Journal of Southeast University (English Edition) Vol. 32, No. 1, pp. 29 - 34 Mar. 2016 ISSN 1003—7985

Image quality assessment based on perceptual grouping

Wang Tonghan[1] Zhang Lu[2] Jia Huizhen[3] Kong Youyong[1] Li Baosheng[1,4] Shu Huazhong[1]

([1]Laboratory of Image Science and Technology, Southeast University, Nanjing 210096, China)
([2]IETR Lab (UMR CNRS 6164), INSA de Rennes, 20 Avenue des Buttes de Coesmes, CS 70839F-35708 Rennes Cedex 7, France)
([3]School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China)
([4]Shandong Cancer Hospital, Jinan 250117, China)

Abstract: To further explore the human visual system (HVS), the perceptual grouping (PG), which has been proven to play an important role in the HVS, is adopted to design an effective image quality assessment (IQA) model. Compared with the existing fixed-window-based models, the proposed one is an adaptive window-like model that introduces the perceptual grouping strategy into the IQA model. It works as follows: first, it preprocesses the images by clustering similar pixels into a group to the greatest extent; then the structural similarity is used to compute the similarity of the superpixels between reference and distorted images; finally, it integrates all the similarity of superpixels of an image to yield a quality score. Experimental results on three databases (LIVE, IVC and MICT) show that the proposed method yields good performance in terms of correlation with human judgments of visual quality.

Key words: perceptual grouping; perceptual image quality assessment; superpixels; full reference

DOI: 10.3969/j.issn.1003-7985.2016.01.006

Image quality assessment occupies a very important position in numerous fields and applications, such as image acquisition, compression, transmission and restoration. Since human beings are the ultimate receivers of any visual stimulus, it is essential to develop a perceptual model to closely correlate with the human visual system (HVS).

Objective quality assessment methods can be classified into three types[1]: 1) Full-reference (FR), where an ideal "reference" image is available for comparison; 2) Reduced-reference (RR), where partial information about the reference image is available; 3) No-reference (NR), where the reference image is not accessible. This paper focuses on the FR methods.

In the past decades, great efforts and huge advances have been made in FR methods. Here, we briefly review some representative ones. The traditional metrics such as the peak signal-to-noise ratio (PSNR) and the mean squared error (MSE) did not correlate well with human opinions[2]. As a milestone in the development of IQA models, the structural similarity (SSIM)[3] surpassed the previous ones since it had a better correlation with the human perception. It was based on the assumption that the HVS was highly adapted for extracting structural information. Then, several SSIM-based metrics were proposed in Refs. [4-6]. Sheikh et al.[7] proposed the visual information fidelity (VIF), which took the FR IQA problem as an information fidelity problem and chose the amount of information shared by the reference image and the distorted one as the similarity. Based on the observation that the visual information in an image is often redundant and the HVS understands an image mainly based on its low-level features, Zhang et al.[8] proposed the feature-similarity (FSIM) index, which employed two features (the phase congruency and the gradient magnitude) to compute the local similarity map. Unlike the SSIM's average pooling, the FSIM adopted a weighting strategy for the pooling. In their later work, Zhang et al.[9] proposed a visual saliency-induced metric (VSI), based on the assumption that an image's visual saliency map had a close relationship with its perceptual quality. In the VSI, three components (visual saliency, gradient modulus and chrominance) were first computed by locally comparing the distorted image with the reference one via similarity function, and then the visual saliency part was used as a weighting function to measure the importance of a local image region. Note that the weighting pooling may improve the IQA accuracy against those with average pooling to some extent, but it may be costly to compute the weights. In addition, this pooling can make the predicted quality scores become more nonlinear to human opinions[10]. The image gradient is a popular feature in IQA since it can effectively capture image local structures, to which the HVS is highly sensitive. Based on these observations, Xue et al.[10] proposed the gradient magnitude similarity deviation (GMSD) index, where image gradient magnitude maps were first computed, then the standard deviations of these maps were treated as the overall im-

Received 2015-09-16.

Biographies: Wang Tonghan (1984—), male, graduate; Shu Huazhong (corresponding author), male, doctor, professor, shu.list@seu.edu.cn.

Foundation items: The National Natural Science Foundation of China (No. 81272501), the National Basic Research Program of China (973 Program) (No. 2011CB707904), Taishan Scholars Program of Shandong Province, China (No. ts20120505).

Citation: Wang Tonghan, Zhang Lu, Jia Huizhen, et al. Image quality assessment based on perceptual grouping[J]. Journal of Southeast University (English Edition), 2016, 32(1): 29 - 34. DOI: 10.3969/j.issn.1003-7985.2016.01.006.

age quality score. A comprehensive survey and comparison of state-of-the-art FR-IQA models can be found in Refs. [11 − 12].

1 Perceptual Grouping Induced IQA Metric

The sophisticated FR-IQA models are normally divided into two steps[10]: local quality computation and pooling. The two-step FR-IQA can be summarized as shown in Fig. 1.

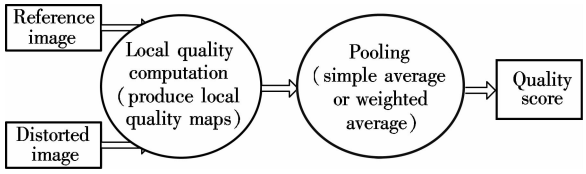

Fig. 1 Architecture of the two-step FR-IQA

During the first step, a local fixed-sized window strategy is usually adopted to filter an image. This strategy was blindly applied to the pixels in the filter window regardless of these pixels' intensity. Take the SSIM as an example, it utilized an 11 × 11 window, within which the SSIM index was calculated for the reference image and the distorted one, respectively. The window moved pixel-by-pixel to traverse the whole image to generate the local quality maps. This fixed-sized window strategy may not correlate well with the HVS. For example, the SSIM was insufficient for assessing images with blur distortion or noise[5]. The reason may be that if the distortion of one pixel in the filtering window makes its intensity very different from its neighborhood, the mean value of pixels in the window will then deviate from the counterpart in the reference image. Thus, this distortion changes the SSIM index value, but may not change the human perceived quality in some cases, e. g., the number of contaminated pixels is much fewer than that of the overall pixels in an image or these pixels are not visually salient.

To solve this problem, we adopt the perceptual grouping (PG) to make the IQA score changes correlate better with the changes in human perception. Wertheimer[13] showed that the PG strategy plays an important role in human visual perception. He also listed several key factors that bring about visual grouping, e. g. similarity, proximity, and good continuation, etc. With the PG strategy, the HVS can easily capture the variation among the changed pixels since these pixels are much more visually salient now. The superpixel algorithm is one of the PG methods that clusters the similar pixels into perceptually meaningful atomic groups. The PG methods capture image redundancy and provide a convenient description to compute image features, and they reduce the complexity of the follow-up image processing tasks[14]. A broad range of computational vision problems are closely connected with the PG. Spatially non-uniform regions of support can be identified using GP techniques[15]. Superpixel algorithms[14−20] have been used as a preprocessing step in the segmentation algorithms. Traditional superpixel algorithms like Ncuts[16] and Trubopixel[18] are too expensive to generate the superpixel of an image. In this paper, we select the simple linear iterative clustering (SLIC)[14] strategy since it yielded the best performance for now.

We propose here integrating the SLIC into the IQA metric and adapt that for grayscale images. To demonstrate the efficiency of the PG strategy, we simply use the SSIM as the IQA method. Our method is, thus, superpixel-wise, instead of pixel-wise. Experimental results show that the proposed method yields better overall performance.

1.1 Simple linear iterative clustering

The simple linear iterative clustering (SLIC) can be implemented by the following steps:

1) In the initial step, the centers are sampled on a regular interval length of $L = \sqrt{N/c}$, where c and N are the number of centers and the total number of pixels of a grayscale image, respectively.

2) For each center, it iteratively searches for the best matching pixels under the guidance of intensity similarity and spatial proximity. By doing this, it guarantees both the homogeneity and the compactness.

3) When the distance of the new clustering center and the previous one is smaller than a threshold, the iteration stops.

By default, the SLIC only needs to specify one parameter, i. e. the number of expected superpixels c. Note that the parameter m in Ref. [14] is a measure of the maximal grayscale distance in a cluster. To demonstrate the importance of the PG strategy to the HVS, we simply set $m = 0$ to not take into account the spatial distance. In other words, we only care about the intensity similarity, i. e. those pixels having the most similar intensity are considered a superpixel. The smaller a value m takes, the more tightly the resulting superpixels will adhere to image boundaries[14].

Fig. 2 shows an example of the generating superpixels for the white noise distorted version of "parrots" with its difference mean opinion score (DMOS) of 0.128 906[21]. We can see from Fig. 2 that the small value of m yields an irregular size and shape but with tight adherence to boundaries. This makes the similar pixels being clustered into a group as much as possible. Thus, the small changes in this group between the reference image and the distorted one can be accurately reflected by this measure. In addition, this is in line with the HVS since it is easier for human beings to capture the changes in the regions with intensity homogeneity.

1.2 Similarity measure

To show the effectiveness of this PG strategy, we apply it to the window-based similarity metric SSIM for

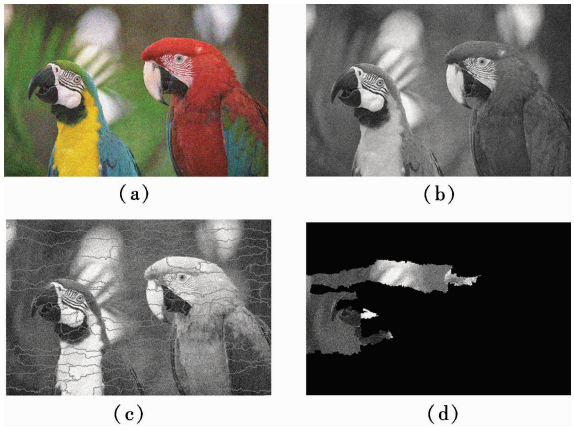

Fig. 2 Illustration of superpixels generated by SLIC. (a) Color version; (b) Grayscale version; (c) Superpixels generated by SLIC with $c=100$ and $m=0$; (d) The 4th superpixel (excluding the black background).

simplicity. As is known, the SSIM can be treated as the milestone of IQA despite its prediction performance which has been outperformed by the later developed metrics, such as weighting SSIM, multiscale SSIM, FSIM, VSI, GMSD, etc.

For each superpixel generated by the SLIC, we compute its SSIM[3]:

$$\mathrm{SSIM}(x,y)=\frac{(2\mu_x\mu_y+K_1)(2\sigma_{xy}+K_2)}{(\mu_x^2+\mu_y^2+K_1)(\sigma_x^2+\sigma_y^2+K_2)} \quad (1)$$

where μ_x and μ_y are the sample means of x and y, respectively; σ_x and σ_y are the standard variances of x and y, respectively; σ_{xy} is the sample correlation coefficient between x and y. K_1 and K_2 are included to avoid instability. In practice, their values depend on the dynamic range of $\mu_x^2+\mu_y^2$ and $\sigma_x^2+\sigma_y^2$, respectively. In our situation, x and y are the corresponding superpixels for the reference image and the distorted one, respectively. For simplicity, we obtain the overall image quality by the mean pooling:

$$\mathrm{Score}(X,Y)=\frac{1}{N}\sum_{i=1}^{N}\mathrm{SSIM}(x_i,y_i) \quad (2)$$

Throughout this paper, we set the parameters $K_1=91$ and $K_2=21$. These values, however, are somewhat arbitrary. For generating the superpixels, we set the parameters $c=800$ and $m=0$, respectively.

In summary, our metric starts with a perceptual grouping of the image (clustering the image pixels into superpixels), then computes the superpixel-wise similarity, and finally adopts the mean pooling to obtain the final score.

2 Performance Evaluation

To test the performance of the PG-based metric, we use the following three databases.

1) The LIVE IQA database[21]. It contains 29 reference images, each with five different types of distortion at 5 to 6 levels. The distortion types include JP2K, JPEG, WN, Gblur, and FF (simulated by JP2K compression followed by channel bit errors). These distortions reflect a broad range of image impairments, such as edge smoothing, block artifacts and random noise. The total number of distorted images (excluding 29 reference images) is 779.

2) MICT[22]. There are 98 images of 768×512 pixels for both JPEG and JP2K groups. Six quality scales are selected for each distortion type.

3) IVC[23]. It consists of 10 original images and 235 distorted images generated from four different processings.

We calculate three commonly used performance indices, i.e. the Spearman's rank ordered correlation coefficient (SROCC) which measures the prediction monotonicity, Pearson's (linear) correlation coefficient (LCC) which is related to the prediction linearity (considered as the measure of prediction accuracy), and the root mean square error (RMSE) which evaluates the prediction consistency. To compute the latter two indices, we use a logistic regression function to reduce the nonlinearity of predicted scores[24]. A value close to 1 for SROCC and LCC indicates good performance for quality prediction. Whereas, for RMSE, the smaller the value, the better prediction consistency it yields.

We compare the proposed method to five state-of-the-art and representative FR-IQA models, including PSNR, SSIM[3], FSIM[8], GMSD[10] and VSI[9]. Note that the source codes of all the other metrics are obtained from the original authors.

We can see from Tabs. 1 to 3 that the top three metrics

Tab. 1 SROCC for different metrics on LIVE database

Metrics	JP2K	JPEG	WN	Gblur	FF	All
PSNR	0.953 8	0.931 5	0.991 5	0.873 0	0.936 2	0.909 2
SSIM	0.971 4	0.958 2	0.978 4	0.938 5	0.965 7	0.925 0
FSIM	0.981 8	0.962 5	0.979 8	0.983 1	0.970 7	0.961 0
GMSD	0.982 3	0.960 7	0.984 7	0.975 1	0.965 8	0.954 6
VSI	0.970 0	0.953 4	0.988 1	0.970 3	0.964 4	0.946 4
PG + SSIM	0.979 7	0.958 6	0.972 2	0.974 7	0.974 3	0.967 6

Tab. 2 LCC for different metrics on LIVE database

Metrics	JP2K	JPEG	WN	Gblur	FF	All
PSNR	0.966 1	0.956 1	0.973 8	0.898 8	0.943 9	0.936 6
SSIM	0.965 5	0.968 2	0.938 2	0.922 7	0.959 7	0.938 8
FSIM	0.961 0	0.948 5	0.976 9	0.956 6	0.953 4	0.949 3
GMSD	0.970 1	0.945 7	0.986 4	0.963 8	0.965 8	0.951 1
VSI	0.957 3	0.961 5	0.971 2	0.935 6	0.933 1	0.943 1
PG + SSIM	0.967 0	0.956 4	0.982 9	0.958 0	0.969 6	0.959 7

Tab. 3 RMSE for different metrics on LIVE database

Metrics	JP2K	JPEG	WN	Gblur	FF	All
PSNR	8.078 1	7.598 2	7.534 8	9.859 8	7.328 1	8.098 3
SSIM	7.162 6	6.613 8	10.971 8	8.405 5	6.416 2	7.961 2
FSIM	7.517 3	8.205 0	6.574 1	6.739 0	6.744 9	7.266 5
GMSD	7.568 1	8.314 3	6.677 8	6.291 3	6.003 3	7.137 4
VSI	7.717 8	7.027 2	8.134 9	7.734 5	7.970 4	7.685 6
PG + SSIM	6.609 6	7.729 7	5.434 8	6.390 4	5.556 2	6.497 3

are the PG-based metric (16 times), GMSD (15 times), and FSIM (12 times) on the whole LIVE database in terms of all the three criteria (SROCC, LCC, RMSE).

In addition, we also draw the scatter plots of subjective scores against objective scores predicted by these FR-IQA metrics in Fig. 3. The greater the linear relationship against DMOS and the tighter clustering it yields, the better performance the metric yields in terms of the correlation with subjective ratings.

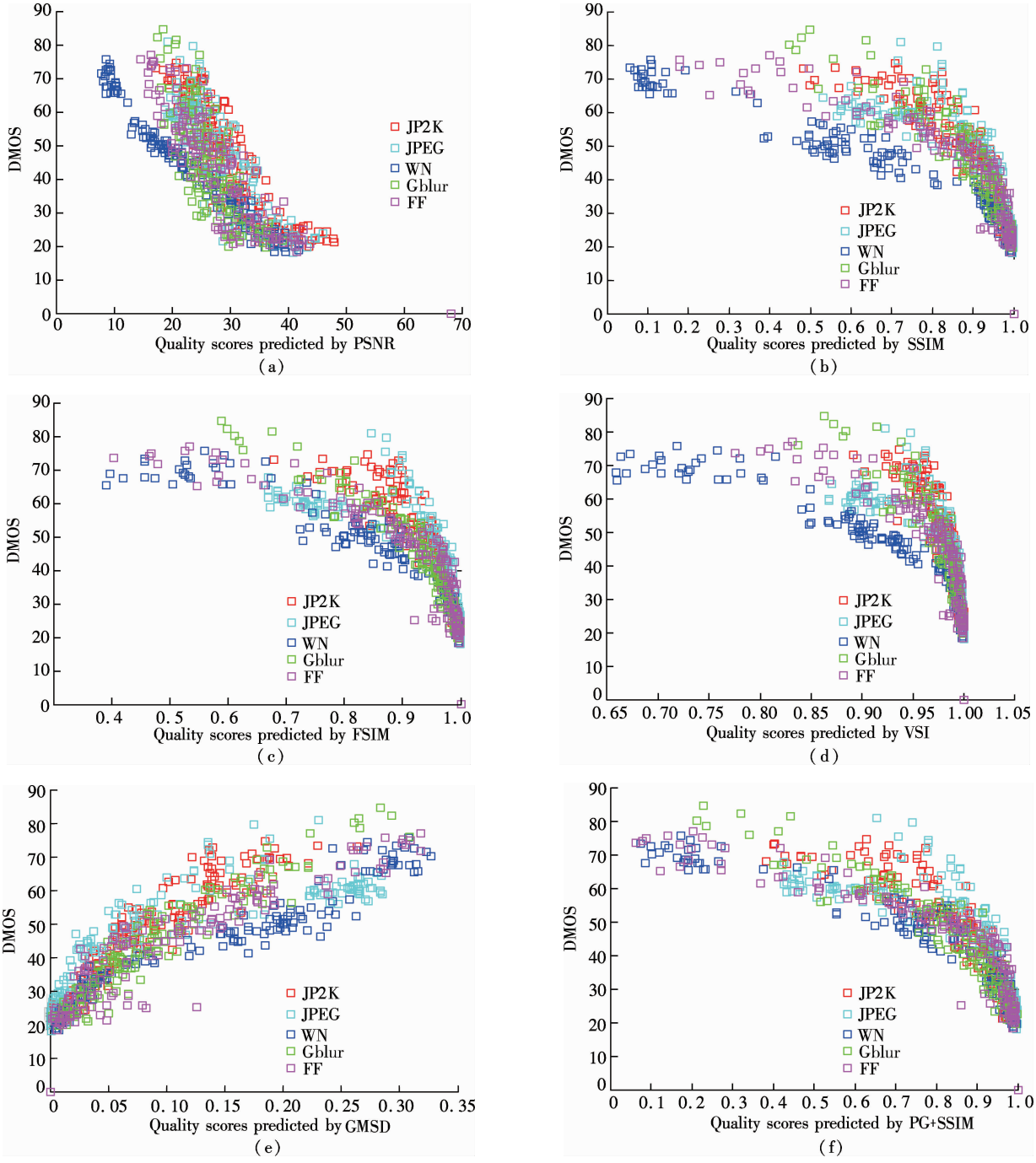

Fig. 3 Scatter plots of predicted quality scores against the subjective quality scores (DMOS) by representative FR-IQA metrics on the whole LIVE database. (a) PSNR; (b) SSIM; (c) FSIM; (d) VSI; (e) GMSD; (f) PG + SSIM

From Fig. 3, we can see that the proposed metric (PG + SSIM) yields good linear prediction for all of the distortion types, including JP2K, JPEG, WN, BLUR and FF. In addition, we also list the performance of different IQA models on the IVC and MICT databases[22] and IVC database[23] in Tab. 4.

To demonstrate the effectiveness of the proposed adaptive window-like model, we tabulate the performance of the improved version of SSIM, namely, MS-SSIM[4] and its perceptual grouping-induced one (PG + MS-SSIM) in Tab. 5.

Tab. 4 Performance on IVC and MICT databases

Method	IVC database			MICT database		
	SROCC	LCC	RMSE	SROCC	LCC	RMSE
PSNR	0.688 4	0.703 2	0.866 2	0.613 2	0.650 3	0.063 4
SSIM	0.901 8	0.911 9	0.499 9	0.879 4	0.888 7	0.573 8
FSIM	0.926 2	0.937 6	0.423 6	0.905 0	0.906 5	0.528 3
GMSD	0.914 6	0.892 6	0.549 4	0.852 8	0.858 2	0.642 4
VSI	0.899 3	0.912 0	0.499 9	0.865 9	0.869 7	0.617 7
PG + SSIM	0.932 6	0.942 2	0.408 2	0.927 6	0.932 3	0.452 6

Furthermore, the PG-based metric is computationally efficient. When combining the PG with the SSIM on a 2.66 GHz Intel Core2 Quad CPU with 4 GB of RAM, the metric takes about 0.51 s to measure the similarity of a pair of 720 × 480 grayscale images (taken from the LIVE database), which is faster than FSIM (It takes about 0.75 s).

Tab. 5 Performance of the original version and perceptual grouping-induced version of MS-SSIM

Method	LIVE database			IVC database			MICT database		
	SROCC	LCC	RMSE	SROCC	LCC	RMSE	SROCC	LCC	RMSE
MS-SSIM	0.951 2	0.946 8	7.438 0	0.884 7	0.893 4	0.547 4	0.886 4	0.893 5	0.562 1
PG + MS-SSIM	0.978 3	0.968 4	6.171 0	0.941 6	0.948 9	0.389 2	0.937 9	0.940 8	0.443 5

3 Conclusion

To make the changed pixels in an image more visually salient to the HVS, we use the PG method to cluster the similar pixels into a group (or superpixel). Then, we calculate the similarity on superpixels, which are more perceptually relevant than the fixed-sized windows commonly used in the existing IQA methods. Compared with the state-of-the-art FR-IQA models, the experimental results show that the proposed metric yields the best overall performance in terms of correlation with human judgment. The proposed framework is generic which allows us to use any PG methods.

Better performance can be expected if the assessment is applied in a multiscale framework[25]. Further work also includes applying the proposed image quality metric to evaluate medical images in different modalities[26].

References

[1] Wang Z, Bovik A C. *Modern image quality assessment* [M]. San Rafael, CA, USA: Morgan & Claypool, 2006.

[2] Wang Z, Bovik A C, Lu L. Why is image quality assessment so difficult[C]//*International Conference on Acoustics, Speech, and Signal Processing*. Orlando, FL, USA, 2002: 3313 - 3316.

[3] Wang Z, Bovik A C, Sheikh H R, et al. Image quality assessment: From error visibility to structural similarity [J]. *IEEE Transactions on Image Processing*, 2004, **13**(4): 600 - 612.

[4] Wang Z, Simoncelli E P, Bovik A C. Multiscale structural similarity for image quality assessment[C]//*37th Asilomar Conference on Signals, Systems, and Computers*. Pacific Grove, CA, USA, 2003: 1398 - 1402.

[5] Li C, Bovik A C. Three-component weighted structural similarity index[C]//*SPIE*. San Jose, CA, USA, 2009: 72420Q1 - 72420Q9.

[6] Wang Z, Li Q. Information content weighting for perceptual image quality assessment [J]. *IEEE Transactions on Image Processing*, 2011, **20**(5): 1185 - 1198. DOI: 10.1109/TIP.2010.2092435.

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

[8] Zhang L, Zhang D, Mou X Q, et al. FSIM: A feature similarity index for image quality assessment [J]. *IEEE Transactions on Image Processing*, 2011, **20**(8): 2378 - 2386. DOI: 10.1109/TIP.2011.2109730.

[9] Zhang L, Shen Y, Li H Y. VSI: a visual saliency-induced index for perceptual image quality assessment [J]. *IEEE Transactions on Image Processing*, 2014, **23**(10): 4270 - 4281. DOI: 10.1109/TIP.2014.2346028.

[10] Xue W F, Zhang L, Mou X Q, et al. Gradient magnitude similarity deviation: a highly efficient perceptual image quality index [J]. *IEEE Transactions on Image Processing*, 2014, **23**(2): 684 - 695.

[11] Zhang L, Zhang L, Mou X Q, et al. A comprehensive evaluation of full reference image quality assessment algorithms[C]//*19th International Conference on Image Processing*. Orlando, FL, USA, 2012: 1477 - 1480.

[12] Lin W S, Jay K C C. Perceptual visual quality metrics: A survey [J]. *Journal of Visual Communication and Image Representation*, 2011, **22**(4): 297 - 312. DOI: 10.1016/j.jvcir.2011.01.005.

[13] Wertheimer M. *Laws of organization in perceptual forms* (*partial translation*) [M]. Harcourt Brace Jovanovich, 1938: 71 - 88.

[14] Achanta R, Shaji A, Smith K, et al. Slicsuperpixels compared to state-of-the-art superpixel methods [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012, **34**(11): 2274 - 2282. DOI: 10.1109/TPAMI.2012.120.

[15] Felzenszwalb P F, Huttenlocher D P. Efficient graph-based image segmentation [J]. *International Journal of Computer Vision*, 2004, **59**(2): 167 - 181. DOI: 10.1023/B: VISI.0000022288.19776.77.

[16] Shi J B, Malik J. Normalized cuts and image segmentation [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000, **22**(8): 888 - 905.

[17] Moore A P, Prince S J, Warrell J, et al. Superpixel lattices[C]//*IEEE Conference on Computer Vision and Pattern Recognition*. Anchorage, AK, USA, 2008: 1 - 8.

[18] Levinshtein A, Stere A, Kutulakos K N, et al. Turbopixels: Fast superpixels using geometric flows [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009, **31**(12): 2290 - 2297. DOI: 10.1109/TPAMI.2009.96.

[19] Veksler O, Boykov Y, Mehrani P. Superpixels and supervoxels in an energy optimization framework[C]//*European Conference on Computer Vision*. Heraklion, Greece, 2010: 211－224.
[20] Kong Y Y, Deng Y, Dai Q H. Discriminative clustering and feature selection for brain MRI segmentation [J]. *IEEE Signal Processing Letters*, 2015, **22**(5): 573－577.
[21] Sheikh H R, Wang Z, Cormack L, et al. LIVE image quality assessment database release 2 [EB/OL]. [2007-06-30]. http://live.ece.utexas.edu/.
[22] Horita Y, Shibata K, Kawayoke Y, et al. MICT image quality evaluation database [EB/OL]. (2000) [2012-11-12]. http://mict.eng.u-toyama.ac.jp/mictdb.html.
[23] Ninassi A, Calet P L, Autrusseau F. Pseudo no reference image quality metric using perceptual data hiding[C]// *SPIE Human Vision and Electronic Imaging*. San Jose, CA, USA, 2006: 146－157.
[24] Sheikh H R, Sabir M F, Bovik A C. A statistical evaluation of recent full reference image quality assessment algorithms [J]. *IEEE Transactions on Image Processing*, 2006, **15** (11): 3440－3451. DOI: 10.1109/TIP.2006.881959.
[25] Chen Y, Huang S Y, Pickwell-MacPherson E. Frequency-wavelet domain deconvolution for terahertz reflection imaging and spectroscopy [J]. *Optical Express*, 2010, **18** (2): 1177－1190. DOI: 10.1364/OE.18.001177.
[26] Chen Y, Shi L Y, Feng Q J, et al. Artifact suppressed dictionary learning for low-dose CT image processing [J]. *IEEE Transactions on Medical Imaging*, 2014, **33**(12): 2271－2292. DOI: 10.1109/TMI.2014.2336860.

基于感知分组理论的图像质量评价

王同罕[1] 张 璐[2] 贾惠珍[3] 孔佑勇[1] 李宝生[1,4] 舒华忠[1]

([1] 东南大学影像科学与技术实验室, 南京 210096)
([2]IETR Lab (UMR CNRS 6164), INSA de Rennes, 20 Avenue des Buttes de Coesmes, CS 70839F-35708 Rennes Cedex 7, France)
([3] 南京理工大学计算机科学与工程学院, 南京 210094)
([4] 山东省肿瘤医院, 济南 250117)

摘要:为了更好地利用人类视觉系统特性,采用已经被证明在人类视觉系统中具有重要作用的感知分组策略来设计图像质量评价模型.与目前的基于固定窗口的方法不同,所提方法通过将感知分组策略融入图像质量评价中,实现了以一种自适应窗口的方式来评价图像质量.算法流程如下:首先,通过超像素方法将相似像素尽最大限度地聚集到一组;其次,对参考图像和失真图像的对应超像素进行基于结构相似度的计算;最后,综合图像中的所有超像素的相似性得到最终的评价结果.在3个图像数据库(LIVE, IVC 和 MICT)上的实验结果显示,该方法具有很好的预测性能.

关键词:感知分组;感知图像质量评价;超像素;全参考
中图分类号:TN911.73