ReSIFT: RELIABILITY-WEIGHTED SIFT-BASED IMAGE QUALITY ASSESSMENT

Dogancan Temel and Ghassan AlRegib

Center for Signal and Information Processing (CSIP)
School of Electrical and Computer Engineering
Georgia Institute of Technology, Atlanta, GA, 30332-0250 USA
{cantemel,alregib}@gatech.edu

ABSTRACT

This paper presents a full-reference image quality estimator based on SIFT descriptor matching over reliability-weighted feature maps. Reliability assignment includes a smoothing operation, a transformation to perceptual color domain, a local normalization stage, and a spectral residual computation with global normalization. The proposed method **ReSIFT** is tested on the LIVE and the LIVE Multiply Distorted databases and compared with 11 state-of-the-art full-reference quality estimators. In terms of the Pearson and the Spearman correlation, **ReSIFT** is the *best performing* quality estimator in the overall databases. Moreover, **ReSIFT** is the best performing quality estimator in at least one distortion group in compression, noise, and blur category.

Index Terms— perceptual image quality assessment, scale invariant feature transform (SIFT), spectral residual, reliable descriptor matching

1. INTRODUCTION

Image qualiy assessment methods algorithmically evaluate quality of images and the definition of the quality depends on a target application. Mean squared error and peak signal-to-noise ratio (PSNR) are commonly used in image coding applications, where the quality criterian is based on fidelity. Human visual system characteristics can be used to enhance fidelity metrics in terms of perceptual correlation as in PSNR-HVS [1], PSNR-HVS-M [1], PSNR-HA [2], and PSNR-HMA [2].

Perception can also be introduced to image quality assessment algorithms using saliency-based approaches such as the spectral residual, which is the residual between a spectrum and an averaged spectrum. The spectral residual is based on the visual system characteristic that corresponds to suppressing responses to frequently occuring features and being sensitive to unexpected changes. The spectral residual approach is used to assign significance to gradient magnitudes to estimate image quality [3].

As an alternative to tracking changes in an intensity channel using a pixel-wise fidelity approach or focusing on sharp changes using a gradient magnitude operator, we can solely focus on discriminative features that are perceptually significant. The authors in [4] use number of matched SIFT features to evaluate image quality. Instead of directly calculating the number of matched features over the whole image, the authors in [5] compute number of SIFT features in a unit region on the first octave of the difference of Gaussian scale space of a preprocessed image to obtain a no-reference quality assessment. In [6], the authors extract and match SIFT features to obtain affine transformation parameters and perform a reverse affine transformation to make full-reference quality assesment methods robust to deformations such as translation, rotation, scaling or skewing. In [7], SIFT features are used to match objects in an original image to potentially deformed objects in a processed image. Standard deviation values between matched SIFT points are used to measure the non-rigid deformation level of an object and a structural-similarity index is computed between regions centered around SIFT features inside object boundaries.

In the proposed work, we combine a smoothing operation, a color domain transformation, a normalization operation, and a spectral residual calculation to assign reliability to pixel maps. SIFT descriptors are extracted from the reliability-weighted pixel maps and percentile thresholds are computed from the distances between matched descriptors to obtain a quality indicator. Finally, a non-linear mapping is used to obtain the `ReSIFT` score. In Section 2, a detailed description of the proposed quality estimator is provided including all the details to guarantee reproducibility in research. A validation analysis including the description of databases, compared quality estimators, and results is given in Section 3 and we conclude our work in Section 4.

2. MAIN

2.1. Introduction to ReSIFT

The descriptor extraction process from an input image is summarized in Fig. 1. At first, an input image is smoothed out using a low-pass filter and then the smoothed image is transformed from the RGB to the La*b* color domain. A light-

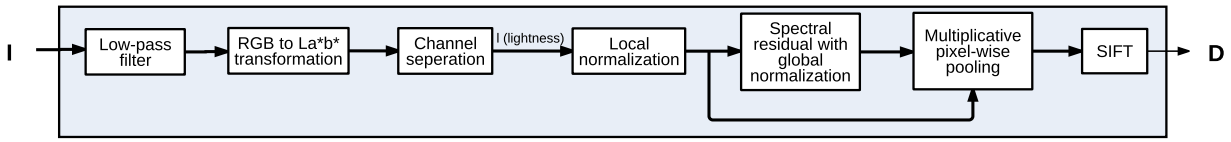


Fig. 1. Reliability-weighted SIFT descriptor extraction

ness map is separated and normalized locally, and a spectral residual with global normalization is computed from the locally normalized map. Then, the normalized lightness map and the normalized spectral residual map are multiplicatively pooled pixel-wise and SIFT descriptors are extracted from the pooled map. The overall ReSIFT pipline is given in Fig. 2, where reliability-weighted SIFT descriptors are matched, pooled, and non-linearly mapped to obtain an estiamted quality score. The proposed quality estimator `ReSIFT` does not use any chroma information.

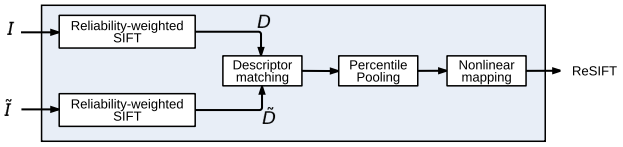


Fig. 2. ReSIFT pipeline

2.2. Low-pass filtering

A rotationally symmetric Gaussian low-pass filter is used to smooth out an image. The cutoff frequency of the low pass filter is a function of the filter size (f_{size}) and the standard deviation (f_σ).

2.3. Perceptually uniform color space transformation

An RGB image is transformed into the perceptually uniform color space La*b* to separate lightness from chroma information. Transformation parameters include a transformation matrix M, CIE standard coefficients κ and ϵ. A lightness channel is fed to the following blocks and chroma channels are not used due to lack of structural information.

2.4. Local normalization

A local normalization operation over the lightness channel is performed by a mean subtraction and a divisive normalization operation. The local mean formulation is given as

$$\mu[m,n] = \frac{1}{W^2} \sum_{\hat{m}=floor(\frac{m}{W})\cdot W+1}^{(floor(\frac{m}{W})+1)\cdot W} \sum_{\hat{n}=floor(\frac{n}{W})\cdot W+1}^{(floor(\frac{n}{W})+1)\cdot W} l[\hat{m},\hat{n}], \tag{1}$$

where m and n are pixel indices, l is the lightness channel, W is the window size and floor is an operator that rounds a number to the next smaller integer. The standard deviation of each window is formualted as

$$\sigma[m,n] = \sqrt{\sum_{\hat{m}=floor(\frac{m}{W})\cdot W+1}^{(floor(\frac{m}{W})+1)\cdot W} \sum_{\hat{n}=floor(\frac{n}{W})\cdot W+1}^{(floor(\frac{n}{W})+1)\cdot W} \frac{(l[\hat{m},\hat{n}]-\mu[\hat{m},\hat{n}])^2}{W^2}}, \tag{2}$$

where σ is a standard deviation map. The local normalization operation is composed of two steps. First, a local mean ($\mu[m,n]$) is subtracted from each pixel in the lightness map. Then, each mean shifted value is divided by a local standard deviation ($\sigma[m,n]$).

2.5. Spectral residual with global normalization

The normalized lightness map (l_{norm}) is transformed from the spatial domain to the frequency domain using the Fourier transform ($\mathscr{F}$). The magnitude component of the transformed map is expressed as

$$|L| = |(\mathscr{F}[l_{norm}])|, \tag{3}$$

where $|\cdot|$ is the magnitude operator and the phase component is given as

$$\angle L = \underline{\angle(\mathscr{F}[l_{norm}])}, \tag{4}$$

where $\angle$ is the phase operator. The spectrum of a signal is computed by the log of the magnitude ($\log|L|$) and the average spectrum is computed by convolving the spectrum with an averaging filter (g). The difference between the spectrum and the averaged spectrum results in the spectral residual, which is fomulated as

$$SR(L) = \log|L| - g * \log|L|, \tag{5}$$

where $*$ is the convolution operator and g is the averaging filter. The spectral residual is combined with the phase of the normalized lightness map, and then an inverse Fourier transform is performed. In [8], a reconstructed image is used as a saliency map, which is referred as the unexpected portion of an image.

A reconstruction operation using the spectral residual is formulated as

$$S = h * \mathscr{F}^{-1}[SR(L)\underline{\angle L}], \tag{6}$$

where $\mathscr{F}^{-1}$ is the inverse Fourier transform operator, h is the Gaussian low-pass filter used to smooth out a reconstructed map and $[\cdot\angle\cdot]$ is the representation of a signal in terms of its magnitude and its phase . The reconstructed map is globally normalized so that the pixel values are in between 0.0 and 1.0.

2.6. Multiplicative pixel-wise pooling

The locally normalized lightness map and the spectral residual-based resonstructed map are multiplicatively pooled pixel-wise to obtain a reliability-weighted lightness map.

2.7. Scale-invariant Feature Transform

Scale-invariant feature transform (SIFT) is an algorithm [9] that takes an image as an input and outputs scale-invariant coordinates relative to local features. These features are desinged to be fully invariant to a translation, a scaling, and a rotation operation, and partially invariant to a change in the illumination and the camera viewpoint. The characteristics

of these features resemble the properties of neurons in the inferrior temporal cortex of primate vision system [10].

To obtain a SIFT feature vector, an input image is convolved using Gaussian kernels with nearby scales to obtain feature maps. A difference between these feature maps is calculated to obtain a difference of Gaussian (DoG) map. Each pixel in the DoG map is compared to its eight neighbors in the current map and nine neighbors in the scales above and below. If the center pixel is larger or smaller than all other pixels, it is selected as an extrema point. Rather than selecting a central point directly as a keypoint, a 3D quadratic function is fitted to sample points to obtain an interpolated location. Directions of local image gradients are used to assign orientations to the keypoints and a histogram orientation is formed using 36 bins that correspond to 360 degrees. Each sample in the historgram is weighted by a gradient magnitude and a Gaussian-weighted circular window. The highest peak and any other local peak within 80% of the highest peak are used in the keypoint description. Therefore, keypoints with different orientations can have the same location and the same scale.

The location and the scale of a keypoint descriptor are used while computing the gradient magnitudes and the orientations. The magnitudes are weighted with a Gaussian function, whose standard deviation is a function of the descriptor window size. The orientation of a keypoint is used to rotate the computed gradient orientations relatively. A SIFT descriptor is obtained using a $4x4$ array of histograms with 8 orientations, which lead to a feature vector of length $4x4x8 = 128$. In the proposed method ReSIFT, SIFT descriptors are extracted over the reliability-weighted lightness maps.

2.8. Descriptor matching

The SIFT descriptors are matched based on the distance between them. Two descriptors are matched only if a threshold ($thresh$) times the Euclidean distance between them is less than the distance between that descriptor and other descriptors. Moreover, clusters of descriptors are analyzed based on their geometric characteristics to reject errenous matches.

2.9. Percentile pooling

The SIFT descriptors are commonly used in object recognition, where target objects can be located anywhere in compared images. However, in case of image quality assessment, objects are only slighly rotated, translated or deformed because of distortions. Therefore, a mean distance between descriptors can be misleading. We use a percentile pooling strategy, which only requires a single parameter ($perc$), to obtain a threshold that only contains relativley small distances among matched descriptors.

2.10. Nonlinear mapping

The percentile pooling threshold is proportional to the distance among the descriptors but the estimated quality is inversly proportional. Therefore, reciprocal of the percentile can be used to estiamte a quality score. An euclidean distance among $128D$ descriptors leads to values that are in the range of tens of thousands. To scale the range of the quality estimator, the percentile threshold is divided by a constant (C_1). We also add a constant (C_2) next to the divison in the reciprocal to avoid instabilities in case of small distance values. The nonlinear mapping function is given as

$$ReSIFT = \frac{1}{\frac{dist}{C_1} + C_2} \tag{7}$$

where $dist$ corresponds to the percentile threshold, C_1 and C_2 are coefficients and $ReSIFT$ is the estiamted quality score.

2.11. Parameter setup

All of the parameters used in the implementation of the proposed quality estimator are listed in Table 1. VLFeat library [11] is used for SIFT without any parameter tuning. The default values are used in the color space transformation and the spectral normalization. In the descriptor matching and the percentile pooling, the parameters are slightly tweaked. The low-pass filtering and the local normalization parameters are selected by visually assessing the feature maps and the nonlinear mapping values are set to fix the highest score to 100 and to stabilize the proposed method.

Table 1. List of the parameters and their values in ReSIFT

Section/Block	Parameter	Value
2.2. Low-pass filtering	f_{size}	4
	f_σ	5
2.3. Color space transformation	κ	903.3 CIE standard
	ϵ	0.008856
	M	Adobe RGB stand. 1998
2.4. Local normalization	W	20
2.5. Spectral residual	g_{size}	3
	h_{size}	10
	h_σ	3.8
2.8. Descriptor matching	$thresh$	1.4
2.9. Percentile pooling	$perc$	5
2.10. Nonlinear mapping	C_1	100,000
	C_2	0.01

3. VALIDATION

3.1. Databases

ReSIFT is validated using the LIVE [12] and the LIVE Multiply Distorted (MULTI) [13] databases. All of the distortion types in these databases can be grouped into four categories. Compression includes Jpeg, Jp2k, and Jpeg of blurred images. Noise contains white noise over reference images and white noise over blurred images. Communication includes Rayleigh fast-fading channel model errors and Blur consists of Gaussian blur. The number of images in each category is summarized in Table 2. Blur-Noise and Blur-Jpeg are included in two different categories since pristine images are processed with both of the distortion types simultaneously.

Table 2. The number of distorted images with respect to degradation categories in each database

	LIVE	MULTI	Total
Compression	460	225	685
Noise	174	225	399
Communication	174	-	174
Blur	174	450	624

Table 3. Performance of IQA methods on different databases

	PSNR	PSNR-HA	PSNR-HMA	SSIM	MS-SSIM	CW-SSIM	IW-SSIM	SR-SIM	FSIM	FSIMc	PerSIM	ReSIFT
	Pearson Correlation Coefficicent											
LIVE	0.927	0.953	0.958	0.945	0.946	0.872	0.951	0.945	0.949	0.950	0.955	**0.961**
MULTI	0.739	0.801	0.821	0.812	0.802	0.379	0.847	0.888	0.818	0.821	0.852	**0.906**
	Spearman Correlation Coefficient											
LIVE	0.909	0.937	0.944	0.949	0.951	0.902	0.960	0.955	0.961	0.959	0.950	**0.962**
MULTI	0.677	0.714	0.742	0.860	0.836	0.630	0.883	0.866	0.863	0.866	0.818	**0.887**

Table 4. Performance of best performing IQA methods on different distortion types

Distortion Types	Databases	Pearson Correlation				Spearman Correlation				
		PSNR-HMA	SR-SIM	PerSIM	ReSIFT	IW-SSIM	FSIM	FSIMc	SR-SIM	ReSIFT
Compression	**Jp2k [LIVE]**	**0.982**	0.957	0.976	0.972	0.979	0.981	**0.982**	0.972	0.971
	Jpeg [LIVE]	**0.972**	0.950	0.959	0.964	0.960	**0.962**	**0.962**	0.954	0.955
	Blur-Jpeg [MULTI]	0.838	0.904	0.864	**0.921**	0.869	0.854	0.855	0.862	**0.886**
Noise	**Wn [LIVE]**	0.985	0.974	0.968	**0.986**	0.982	0.980	0.979	0.983	**0.984**
	Blur-Noise [MULTI]	0.803	0.871	0.839	**0.897**	**0.893**	0.864	0.869	0.863	0.882
Communication	**FF [LIVE]**	**0.954**	0.943	0.946	0.949	0.967	0.970	**0.971**	0.966	0.959
Blur	**GBlur [LIVE]**	0.950	0.949	0.967	**0.971**	**0.983**	**0.983**	**0.983**	0.978	0.979
	Blur-Jpeg [MULTI]	0.838	0.904	0.864	**0.921**	0.869	0.854	0.855	0.862	**0.886**
	Blur-Noise [MULTI]	0.803	0.871	0.839	**0.897**	**0.893**	0.864	0.869	0.863	0.882

3.2. Performance metrics

The performance of the quality estimators are validated using the Pearson and the Spearman correlation coefficients. Since linearity-based Pearson correlation is sensitive to the range and to the distribution of the scores, a monotonic regression can be used before the correlation calculation. We use the monotonic formulation that can be expressed as

$$S = \beta_1 \left(\frac{1}{1} - \frac{1}{2 + exp(\beta_2 (S_0 - \beta_3))} \right) + \beta_4 S_0 + \beta_5 \quad (8)$$

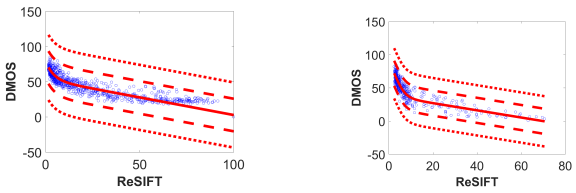
where S_0 is the input (raw value), S is the regressed output, and βs are the tuning parameters that are set according to the relationship between the quality estimates and the mean opinion scores.

3.3. State-of-the-art quality estimators

In the performance comparison, we use full-reference quality estimators based on fidelity, perceptually-extended fidelity, structural similarity, feature similarity, and perceptual similarity, which include PSNR, PSNR-HA, PSNR-HMA, SSIM, MS-SSIM, CW-SSIM, IW-SSIM, SR-SIM, FSIM, FSIMc, and PerSIM.

3.4. Results

Performances of the quality estimators in overall databases are summarized in Table 3. `ReSIFT` is the highest performing quality estimator in the LIVE and the MULTI databases in terms of the Pearson and the Spearman correlation as highlighted with bold. The best performing estimators in the LIVE database are not significantly different from eachother but the difference becomes significant in the MULTI. Three best-performing quality estimators in each database and correalation categoy include PSNR-HMA, SR-SIM, PerSIM, IW-SSIM, FSIM, FSIMc, and `ReSIFT`. Performances of these quality estimators in various distortion types are given in Table 4, where the highest performance in each category is highlighted. `ReSIFT` is the best performing quality estimator in at least one distortion group in compression, noise, and blur category. Scatter plots of `ReSIFT` corresponding to the LIVE and the MULTI databases are given in Fig. 3.4. The range of estimated quality socres are in between zero and hundred, and almost all the estimates are in the one standard deviation range with respect to the regression curve. Estimated quality scores use a wider quality range in the LIVE database compared to the MULTI database.



(a) LIVE-ReSIFT (b) MULTI-ReSIFT

Fig. 3. Scatter plots of ReSIFT

4. CONCLUSION

We proposed a reliability-weighted SIFT descriptor matching-based image quality estimator. The propsed method `ReSIFT` is used to quantify the perceptual quality of images under compression, noise, communication, and blur-based distortion types. When the LIVE and the LIVE Multiply Distorted image quality databases are considered, `ReSIFT` is the best performing quality estimator in at least one distortion group in compression, noise, and blur in terms of the Pearson and the Spearman correlation coefficients. Since proposed approach extracts features all over an image and relies solely on a lightness channel, `ReSIFT` is inherently not designed for local, global, and color-based distortions.

5. REFERENCES

[1] N. Ponomarenko, F. Silvestri, K. Egiazarian, M. Carli, J. Astola, and V. Lukin, "On between-coefficient contrast masking of dct basis functions," in *Proceedings of the 3rd Int Workshop on Video Processing and Quality Metrics for Consumer Electronics*, 2007, pp. 1–4.

[2] N. Ponomarenko, O. Ieremeiev, V. Lukin, K. Egiazarian, and M. Carli, "Modified image visual quality metrics for contrast change and mean shift accounting," in *International Conference The Experience of Designing and Application of CAD Systems in Microelectronics (CADSM)*, Feb 2011, pp. 305–311.

[3] L. Zhang and H. Li, "SR-SIM: A fast and high performance IQA index based on spectral residual," in *Image Processing (ICIP), 2012 19th IEEE International Conference on*, Sept 2012, pp. 1473–1476.

[4] G.-L. Wen, G. Liu, S.-G. Zheng, and S.-K. Ning, "Enhanced Image Quality Evaluation based on SIFT Feature," in *Machine Learning and Cybernetics (ICMLC), 2014 International Conference on*, July 2014, vol. 1, pp. 221–226.

[5] T. Sun, S. Ding, and X. Xu, "No-Reference Image Quality Assessment through SIFT Intensity," *Applied Mathematics and Information Sciences*, vol. 8(4), pp. 1925–1934, 2014.

[6] G. Chen and S. Coulombe, "An Image Visual Quality Assessment Method Based on SIFT Features," *Journal of Pattern Recognition Research*, vol. 1, pp. 85–97, 2013.

[7] M. Decombas, F. Dufaux, E. Renan, B. Pesquet-Popescu, and F. Capman, "A new object based quality metric based on SIFT and SSIM," in *Image Processing (ICIP), 2012 19th IEEE International Conference on*, Sept 2012, pp. 1493–1496.

[8] X. Hou and L. Zhang, "Saliency Detection: A Spectral Residual Approach," in *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, June 2007, pp. 1–8.

[9] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91–110, 2004.

[10] T. Serre, M. Kouh, C. Cadieu, U. Knoblich, G. Kreiman, and T. Poggio, "A Theory of Object Recognition: Computations and Circuits in the Feedforward Path of the Ventral Stream in Primate Visual Cortex," 2005.

[11] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," http://www.vlfeat.org/, 2008.

[12] H. R. Sheikh, L. Cormack, and A. C. Bovik, "LIVE Image Quality Assessment Database Release 2," http://live.ece.utexas.edu/research/quality, 2006.

[13] D. Jayaraman, A. K. Moorthy A. Mittal, and A. C. Bovik, "Objective Quality Assessment of Multiply Distorted Images," 2012, Proceedings of Asilomar Conference on Signals, Systems and Computers.