Visual Attention Fusion Framework for Image Retargeting Quality Assessment

Visual Attention Fusion Framework for Image Retargeting Quality Assessment

Shuai Zhang, Yuzhen Niu, Jiawen Lin*and Junhao Chen

College of Mathematics and Computer Science, Fuzhou University, Fuzhou, China
zs614125016@gmail.com

Abstract. To adapt images for diversified digital devices, researchers have presented many image retargeting methods. However, the consistency between results of objective image retargeting quality assessment (IRQA) metrics and subjective perception is still low. In this paper, we propose a visual attention fusion (VAF) framework to assist IRQA metrics in better understanding the features of images such as image saliency, faces, and lines. First, we combine the results of multiple salient object detection algorithms to reduce the limitations of a single algorithm. Second, faces and lines are considered in our framework to measure deformations to these visually sensitive regions. Finally, we propose a saliency enhancement model to simulate human visual attention for IRQA. We combine the proposed VAF framework with some state-of-the-art IRQA metrics. Experimental results show that the proposed VAF framework can improve the consistency between the results of objective IRQA metrics and subjective opinion scores.

1. Introduction

Differences exist between IRQA and general-purpose image quality assessment (IQA), which makes IRQA a challenging research problem. First, there are resolution and aspect ratio differences between the retargeted and original images. Second, the types of distortion in the retargeted image are different from general image distortion types, such as blur, noise, and compression. Distortions in a retargeted image mainly include geometric distortion and information loss. Finally, subjective perceptual quality of a retargeted image is closely related to humans' cognition of this image, such as that of the structure and integrity of an object [7]. Human beings have comprehensive priori knowledge of the structure and integrity of human faces and common natural objects. Therefore, when evaluating the visual distortion of an image, special attention should be given to faces, dominant lines, and salient objects in an image.

Early IRQA works such as the Edge Histogram (EH) [1], Color Layout (CL) [2], Earth Mover's Distance (EMD) [3], and SIFT-flow [4] evaluated the quality of a retargeted image by measuring the distance between the original and retargeted images. These metrics can effectively measure the similarities of image content and structure between two images. However, the consistency values between the results of objective IRQA metrics and subjective scores are usually low, due to the negligence of human visual attention mechanism for IRQA.

Therefore, to better evaluate the visual quality of retargeted images, recent works have incorporated human visual attention into IRQA. Liu *et al.* [5] used SIFT-flow to establish local pixel correspondence between two images, and then exploited a saliency weighted similarity metric to measure the quality of retargeted images. Hsu *et al.* [6] introduced saliency maps to simulate the subjective perception of geometric distortions. They also used saliency loss to measure information loss in retargeted images. Chen et al. [7] proposed a Bi-Directional Salient Information Loss (BDSIL) measurement to measure the salient information loss in a bi-directional manner. Zhang et al. [8] proposed to use backward registration to simulate the geometric transformations that an image experienced during retargeting. A saliency-weighted aspect ratio similarity was defined as the quality of the retargeted image.

In this paper, we propose a visual attention fusion (VAF) framework that is suitable for objective quality assessment of retargeted images. The main contributions are as follows. First, we combine the results of multiple salient object detection algorithms to reduce the limitations of a single algorithm. Second, considering human's priori knowledge of faces and lines, our method adaptively increases the visual attention values of faces and lines. Finally, we propose a saliency enhancement model to produce a visual importance map that is more consistent with human visual attention for IRQA.

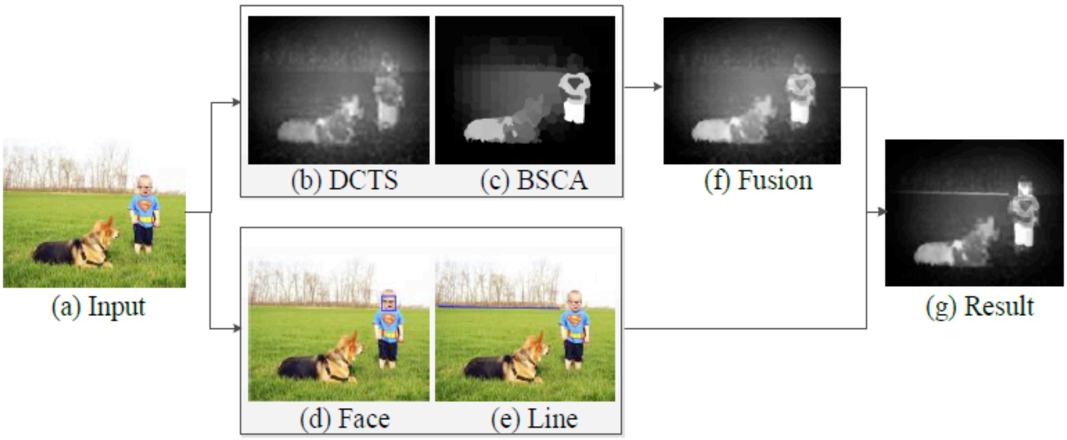
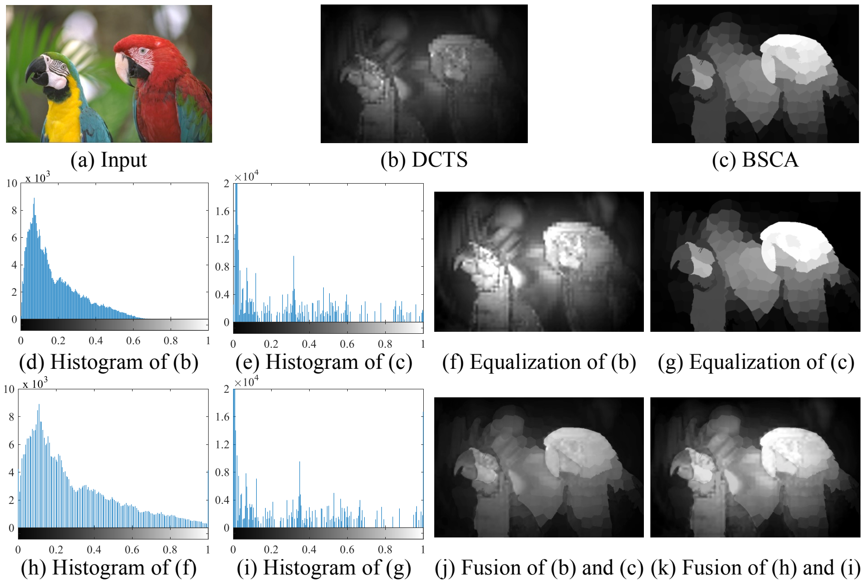


Figure 1. Overall framework of the proposed VAF framework.

2. Proposed Method

Human visual attention plays a major role in the objective quality assessment of retargeted images. Subjective perception of retargeted images mainly considers shape distortion and information loss. Therefore, to obtain more accurate assessment results,
computing a visual importance map, which is more in line with human visual attention for IRQA, is necessary. The overall framework of our proposed VAF framework is shown in Figure. 1.


(a) Input (b) DCTS (c) BSCA

(d) Histogram of (b) (e) Histogram of (c) (f) Equalization of (b) (g) Equalization of (c)

(h) Histogram of (f) (i) Histogram of (g) (j) Fusion of (b) and (c) (k) Fusion of (h) and (i)

Figure 2. Example of equalization operation

2.1. Saliency Map Fusion

Work [6] found that different salient object detection algorithms achieve different performance in IRQA. Because each salient object detection algorithm focuses on certain aspects of image content, fusing saliency maps computed by different salient object detection algorithms may obtain a more comprehensive saliency map. Therefore, in our framework, we first fuse two saliency maps that are computed by two salient object detection algorithms.

The first selected algorithm is based on Discrete Cosine Transform (DCTS) [9] coefficients, and is widely used for image retargeting and IRQA [6,8]. The second algorithm, namely, the Background-based map optimized via Single-layer Cellular Automata (BSCA) [10], previously showed good performance on six public datasets. We fuse the result saliency maps of these two algorithms to obtain a more accurate saliency map. A straightforward fusion method is to take the mean or max saliency value at each pixel directly. The straightforward fusion method does not work well because large differences exist in the saliency distribution of the two saliency maps, as shown in Figure. 2 (d), (e), (h), and (i).

To minimize the distribution differences while preserving the overall distributions of the original saliency maps, we first equalize the two saliency maps. Figure. 2 shows an example of our equalization operation, and the specific calculation process is as follows:

$$S'_p = \frac{\max(\min(S_p, t), b) - b}{t - b}, \tag{1}$$

where S_p and S'_p represent the saliency values of the pixel p before and after equalization, respectively. Parameters t and b are defined as follows:

$$\begin{cases} t = S^d_{p_t}, \quad p_t = floor(w \times h \times k) \\ b = S^d_{p_b}, \quad p_b = floor(w \times h \times (1-k)) \end{cases}, \tag{2}$$

where w and h represent the width and height of the original image, respectively, and S^d is a list of saliency values of saliency map S in descending order. In this paper, the default values of k are 0.02 and 0.05 for the equalizations of DCTS and BSCA saliency maps, respectively.

We then calculate the mean saliency value at each pixel of the equalized saliency maps of DCTS and BSCA, that is, $S^{D'}$ and $S^{B'}$. Subsequently, the mean saliency map S^m is normalized and the fusion saliency map S^F is obtained. The processes of fusion and normalization are as follows:

$$S^m_p = \frac{S^{D'}_p + S^{B'}_p}{2}, \tag{3}$$

$$S^F_p = \frac{S^m_p - \min(S^m)}{\max(S^m) - \min(S^m)}, \tag{4}$$

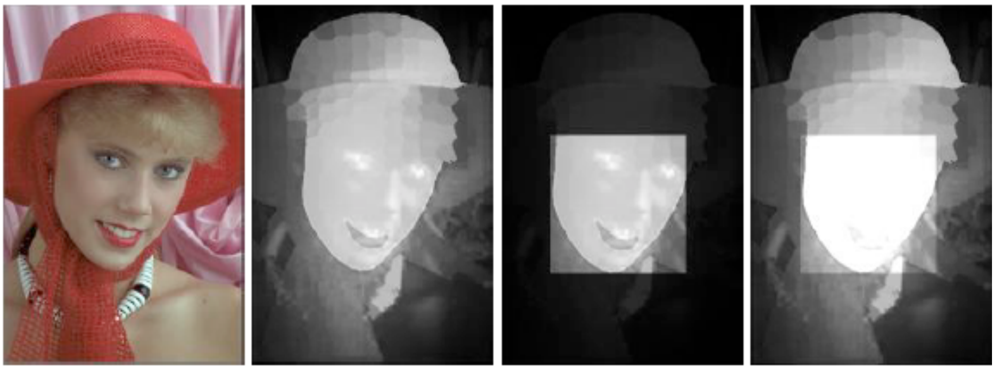
where $\min(S^m)$ and $\max(S^m)$ compute the minimum and maximum values of saliency map S^m, respectively.

2.2. Face and Line Enhancement

Considering only low-level features is not sufficient to cover all the factors that contribute to the degraded visual quality of retargeted images, and salient object detection algorithms do not specifically consider face and line information, they cannot detect faces and lines well. Therefore, to bring the results of objective IRQA metrics closer to subjective perception, and to enhance the sensitivity to the deformations in these regions, we consider face and line information in our proposed VAF framework.

We adopt the face++ toolkit [13] and Line Segment Detector (LSD) [14] to detect human faces and lines, respectively, in the original image. The face detection results include the position of the upper left vertex as well as the width and height of the face rectangle. The results of line detection include the coordinates of the start and end points of a line as well as the line width. In addition, because shorter lines may not receive much attention, we only consider lines that are longer than 1/3 the length of the diagonal of the original image.

Our method adaptively magnifies the saliency values of the regions covered by faces and lines to increase the importance of the corresponding regions. The purpose of our VAF framework is to encourage the face and line regions to become more important but not the only visual important regions in the saliency map. Over-emphasizing these regions will underestimate the importance of other salient regions.



(a) Input (b) Fusion (c) Results without/with constraint

Figure 3. Example of constraining the maximum value.

Therefore, we avoid over-emphasizing the face and line regions by constraining the magnified saliency value not to exceed the maximum value of 1:

$$S_p^F = \min_{p\in R_{F_i}}(S_p^F \times e^{C_1 \cdot (a_i/A)^{\frac{1}{4}}} ,1), \tag{5}$$

$$S_p^F = \min_{p\in R_{L_i}}(S_p^F \times e^{C_2 \cdot (l_i/L)^{\frac{1}{2}}} ,1) \tag{6}$$

,

where R_{F_i} and R_{L_i} represent the i-th face and line regions detected in the original image, respectively, and a_i and l_i represent the area of the face region and length of the line, respectively. A and L represent the area of the original image and length of the diagonal of original image, respectively, and C_1 and C_2 are the corresponding weight values, the defaults for which are 1.

In Figure. 3, we show an example of with and without constraining the maximum value. First, we can see that the fused saliency map of Figure. 3 (b) does not fully achieve the desired effect because the human face is not conspicuously prominent in relation to other regions. The left image of Figure. 3 (c) shows the result without constraining the maximum value, where the importance of the cap and other body parts are diminished considerably as a result of the over-magnification of the saliency values of the human face. We constrain the maximum value of the saliency map to avoid the aforementioned situation. The result is shown in the right image of Figure. 3 (c).

2.3. *Visual Attention Enhancement Model*

Work [6] explained that it is usually difficult for human beings to give consistent subjective scores for a retargeted image whose original has no obvious visual characteristics. By contrast, the subjective scores are generally consistent when dominant visual characteristics are present in the original image. Therefore, we further enhance the contrast of saliency maps after conducting salient feature detection. The saliency enhancement model is designed as follows:

$$\tilde{S}_p = S_p^F \times e^{C_3 \cdot S_p^F}, \tag{7}$$

Table 1. Performance comparison on MIT RetargetedMe dataset. The best results are formatted in boldface.

Metric	Mean KRCC in each subset						Total			
	Line Edge	*Face People*	*Foreground Objects*	*Texture*	*Geometric Structure*	*Symmetry*	Mean KRCC	Std KRCC	LCC	*p*-val
CSim [5]	0.097	0.290	0.293	0.161	0.085	0.071	0.164	0.263	0.274	0.028
VAF+[5]	**0.192**	**0.314**	**0.304**	**0.214**	**0.139**	**0.203**	**0.216**	**0.239**	**0.300**	**2e-4**
PGDIL [6]	0.431	0.390	0.389	0.286	**0.438**	**0.523**	0.415	0.296	**0.468**	6e-10
VAF+[6]	**0.437**	**0.504**	**0.458**	**0.356**	0.405	0.369	**0.438**	**0.294**	0.453	**4e-10**
ARS [8]	0.463	0.519	0.444	0.330	0.505	**0.464**	0.452	0.283	0.567	1e-11
VAF+[8]	**0.476**	**0.572**	**0.520**	**0.384**	**0.527**	0.372	**0.485**	**0.264**	**0.630**	**6e-14**

Table 2. Performance comparison on CUHK dataset. The best results are formatted in boldface.

Metric	PLCC	SRCC	RMSE	OR
CSim [5]	0.4374	0.4760	12.141	**0.1520**
VAF+[5]	**0.5361**	**0.5061**	**11.396**	**0.1520**
PGDIL [6]	0.5403	0.5409	11.361	0.1520
VAF+[6]	**0.5801**	**0.5807**	**10.997**	**0.1170**
ARS [8]	0.6835	0.6693	9.855	0.0702
VAF+[8]	**0.7157**	**0.6911**	**9.289**	**0.0585**

where S' represents the enhanced saliency map, and C_3 is a weight value, the default for which is 0.8. Finally, we normalize the enhanced saliency map to obtain our final visual attention fusion map.

3. Experimental Results

3.1. Image Retargeting Datasets

MIT RetargetMe Dataset. There are 37 original images in the MIT RetargetMe dataset [15], which are classified into six image attributes including: Line/Edge, Face/People, Texture, Foreground Objects, Geometric Structure and Symmetry, and each image may contain multiple attributes. The subjective scores of each retargeted image in the dataset were obtained by pairwise comparison. We adopted the mean and standard deviation values of the Kendall Rank Correlation Coefficient (KRCC), Linear Correlation Coefficient (LCC) and p-value to measure the consistency between the objective and subjective scores.

CUHK Dataset. The CUHK dataset [16] consists of 57 original and 171 retargeted images. The subjective tests of the dataset use a 5-level quality quantization strategy.

Table. 3. Performance comparison on two datasets with difference saliency map. The best results are formatted in boldface.

Metric	DCTS		BSCA		VAF	
	KRCC	PLCC	KRCC	PLCC	KRCC	PLCC
CSim [5]	0.174	0.4520	0.168	0.4412	**0.216**	**0.5361**
PGDIL [6]	0.415	0.5403	0.423	0.5613	**0.438**	**0.5801**
ARS [8]	0.452	0.6835	0.474	0.7016	**0.485**	**0.7157**

For this dataset, we employ four widely used evaluation metrics to evaluate the performance of objective IRQA metrics, including Pearson Linear Correlation Coefficient (PLCC), Spearman Rank-order Correlation Coefficient (SRCC), Root Mean Squared Error (RMSE) and Outlier Ratio (OR).

3.2. Results Analysis

We combined the proposed VAF framework with three IRQA metrics, including CSim [5], PGDIL [6], and ARS [8]. For the sake of fairness, we only replaced the saliency map with the visual attention map generated by the proposed VAF framework in the source code, whereas the parameters remained unchanged.

TABLE I shows the performance comparison on MIT dataset. First, the overall performance of all three IRQA metrics improves after being combined with the proposed VAF framework. Second, the performance of the IRQA metrics in the Line/Edge, Face/People, Foreground Objects and Texture categories has also been improved.

TABLE II shows the performance comparison on CUHK dataset. The experimental results show that the performance of all three IRQA metrics improves in all four evaluation metrics after being combined with our VAF framework.

In TABLE III, we report the statistical results for the influence of combining the saliency maps computed by DCTS and BSCA algorithms with different IRQA metrics. The experimental results on both MIT and CUHK datasets validate the effectiveness of the proposed VAF framework.

4. Conclusions

In this paper, we proposed a VAF framework for IRQA. We obtained more comprehensive salient object detection results by combining the saliency maps computed by DCTS and BSCA algorithms. Our method adaptively magnifies the importance of the human visual sensitive face and line regions, and avoids over-magnification by constraining the maximum value of the saliency map. Finally, to simulate human visual attention in IRQA, we proposed a saliency enhancement model. We combined the proposed VAF framework with three state-of-the-art IRQA metrics. Experimental results on two widely used public datasets showed that the proposed VAF framework can improve the consistency between the results of objective IRQA metrics and subjective opinion scores.

5. References

[1] B. S. Manjunath, J.-R. Ohm, V. V. Vasudevan, and A. Yamada,: Color and texture descriptors Circuits and Systems for Video Technology. IEEE Transactions on Circuits and Systems for Video Technology, vol. 11, no. 6, pp. 703–715 (2001).

[2] E. Kasutani and A. Yamada,: The MPEG-7 color layout descriptor: a compact image feature description for high-speed image/video segment retrieval. IEEE International Conference on Image Processing, vol. 1, pp. 674–677 (2001).

[3] Y. Rubner, C. Tomasi, and L. J. Guibas,: The earth mover's distance as a metric for image retrieval. IEEE International Journal of Computer Vision, vol. 40, no. 2, pp. 99–121 (2000).

[4] C. Liu, J. Yuen, A. Torralba, J. Sivic, and W. T. Freeman,: Sift flow: Dense correspondence across different scenes. European Conference on Computer Vision, pp. 28–42 (2008).

[5] Y.-J. Liu, X. Luo, Y.-M. Xuan, W.-F. Chen, and X.-L. Fu,: Image retargeting quality assessment. Computer Graphics Forum, vol. 30, no. 2, pp. 583–592 (2011).

[6] C.-C. Hsu, C.-W. Lin, Y. Fang, and W. Lin,: Objective Quality Assessment for Image Retargeting Based on Perceptual Geometric Distortion and Information Loss. IEEE Journal of Selected Topics in Signal Processing, vol. 8, no. 3, pp. 377–389 (2014).

[7] Z. Chen, and J. Lin, and N. Liao, and C. W. Chen,: Full Reference Quality Assessment for Image Retargeting Based on Natural Scene Statistics Modeling and Bi-Directional Saliency Similarity. IEEE Transactions on Image Processing, vol. 26, no. 11, pp. 5138–5148 (2017).

[8] Y. Zhang, Y. Fang, W. Lin, X. Zhang, and L. Li,: Backward registration based aspect ratio similarity for image retargeting quality assessment. IEEE Transactions on Image Processing, vol. 25, no. 9, pp. 4286–4297 (2016).

[9] Y. Fang, Z. Chen, W. Lin, and C.-W. Lin,: Saliency detection in the compressed domain for

adaptive image retargeting. IEEE Transactions on Image Processing, vol. 21, no. 9, pp. 3888–3901 (2012).

[10] Y. Qin, H. Lu, Y. Xu, and H. Wang,: Saliency detection via cellular automata. Computer Vision and Pattern Recognition, pp. 110–119 (2015).

[11] Downing PE, Bray D, Rogers J, and Childs, C,: Bodies capture attention when nothing is expected. Cognition, vol. 93, no. 1, pp. B27 (2004).

[12] Biederman I,: Recognition-by-components: a theory of human image understanding. Psychological Review, vol. 94, no. 2, pp. 115-47 (1987).

[13] Face++ Research Toolkit, www.faceplusplus.com, last accessed 2018/4/24.

[14] R. G. von Gioi, J. Jakubowicz, J.-M. Morel, and G. Randall,: LSD: A fast line segment detector with a false detection control. IEEE Transactions on Pattern Analysis & Machine Intelligence, vol. 32, no. 4, pp. 722–732 (2010).

[15] M. Rubinstein, D. Gutierrez, O. Sorkine, and A. Shamir,: A comparative study of image retargeting. ACM SIGGRAPH Asia, vol. 29, no. 6, pp. 160 (2010).

[16] L. Ma, W. Lin, C. Deng, and K. N. Ngan,: Image retargeting quality assessment: A study of subjective scores and objective metrics. IEEE Journal of Selected Topics in Signal Processing, vol. 6, no. 6, pp. 626–639 (2012).