
	FSIM [13]	FR	0.5664	0.4830
	MSSSIM [14]	FR	−0.0178	−0.0076
	PSNR [15]	FR	−0.3560	−0.3213
	VSI [16]	FR	**0.6576**	**0.4830**
IIQA	PWIIQ [2]	FR	0.6101	0.5905
	ASVS [3]	NR	−0.4528	−0.3944
	DN [3]	FR	−0.4774	−0.4135
	GDin [4]	FR	0.5450	0.4674
	GDout [4]	FR	0.0485	0.0365
	BorSal [5]	FR	0.4980	0.4282
	StructBorSal [5]	FR	0.6017	0.5174
	GRS_f_1	FR	0.7904	0.7193
	GRS_f_2	FR	0.7963	0.7340
	GRS	FR	**0.8434**	**0.7781**

the prediction monotonicity between the subjective rankings and the objective scores generated by IQA metrics [12]. Note that the calculation of both SROCC and KROCC is based on a group of images, which have the same content but processed by different inpainting algorithms. The TUM-IID database consists of 68 groups, so the mean values are reported here. The final value ranges from −1 to 1, where "1" indicates that the predicted scores are totally consistent with the subjective ratings, and "−1" is opposite.

3.2 Performance Evaluation

In this part, we compare the performance of the proposed method with the relevant state-of-the-art quality assessment metrics, including both traditional IQA and IIQA. The experimental results are listed in Table 1.

It is observed from Table 1 that the proposed GRS metric achieves the best performance in terms of both SROCC and KROCC, and it also outperforms the general IQA metrics and other IIQA metrics. Specifically, from the results, the quality scores of several metrics show weak correlations (negative correlation values) with the subjective rankings, especially for ASVS [3] and DN [3]. In addition, the performance of the proposed two component features (GRS_f_1 and GRS_f_2) is also superior to other metrics. At the quality score pooling stage, the influence of different regions on the quality perception of HVS is taken into account, so combining the two features is more effective.

4. Conclusion

In this letter, a new IIQA metric has been proposed by comprehensively considering the changes of statistical information around and inside the inpainted regions. Boundary features and inpainting region naturalness features are utilized to measure the boundary and regional statistics under the guidance of the inpainting mask. Experiments have been conducted based on the TUM-IID database and the results confirm the superiority of the proposed metric in contrast to the state-of-the-art quality metrics.

Acknowledgements

This work was supported by Natural Science Foundation of Jiangsu Province (BK20181354), National Natural Science Foundation of China (61771473 and 61379143), the Six Talent Peaks High-level Talents in Jiangsu Province (XYDXX-063) and the Qing Lan Project.

References

[1] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," IEEE Trans. Image Process., vol.13, no.4, pp.600–612, 2004.
[2] S. Wang, H. Li, X. Zhu, and P. Li, "An evaluation index based on parameter weight for image inpainting quality," The 9th International Conference for Young Computer Scientists, pp.786–790, 2008.
[3] P.A. Ardis and A. Singhal, "Visual salience metrics for image inpainting," Visual Communications and Image Processing, vol.7257, pp.72571W-72571W-9, 2009.
[4] M.V. Venkatesh and S.-C.S. Cheung, "Eye tracking based perceptual image inpainting quality analysis," IEEE International Conference on Image Processing, vol.119, no.5, pp.1109–1112, 2010.
[5] A.I. Oncu, F. Deger, and J.Y. Hardeberg, "Evaluation of digital inpainting quality in the context of artwork restoration," International Conference on Computer Vision, vol.7583, pp.561–570, 2012.
[6] A.D.T. Trung, B.A. Beghdadi, and C.C. Larabi, "Perceptual quality assessment for color image inpainting," IEEE International Conference on Image Processing, pp.398–402, 2013.
[7] M. Isogawa, D. Mikami, K. Takahashi, and H. Kimata, "Image quality assessment for inpainted images via learning to rank," Multimedia Tools and Applications, vol.78, no.2, pp.1399–1418, 2019.
[8] P. Tiefenbacher, V. Bogischef, D. Merget, and G. Rigoll, "Subjective and objective evaluation of image inpainting quality," IEEE International Conference on Image Processing, pp.447–451, 2015.
[9] W. Xue, X. Mou, L. Zhang, A.C. Bovik, and X. Feng, "Blind image quality assessment using joint statistics of gradient magnitude and laplacian features," IEEE Trans. Image Process., vol.23, no.11, pp.4850–4862, 2014.
[10] S. Kullback, "The Kullback-Leibler distance," American Statistician, vol.41, no.4, pp.340–341, 1987.
[11] Y. Gong and I.F. Sbalzarini, "Image enhancement by gradient distribution specification," Asian Conference on Computer Vision, vol.9009, pp.47–62, 2015.
[12] L. Li, Y. Zhou, W. Lin, J. Wu, X. Zhang, and B. Chen, "No-reference quality assessment of deblocked images," Neurocomputing, vol.177, pp.572–584, 2016.
[13] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: a feature similarity index for image quality assessment," IEEE Trans. Image Process., vol.20, no.8, pp.2378–2386, 2011.
[14] Z. Wang, E.P. Simoncelli, and A.C. Bovik, "Multiscale structural similarity for image quality assessment," IEEE Conference on Signals, Systems and Computers, vol.2, no.2, pp.1398–1402, 2004.
[15] Y. Fisher, "Fractal Image Compression," Fractals-complex Geometry Patterns and Scaling in Nature and Society, vol.2, no.03, pp.347–361, 1994.
[16] L. Zhang, Y. Shen, and H. Li, "VSI: a visual saliency-induced index for perceptual image quality assessment," IEEE Trans. on Image Process., vol.23, no.10, pp.4270–4281, 2014.