Journal of Computational Mathematics
Vol.34, No.6, 2016, 703–720.

http://www.global-sci.org/jcm
doi:10.4208/jcm.1611-m2016-0534

IMAGE QUALITY ASSESSMENT BASED ON CONTOUR AND REGION *

Chen Huang
LMAM, School of Mathematical Sciences, Beijing International Center for Mathematical Research, and Cooperative Medianet Innovation Center, Peking University, Beijing 100871, China
Email: chenhuang@pku.edu.cn
Ming Jiang
LMAM, School of Mathematical Sciences, Beijing International Center for Mathematical Research, and Cooperative Medianet Innovation Center, Peking University, Beijing 100871, China
Email: ming-jiang@pku.edu.cn
Tingting Jiang
NELVT, National Engineering Laboratory for Video Technology, School of Electronics Engineering and Computer Science, and Cooperative Medianet Innovation Center, Peking University, Beijing 100871, China
Email: ttjiang@pku.edu.cn

Abstract

Image Quality Assessment (IQA) is a fundamental problem in image processing. It is a common principle that human vision is hierarchical: we first perceive global structural information such as contours then focus on local regional details if necessary. Following this principle, we propose a novel framework for IQA by quantifying the degenerations of structural information and region content separately, and mapping both to obtain the objective score. The structural information can be obtained as contours by contour detection techniques. Experiments are conducted to demonstrate its performance in comparison with multiple state-of-the-art methods on two large scale datasets.

Mathematics subject classification: 68U10, 94A08.
Key words: Image quality assessment, Contour detection, Image segmentation.

1. Introduction

With the coming of information era, multimedia has become the primary carrier of information in our daily life. Digital images, as an important part, attracts tremendous attention. However, digital images are subject to a wide variety of distortions during the procedure such as acquisition, processing, compression, storage, transmission, and display. Image Quality Assessment as one of the fundamental problems attracts tremendous interest in recent years.

Generally IQA methods are classified into two categories: one is subjective assessment by humans and the other is objective assessment by human-designed algorithms. Image quality depends on its ultimate receiver, therefore subjective evaluation by humans is a correct criterion. Nevertheless, it's time-consuming, expensive and unable to be implemented in a real-time system. The objective assessment aims to develop computational models to automatically predict the image quality in consistent with subjective assessment.

* Received February 1, 2016 / Revised version received October 27, 2016 / Accepted November 11, 2016 / Published online December 16, 2016 /

According to the availability of reference source, objective assessment is classified as full-reference, reduced-reference and no-reference. In this paper, we only focus on full-reference methods. The interested readers are referred to the book of Wang and Bovik [1] for more details.

Due to the availability of datasets with human-labeled groundtruth, a variety of full-reference IQA methods are proposed in last decade. Existing approaches could roughly be divided into the following categories: the statistic of errors based model (e.g. Mean Squared Error (MSE), Peak Signal to Noise Ratio (PSNR)), Human Visual System (HVS) based model [2-4], structural similarity based model [5, 6], information theory based model [7, 8], and visual saliency based model [9-11]. Most approaches evaluate local quality at pixels based on patches, and generate a quality map after all pixel evaluation is finished.

In our opinion, a pixel as a low level representation unit is too fine to assess image quality. Human visual perception is adapted for extracting structural information such as contour and segmentation. The perceptual process of images is hierarchical: human first perceive global structural information such as contours and further focus on local regional details such as texture. In this paper, we propose a contour and region based framework for full-reference IQA. Our model separates an image into structure part and local regions. We detect the contour for representing structure, and use local descriptors for representing local region content. Existing models weight each pixel by information content in a low level. As opposed to this, we assess image quality in a higher level: on the one hand, we consider the contour as a whole and try to quantify the degeneration of the contour, on the other hand, we measure the degeneration of the region content. Finally we map both to obtain the objective score.

To evaluate the performance, we test our model on two large-scale benchmark datasets, LIVE2 [12] and TID2013 [13]. We demonstrate its promise through the comparison with multiple state-of-the-art objective methods.

The remainder of this paper is as follows. In Section 2, we review the related works. Next, we introduce our model in detail in Section 3. Section 4 covers the algorithms of two modules. We present experiments in Section 5, and discuss in Section 6.

2. Related Works

In this section, we first cover the review of full-reference IQA, contour detection and image segmentation, then introduce the most related work to our model.

2.1. Full-reference Image Quality Assessment

Traditional full-reference methods, including MSE, PSNR, base on the statistic of errors. They have been the dominant quantitative metrics in image quality assessment for decades due to "their simplicity to calculate, clear physical meanings, and mathematical convenience in the context of optimization" [5]. But these methods focus on the difference on single pixel independently, ignoring the fact that neighbouring pixels in an image are not independent but highly correlated. Furthermore, spatial structure in an image contains abundant visual information. So it's inevitable these metrics don't correlate well with perceptual image quality.

To solve above problems, researchers make great efforts to take the characteristics of HVS into account. Representative work are referred to Just Noticeable Difference (JND) model [2], Noise Quality Measure (NQM) [3], Visual Signal-to-Noise Ratio (VSNR) [4] and so on. JND penalizes the errors in accordance with visibility, considering the spatial contrast sensitivity

and contrast masking. NQM and VSNR emphasize the sensitivity to luminance, contrast, and frequency content. Overall HVS-based model attempts to model the properties of HVS, and makes an improvement compared with traditional methods. Nevertheless, HVS is a complex and highly nonlinear system, we have little knowledge to model its properties, hence it's still a challenging problem. This issue limits the development of HVS-based methods, but subsequent methods all take use of the knowledge of HVS implicitly.

Under the assumption that human visual perception is highly adapted for extracting structural information from a scene, Wang *et al.* [5] propose the structural similarity index (SSIM). SSIM tries to measure the degradation of local structural information as the perceptual quality and surpasses the traditional methods across-the-aboard. Later, Wang *et al.* [6] propose a multiscale extension of SSIM which produces a better result than SSIM in single scale.

From the view of information theory, Sheikh *et al.* [7] regard the full-reference quality assessment as the measure of information fidelity, which measures the shared information between the reference image and the distorted image. They propose the Information Fidelity Criterion (IFC) and its extension, Visual Information Fidelity (VIF) [8].

Another line is to adopt appropriate strategies at the pooling stage of quality score. Wang and Li [9] propose a modified version of SSIM weighted by information content. Zhang *et al.* [10] choose phase congruency as a low-level feature and a weighting function to derive a single similarity score. In their later work [11], they replace the phase congruency with visual saliency, proposing a metric called Visual Saliency-Induced (VSI), which is confirmed to get beneficial.

2.2. Contour Detection and Image Segmentation

Contour detection and image segmentation are related but not identical problems. They both correspond to the edges of objects in some way. However, contour detection may produce discontinuous result but image segmentation which aims to partition an image into several regions, leads to closed result. There are different lines of approaches to investigate those problems.

Traditional methods of contour detection base on first-order and second-order derivative of the gray-value of neighborhoods. Local operator methods including Prewitt, Sobel, Log, Roberts and Canny operator [14], detect contour by convolving a gray-scale image with special local operators. Later, filters of multiple scales and orientations are adapted to describe the contour. Morrone and Owens [15] propose the Oriented Energy model by using quadrature pairs of even and odd symmetric filters to detect contour.

In recent years, researchers try to combine the features in different channels and detect contour in a learning method. Martin *et al.* [16] propose the Pb feature which combines gradient in brightness, color, and texture channels. They use a logistic regression to predict the probability of a pixel belonging to the contour. Inspired by their work, subsequent researches go further. In [17], a multi-scale extension of Pb is proposed to work better. Arbelaez *et al.* [18] integrate multiple local cues into a globalization framework by spectral clustering, and recover regions from a contour detector. Ren and Bo [19] compute Sparse Code Gradients (SCG) which measures contrast using patch representations automatically learned through sparse coding.

On the other hand, different approaches have been proposed to segment images.

Image threshold segmentation is a simple and common method based on the threshold of the magnitude of an image. It's appealing for simplicity to calculate and high efficiency. Researchers have developed many variants such as global threshold, adaptive threshold and so on. However it's sensitive to noise and fails in complex images.

In recent years, some complex and robust algorithms have been proposed. The region-based model integrates features such as color, texture, shape, then segments pixels into different regions. Region growing [20] is representative. It initializes segmentation with a seed set, then merges similar pixels or regions step by step.

The graph-based model, e.g. [21], regards the image segmentation problem as a min-cut problem in mathematics. Pixels in an image are treated as nodes, and the weight on edges measures the dissimilarity between pixels. Furthermore, spectral graph theory is proposed to integrate global image information into the grouping process. Shi and Malik [22] propose the normalized cut criterion to "measure both the total dissimilarity between the different groups as well as the total similarity within the groups".

2.3. Existing Contour and Region Based Model for IQA

In some variants of SSIM, extra information of contour and region is utilized. Three components weighted SSIM [23] pools quality score by assigning different weights to different type (edge, texture, smooth area) of local region.

Similar idea is applied in video quality assessment. Pessoa *et al.* [24] introduce a region-based objective measurement for video quality assessment. They segment natural scenes into plane, edge and texture regions with different objective parameters. Then, a logistic regression is applied to approximate the relationship between the objective parameters and the subjective impairment. Telecommunications Research and Development Center proposes the CPqD-IES algorithm [25] which measures the objective impairment on plane, edge, and texture regions likewise.

Our model is inspired by Guo *et al.* [26]. They propose the primal sketch model, where they use sparse coding model and Markov random field model for representing geometric structure and stochastic texture respectively. They use primal sketch model to represent real images and provide a lossy image coding scheme.

3. Contour and Region Based Framework for IQA

Our model separates an image into structure part and local regions. Specially, we detect the contour for representing the structure, and use local descriptors for representing local region content. For each component, we quantify the degeneration by measuring the dissimilarity between the reference image and the distorted image.

3.1. The Representation of Image Contour

With the availability of public datasets which offer human-marked groundtruth contours, various approaches of contour detection are able to be compared with each other. To keep consistent with human perception as much as possible, we try several top-ranking algorithms on the Berkeley benchmarks [18] and finally select Ren's method [19] which has the highest score.

Built on the top of gPb [18], they replace hand-designed features with representation automatically learned through sparse coding. Orthogonal Matching Pursuit [27] and K-SVD [28] are taken to learn the dictionary and extract sparse codes at every pixel. They pool the sparse codes on the pixels over all scales for each orientation. Afterwards power transforms are applied before classifying them with a linear SVM.

The detected contour is represented as a matrix, where each entry denotes the probability of a pixel being the contour. A threshold is set to obtain the binary map of the contour. Figure 3.1 shows an example. (a) is the reference image "ocean" from LIVE2 database. (b) is the distorted image. (c,d) are the contours of (a,b) detected by Ren's method accordingly. (e,f) are the binary maps of the contours. The difference between the detected contours reveals the degradation of structure.

3.2. Measuring the Dissimilarity between Contours

To measure the dissimilarity between two contours, we choose shape context [29] as shape descriptor. The shape context captures the spatial distribution of the shape relative to the reference point on the contour. It's appealing for invariance to scaling, rotation, translation, without the requirement that the contour must be closed.

The processing pipeline includes the following steps:

First, for each contour, sample N points on the contour uniformly to obtain a sample set.

Second, for each point in the sample set, we can construct a vector set from this point to the other $N-1$ points. This vector set describes the spatial distribution of the shape relative to the point. To get a compact and discriminative descriptor, the histogram of the relative coordinates of the other $N-1$ points is computed.

Third, for all pairs of points p on the first contour and q on the second contour, calculate the matching cost $C_{p,q}$, which comprises of the shape context term C_S and local appearance term C_A by

$$C_S = \frac{1}{2}\sum_{k=1}^{K}\frac{[g(k)-h(k)]^2}{g(k)+h(k)}\ , \tag{3.1}$$

$$C_A = \frac{1}{2}\left\|\begin{pmatrix}\cos(\theta_p)\\ \sin(\theta_p)\end{pmatrix} - \begin{pmatrix}\cos(\theta_q)\\ \sin(\theta_q)\end{pmatrix}\right\|, \tag{3.2}$$

$$C_{p,q} = \alpha C_S + (1-\alpha)C_A, \tag{3.3}$$

where $g(k), h(k)$ denote the two K-bin (normalized) histograms, θ_p, θ_q denote the tangent orientations of two points on the images. C_S measures the difference on spatial distribution by χ^2 distance [30] and C_A measures the difference on local orientation. α balances their contribution.

Finally, given the set $\{C_{p,q}\}$ between two contours as the weights, solve the weighted bipartite graph matching problem to get optimal matching π^* and the minimum of the total cost SC as (3.4-3.5).

$$SC = \sum_i C_{i,\pi(i)}, \tag{3.4}$$

$$\pi^* = argmin_\pi SC, \tag{3.5}$$

where π is a permutation of $1, 2, \cdots N$.

In our experiment, to improve efficiency, we cut the binary map into blocks, calculate the matching cost between the binary maps block-by-block and use average pooling on non-zero entries. The dissimilarity score of contour DSC is defined as

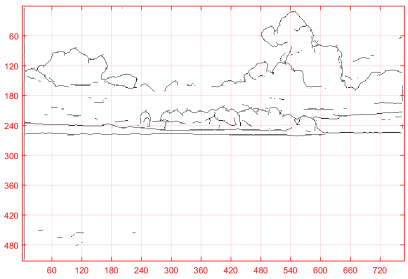
$$DSC = \frac{\sum_i SC_i}{\sum_i I_{SC_i>0}}, \tag{3.6}$$

(a)

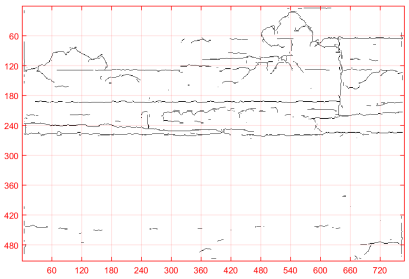
(b)

(c)

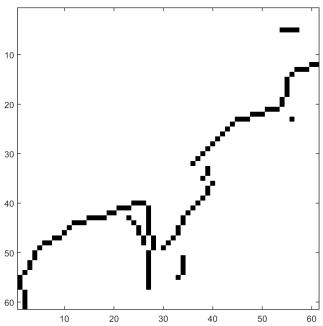
(d)


60
120
180
240
300
360
420
480
60 120 180 240 300 360 420 480 540 600 660 720

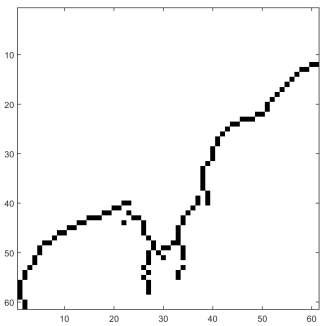
(e)


60
120
180
240
300
360
420
480
60 120 180 240 300 360 420 480 540 600 660 720

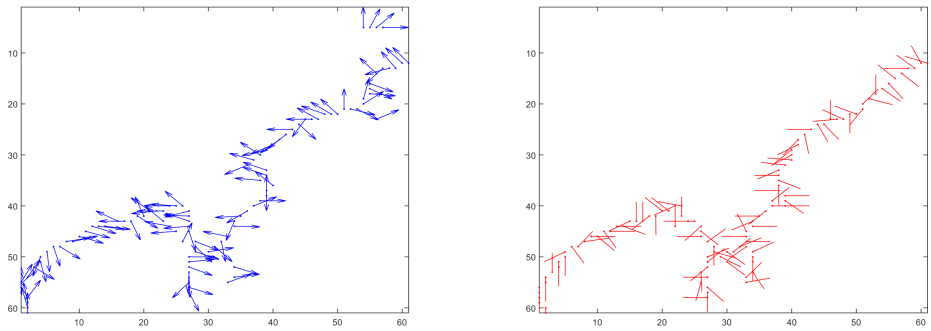
(f)


10
20
30
40
50
60
10 20 30 40 50 60

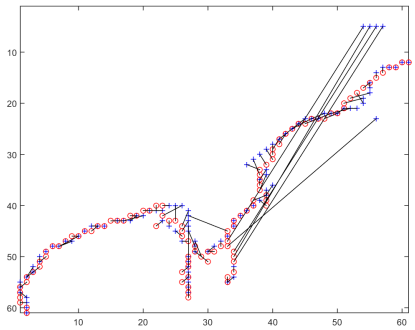
(g)


10
20
30
40
50
60
10 20 30 40 50 60

(h)



(i) (j)



(k)

Fig. 3.1. Calculate the dissimilarity between contours. (a) reference image. (b) distorted image. (c,d) contours detected by Ren's method. The value of each pixel predicts the probability of the boundary. (e,f) the binary maps of the contours by threshold 0.1. (g,h) the block of the first row of the ninth column of (e,f) accordingly. (i,j) the sample sets with the orientation. The arrow shows the tangent orientation of the point on the image. (k) the matching of two contours. Black line connecting two points reflects the correspondence.

where I is the indicator function whether SC_i is nonzero. Higher DSC means more severe degeneration in the view of the contour.

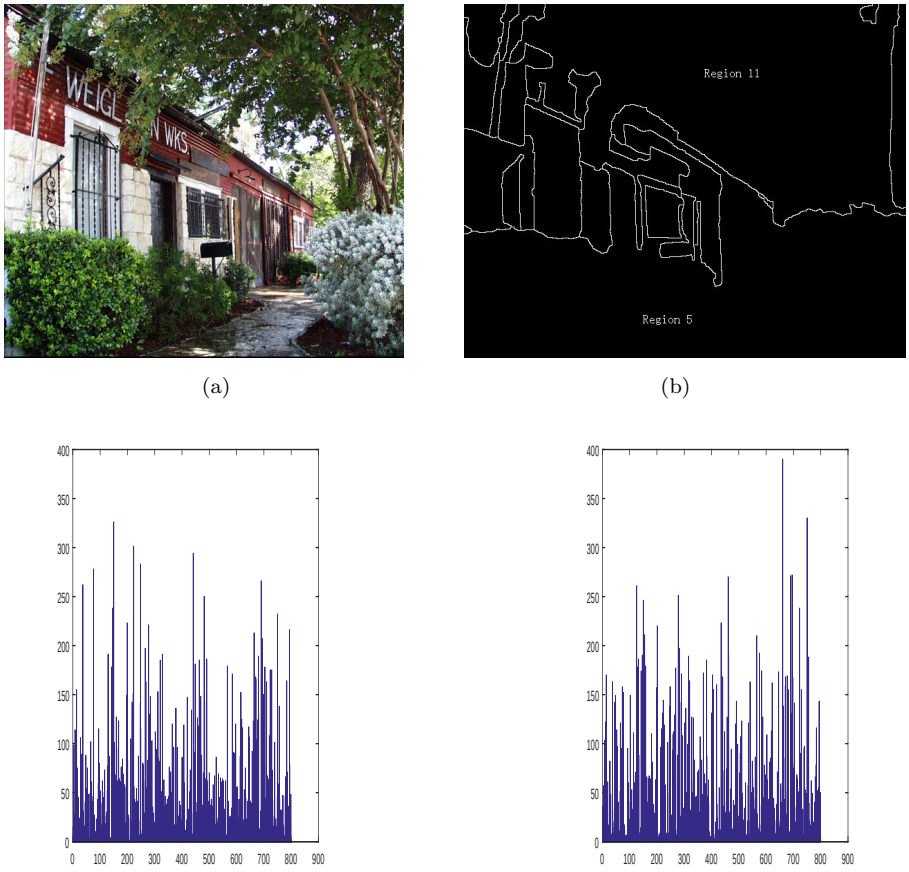
Figure 3.1 demonstrates the whole flow. (h) has similar structure with (g). In (k) most point pairs are well matched and there are few outliers. The match cost SC of this block is 0.18 and DSC of the distorted image is 0.55.

3.3. The Representation of Region Content

Image segmentation reduces the complexity for representation of images by clustering millions of pixels into hundreds of regions. Every region corresponds to the whole or parts of objects, provides various domains for details such as color, texture, and complements the representation with the contour.

We use the gPb-ucm [18] to segment images. This method integrates multiple local cues into a globalization framework by spectral clustering, and recovers hierarchical segmentations from a contour detector. More specifically, it carries the Oriented Watershed Transform to take the

contour signal to produce initial regions. Then an agglomerative clustering method proceeds to construct hierarchical segmentations from the boundaries of the initial regions. Figure 3.2 (b) is the segmentation of (a) which has been segmented into more than twenty regions.



(a) (b)

(c) (d)

Fig. 3.2. The representation of region content. (a) original image. (b) the segmentation. 5th and 11th regions are labeled. (c,d) the histograms of the 5th and 11th region. The x-axis represents for the index of visual words, and the y-axis represents for the number of the visual words.

Within each region, local descriptors are extracted and pooled for the representation of region content. Scale Invariant Feature Transform (SIFT) [31] is a popular method which extracts distinctive invariant features from images. In this paper, we use dense SIFT [32], a variant of SIFT as local descriptors. It should be pointed out that original SIFT and dense SIFT both operate on gray-scale images.

It's improper to make use of the full set of dense SIFT because it could lead to great cost in memory and time. We apply Bag Of Words (BOW) model [33] to produce compact representation. Dense SIFT is extracted from all images and clustered to form the visual dictionary. Within each region the dense SIFT is mapped into the words of the the visual dictionary and pooled to get a histogram. An example of such a statistical histogram computed on the single region is shown in Figure 3.2.

3.4. Measuring the Dissimilarity between Regions

The BOW model based on dense SIFT describes the region content as low dimension vector-valued attribute. The χ^2 distance [30] is calculated as the measurement of the dissimilarity between regions by

$$d(g,h) = \frac{1}{2}\sum_{k=1}^{K}\frac{[g(k)-h(k)]^2}{g(k)+h(k)}\ , \tag{3.7}$$

where $g(k), h(k)$ denote the two K-bin (normalized) histograms of visual words.

For the sake of simplicity, we weight every region according to their area in the whole image. In conclusion, the dissimilarity score of the region content DSR between images is the function of the dissimilarity of all regions as follow:

$$DSR = \sum_i w_i d_i, \tag{3.8}$$

where w_i is the proportion of i-th region in the image, d_i measures dissimilarity between the i-th region of the reference image and the corresponding region of the distorted image . Higher DSR means more severe degeneration in the view of the region content.

3.5. Integrated Model

Considering either the contour or the region content is not enough because the quality of images depends on both. Formally, the dissimilarity between contours (3.6) and the dissimilarity between region content (3.8) are integrated to get the image quality score:

$$Q = DSC^{\gamma} \cdot DSR^{(1-\gamma)}, \tag{3.9}$$

where $0 < \gamma < 1$, is a parameter to adjust the relative importance of both components. Lower Q corresponds to better quality.

4. Computational Algorithm

The primary algorithms are summarized in Algorithm 4.1 and 4.2. In Algorithm 4.1, it works well in the case that there are enough non-zero points in the block, however it doesn't work when the block of either the reference or the distorted image contains too few even no non-zero points. In this case, we calculate SC with the nonlinear function:

$$SC = \begin{cases} \frac{1}{1+\exp\left[-a(|n_1-n_2|-b)\right]} & |n_1-n_2| \geq c, \\ 0 & |n_1-n_2| < c, \end{cases} \tag{4.1}$$

where a, b are parameters and n_1, n_2 count the number of non-zero points in two blocks. If n_1 is closed to n_2, we get small SC.

Algorithm 4.1. Compute dissimilarity between contours

Input: (1) reference image x. (2) distorted image y. (3) threshold. (4) block size. (5) sample number.
Output: DSC

```
Detect the contour of x, y by Ren's method [19] and get the binary map x', y' according
 to the threshold.
Cut x', y' into blocks {x'_i}, {y'_i} without overlap respectively.
for each block pair x'_i, y'_i
   if the number of non-zero points in x'_i or y'_i is too small
      Handle exception.
   else
      Sample non-zero points of x'_i, y'_i to get point sets s(x'_i), s(y'_i) respectively.
      Compute the orientation of each point in point sets s(x'_i), s(y'_i).
      for all point pairs p ∈ s(x'_i), q ∈ s(y'_i)
          Compute the matching cost by (3.3)
      end for
      Solve the weighted bipartite graph matching problem by Hungarian method [34] to
      get the minimum of the total cost SC by (3.4).
   end if
end for
Compute DSC by (3.6)
```

Algorithm 4.2. Compute dissimilarity between regions

Input: (1) reference image x. (2) distorted image y. (3) image set I. (4) number of visual words.
Output: DSR

```
for each image I_i ∈ I do
   Extract dense SIFT on I_i
end for
 Run K-means on all dense SIFT to construct the visual dictionary D.
 Segment x into regions R by gPb-ucm [18].
 for each region R_i ∈ R
    Compute weight w_i of R_i.
    Compute the histogram g, h of dense SIFT in R_i for x, y.
    Compute the the dissimilarity d_i between regions by (3.7).
 end for
Compute DSR by (3.8).
```

In Algorithm 4.2, the visual dictionary D is constructed on all images including reference images and distorted images. Note that segmenting the reference image and the distorted image separately may lead to different segmentations. Under the assumption that the distortion gives rise to little shift on images, we just use the segmentation of the reference image as the one of the distorted image. To control the number of the segments, the adaptive threshold is used to segment an image to obtain about 20-60 regions.

5. Experiments

5.1. Benchmark Dataset

We evaluate our algorithm on two large-scale image datasets, LIVE2 and TID2013. In LIVE2, 29 reference images are distorted at different levels to generate 779 distorted images. The distortion types include JPEG2000 (JP2K), JPEG, White noise in the RGB components (WN), Gaussian blur in the RGB components (GB), and bit errors in JPEG2000 bitstream when transmitted over a simulated fast-fading Rayleigh channel (FF). For each distorted image, more than 20 subjects judgment the quality to obtain the convincing scores. The TID2013 database consists of 24 types of distortion, up to 3000 distorted images as the largest database in IQA community. The information about the datasets is summarized in Table 5.1.

To evaluate the performance of various models, four common metrics are employed on the datasets including Spearman Rank-Order Correlation Coefficient (SROCC), Kendall Rank-Order Correlation Coefficient (KROCC), Pearson Linear Correlation Coefficient (PLCC) [35] , and Root Mean Squared Error (RMSE). SROCC and KROCC measure the monotonicity of prediction, which focus on the rank rather than the error of the prediction. The other two metrics are computed after applying non-linear regression between the objective scores and the subjective Mean Opinion Scores (MOS). PLCC measures the linear correlation and RMSE measures the error of the prediction. Specially, we choose the 5-parameter logistic function as suggested by [12] for regression:

$$Quality(o) = \beta_1 \left(\frac{1}{2} - \frac{1}{1 + e^{\beta_2 (o - \beta_3)}} \right) + \beta_4 o + \beta_5, \tag{5.1}$$

where o is the objective score, $\beta_i, i = 1, 2, \cdots, 5$ are parameters.

Table 5.1: Benchmark datasets.

Dataset	Reference Images No.	Distorted Images No.	Distortion Types No.	Observers No.
LIVE2	29	779	5	161
TID2013	25	3000	24	971

5.2. Parameter Setting

We fix the parameters for each experiment as follow: In Algorithm 4.1, we set threshold as 0.1, block size as 60×60, sample number as 100, and a, b, c in (4.1) as 0.05, 30, 10; In Algorithm 4.2, the number of visual words is 1200 in LIVE2 and 2000 in TID2013; γ is set as 0.7 in the integrated model.

In the following experiments, our model is compared with multiple state-of-the-art models such as MS-SSIM [6], IFC [7], VIF [8], VSNR [4], MAD [36], IW-SSIM [9], RFSIM [37], FSIM_c [10], VSI [11].

5.3. Experiment on LIVE2

The SROCC for each type of distortion on LIVE2 is demonstrated in the upper part of Table 5.2. The highest value for each distortion type is highlighted in boldface. From Table 5.2, we could make the conclusion that our model performs well on most distortion types on LIVE2. There is a slight gap between our model and the best result on types such as JP2K,

JPEG, FF. And VIF almost performs best on individual type on LIVE2. Similar conclusion is easy to be drawn with the other three metrics.

The performance comparison with the other models on the mixture of all distortion types on LIVE2 is listed in Table 5.3. As opposed to the case of individual distortion, our model performs rather worse than VIF in the mixture case. Further researches are conducted to demonstrate the comprehensive comparison between our model and VIF, and the details are shown in Table 5.4 and Figure 5.1. The better results are underlined. Figure 5.1 plots the scatter diagram for each distortion type. Our model performs consistently on individual distortion type, but data points of WN are not well aligned with other four types in mixture case. Furthermore, we calculate SROCC on the mixture of the other four types, and it rises greatly to 0.948 from 0.907.

The sensitivity to noise in contour detection is the leading cause. It is worth mentioning that Ren's method is trained on natural images without distortion, so it's sensitive to noise. Figure 5.2 demonstrates an example comparing the distorted image of WN with the distorted image of JP2K. The detailed information of two distorted images is listed in Table 5.5. It can be seen that Figure 5.2 (f) keeps less structure than (h) in the stage of contour detection. Some blocks in (f) even contain no contour and this is inconsistent with our perception. Among distorted images with the approximate DMOS, images of WN prefer to suffer more severe degradation on the contour in our model because the method of contour detection is sensitive to noise.

Table 5.2: SROCC for each distortion type.

Dataset	Type	MS-SSIM	IFC	VIF	VSNR	MAD	IW-SSIM	RFSIM	$FSIM_c$	VSI	Ours
LIVE2	JP2K	0.963	0.911	0.970	0.955	0.968	0.965	0.932	**0.972**	0.960	0.961
	JPEG	0.982	0.947	**0.985**	0.966	0.976	0.981	0.958	0.984	0.976	0.975
	WN	0.973	0.938	**0.986**	0.979	0.984	0.967	0.980	0.972	0.984	0.953
	GB	0.954	0.958	**0.973**	0.941	0.947	0.972	0.907	0.971	0.953	0.948
	FF	0.947	0.963	**0.965**	0.903	0.957	0.944	0.924	0.952	0.943	0.964
TID2013	AGN	0.865	0.661	0.899	0.827	0.884	0.844	0.888	0.910	**0.946**	0.754
	ANC	0.773	0.535	0.830	0.731	0.802	0.752	0.848	0.854	**0.871**	0.730
	SCN	0.854	0.660	0.884	0.801	0.891	0.817	0.883	0.890	**0.937**	0.749
	MN	0.807	0.693	**0.845**	0.707	0.738	0.802	0.837	0.809	0.770	0.791
	HFN	0.860	0.741	0.897	0.846	0.888	0.855	0.915	0.904	**0.920**	0.826
	IN	0.763	0.641	0.854	0.736	0.277	0.728	**0.906**	0.825	0.874	0.841
	QN	0.871	0.628	0.785	0.836	0.851	0.867	**0.897**	0.881	0.875	0.762
	GB	**0.967**	0.891	0.965	0.947	0.932	0.970	0.970	0.955	0.961	0.950
	DEN	0.927	0.778	0.891	0.908	0.925	0.915	0.936	0.933	**0.948**	0.877
	JPEG	0.927	0.836	0.919	0.901	0.922	0.919	0.940	0.934	**0.954**	0.921
	JP2K	0.950	0.908	0.952	0.927	0.951	0.951	0.952	0.959	**0.971**	0.941
	JGTE	0.848	0.743	0.841	0.791	0.828	0.839	0.831	0.861	**0.922**	0.884
	J2TE	0.889	0.777	0.876	0.841	0.879	0.866	0.906	0.892	**0.923**	0.848
	NEPN	0.797	0.574	0.772	0.665	**0.832**	0.801	0.771	0.794	0.806	0.815
	Block	0.480	0.241	0.531	0.177	0.281	0.372	0.034	**0.553**	0.171	0.495
	MS	**0.791**	0.552	0.628	0.487	0.645	0.783	0.555	0.749	0.770	0.698
	CTC	0.463	0.180	**0.839**	0.332	0.197	0.459	0.399	0.468	0.475	0.256
	CCS	0.410	0.403	0.310	0.368	0.058	0.420	0.020	**0.836**	0.810	0.486
	MGN	0.779	0.614	0.847	0.764	0.841	0.773	0.846	0.857	**0.912**	0.718
	CN	0.853	0.816	0.895	0.868	0.906	0.876	0.892	0.914	**0.924**	0.890
	LCNI	0.907	0.818	0.920	0.882	0.944	0.904	0.901	0.947	**0.956**	0.890
	ICQD	0.856	0.601	0.841	0.867	0.875	0.840	**0.896**	0.882	0.884	0.828
	CHA	0.878	0.821	0.885	0.865	0.831	0.868	**0.899**	0.892	0.891	0.897
	SSR	0.948	0.889	0.935	0.934	0.957	0.947	0.933	0.958	**0.963**	0.928

Table 5.3: Comparision of different models on LIVE2.

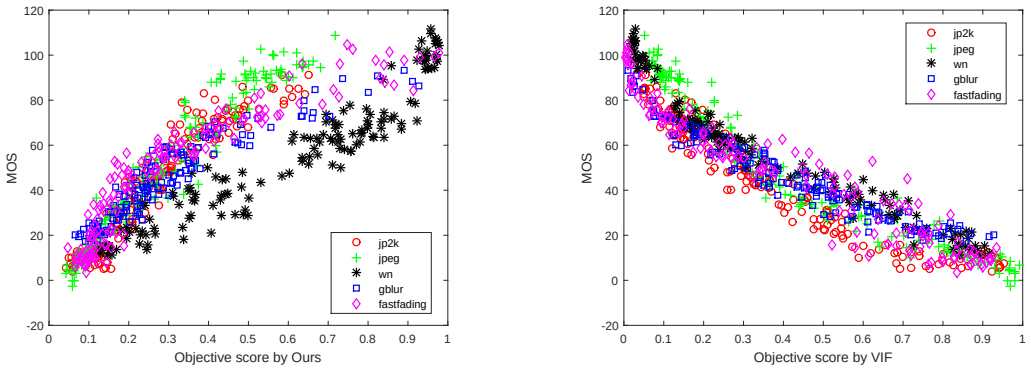
	MS-SSIM	IFC	VIF	VSNR	MAD	IW-SSIM	RFSIM	$FSIM_c$	VSI	Ours
SROCC	0.951	0.926	0.964	0.927	**0.967**	0.957	0.940	0.965	0.952	0.907
KROCC	0.805	0.758	0.828	0.762	**0.842**	0.818	0.782	0.836	0.806	0.737
PLCC	0.949	0.927	0.960	0.923	**0.968**	0.952	0.935	0.961	0.948	0.853
RMSE	8.619	10.264	7.614	10.506	**6.907**	8.347	9.664	7.530	8.682	14.240

Table 5.4: Comparison between our model and VIF on LIVE2.

VIF/Ours	JP2K	JPEG	WN	GB	FF	All Types
SROCC	0.970/0.961	0.985/0.975	0.986/0.953	0.973/0.948	0.965/0.964	0.964/0.907
KROCC	0.847/0.825	0.894/0.864	0.898/0.809	0.859/0.802	0.840/0.835	0.828/0.737
PLCC	0.948/0.961	0.987/0.968	0.988/0.969	0.975/0.954	0.970/0.964	0.960/0.853
RMSE	8.060/6.950	5.066/8.012	4.274/6.933	4.146/5.514	6.974/7.516	7.614/14.240

Table 5.5: Comparison of WN and JP2K.

Distortion Type	Image	Ref Image	DSC	DSR	Q	DMOS
WN	img74	rapids	0.78	0.39	0.64	54.63
JP2K	img138	sailing2	0.43	0.14	0.31	56.33



(a) (b)

Fig. 5.1. Scatter plot of DMOS against objective scores predicted by models on LIVE2. (a) our model. (b) VIF. The x-axis indicates the object scores by models, and the y-axis indicates the Difference Mean Opinion Score (DMOS) [12]. Each color represents one distortion type. Lower DMOS value corresponds to higher quality. Note that lower VIF index corresponds to higher quality and object score predicted by our model does conversely.

5.4. Experiment on TID2013

Our next experiment evaluates our model on TID2013 database. The TID2013 database consists more complex distortion types and some types such as ANC, CCS, ICQD are color-dependent distortion. The SROCC for each type is summarized in Table 5.2. Our model performs consistently on parts of all types, even only slightly worse than the best results in GB, NEPN, Block, and CHA.

Table 5.6: SROCC for all components on TID2013.

Dataset	Type	contour	region content	integrated model
LIVE2	JP2K	0.901	0.940	0.961
	JPEG	0.921	0.968	0.975
	WN	0.936	0.870	0.953
	GB	0.903	0.945	0.948
	FF	0.911	0.954	0.964
TID2013	AGN	0.633	0.599	0.754
	ANC	0.567	0.678	0.730
	SCN	0.692	0.563	0.749
	MN	0.566	0.770	0.791
	HFN	0.801	0.713	0.826
	IN	0.867	0.686	0.841
	QN	0.640	0.732	0.762
	GB	932	0.923	0.950
	DEN	0.867	0.820	0.877
	JPEG	0.875	0.907	0.921
	JP2K	0.883	0.916	0.941
	JGTE	0.844	0.858	0.884
	J2TE	0.761	0.776	0.848
	NEPN	0.776	0.783	0.815
	Block	0.331	0.628	0.495
	MS	725	0.614	0.698
	CTC	0.234	0.202	0.256
	CCS	0.610	0.250	0.486
	MGN	0.532	0.515	0.718
	CN	0.734	0.858	0.890
	LCNI	0.793	0.828	0.890
	ICQD	0.747	0.695	0.828
	CHA	0.815	0.876	0.897
	SSR	0.865	0.906	0.928

To analyse the contribution of the contour and region content, we separate these two components and compare them with the integrated model. The result is shown in Table 5.6. The integrated model almost outperforms the single feature in all types except IN, MS, CCS. It's confirmed to obtain performance gains by combining contour and region content.

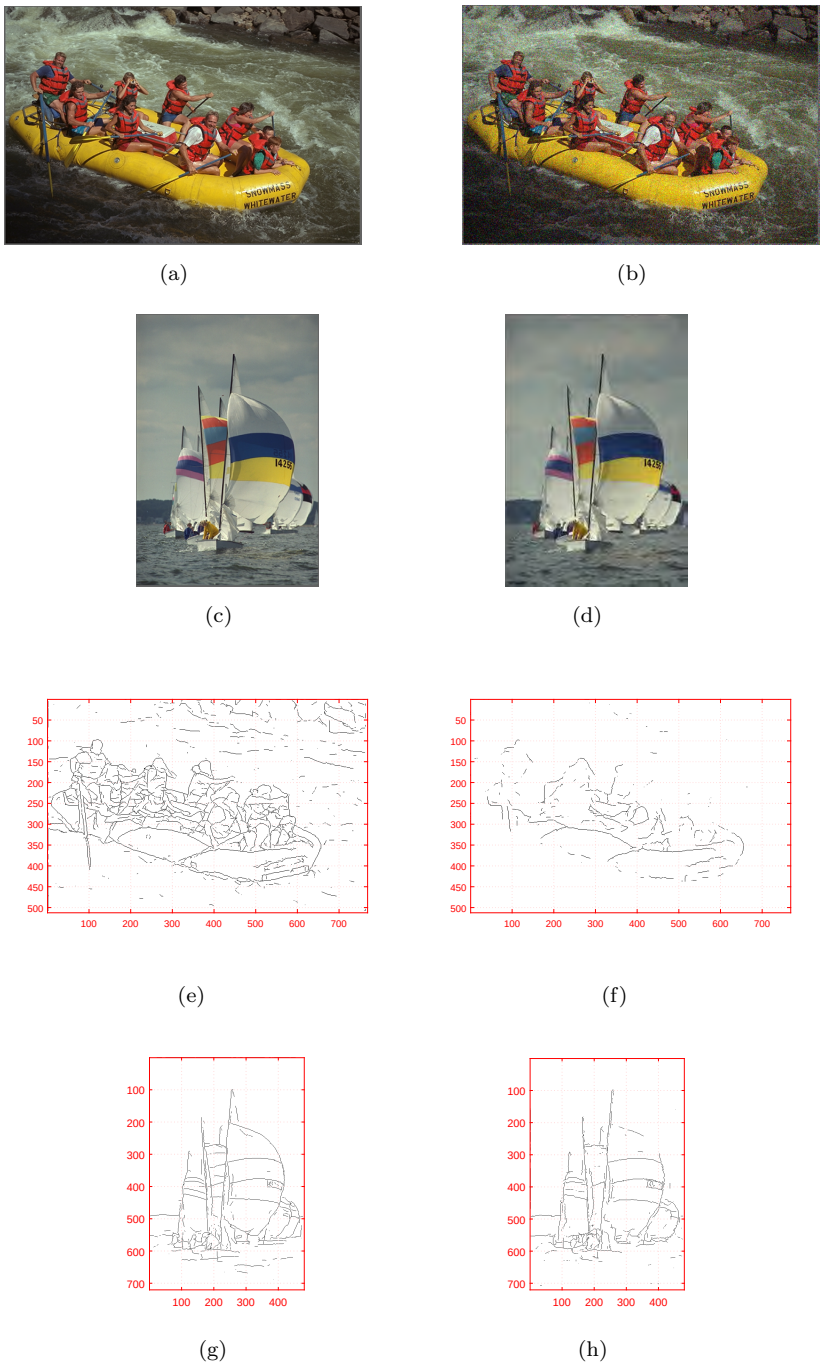


Fig. 5.2. WN vs JP2K. (a) reference image, “rapids”. (b) distorted image, “img74”. (c) reference image, “sailing2”. (d) distorted image, “img138”. (e,g) the binary map of (a,c). (f,h) the binary map of (b,d). (b) obtains approximate DMOS to (d) but (f) keeps less contour than (h) after being distorted.

6. Conclusion and Discussion

In this paper, we propose a novel framework combining the contour and region content to assess image quality. Comparing to weighting pixels by information content (e.g. [23]) in low level, we consider the contour as a whole and try to quantify the degeneration of contour between the reference and the distorted image. Then clustering technology applied in regions offers rich representation to measure the degeneration after distortion. In our experiments we have demonstrated contour information and region content complement with each other to boost performance. The framework opens up possibility for IQA in middle level.

There are still many problems remained to be explored in our framework, and our future work focuses on the following aspects:

- Contour detection. Most approaches of contour detection are trained on natural images without distortion, so they fail to extract contour in line with human perception in some distortion type (e.g. WN). More work should be done to push contour detection closer to human perception.

- Region assignment. Distortion may lead to different segmentations from the reference. In this paper, we approximate the segmentation of the distorted image with the one of the reference image and get satisfactory result in some cases. However whether the approximation works on images of worse quality requires a further research.

- Local descriptors. In our experiment, we extract dense SIFT on gray-scale images. One can replace dense SIFT with other sophisticated descriptors which contain color information.

- Multiscale analysis. Image quality greatly depends on the scale so it's worth exploring the framework in multiscale.

- Weight. In Algorithm 4.2, we weight each region by its area. This pooling strategy is convenient for computing but too simple for visual perception. Next, we intend to introduce visual saliency in the pooling stage.

- Integrated model. The relation between two components and image quality is complicated. The weighted geometric mean may be too simple to model their relation. The learning approaches such as SVM are going to be tested.

- Axiomatic characterization. Axiomatic approaches as in [38-40] can be integrated into the current framework. The results in [39] can be used to improve the integrated model. The approaches in [40] can be adapted to evaluate image sharpness in addition to the current two evaluations. Nevertheless, it will be a challenging problem to establish an axiomatic IQA theory.

Acknowledgments. This work is partially supported by the National Basic Research Program of China (973 Program) (2015CB351803), National Science Foundation of China (61421062, 61520106004, 61572042, 61390514, 61210005, 61527084), and Sino-German Center (GZ 1025).

References

[1] Z. Wang and A.C. Bovik, Modern image quality assessment, *Synthesis Lectures on Image, Video, and Multimedia Processing*, **2**:1 (2006), 1–156.

[2] J. Lubin, Digital images and human vision, chapter The Use of Psychophysical Data and Models in the Analysis of Display System Performance, pages 163–178, MIT Press, Cambridge, MA, USA, 1993.

[3] N. Damera-Venkata, T.D. Kite, W.S. Geisler, B.L. Evans and A.C. Bovik, Image quality assessment based on a degradation model, *IEEE Transactions on Image Processing*, **9**:4 (2000), 636–650.

[4] D.M. Chandler and S.S. Hemami, VSNR: A wavelet-based visual signal-to-noise ratio for natural images, *IEEE Transactions on Image Processing*, **16**:9 (2007), 2284–2298.

[5] Z. Wang, A.C. 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.

[6] Z. Wang, E.P. Simoncelli and A.C. Bovik, Multiscale structural similarity for image quality assessment, Proceedings of the 37th IEEE Asilomar Conference on Signals, Systems and Computers, volume 2, pages 1398–1402, IEEE, 2003.

[7] H.R. Sheikh, A.C. Bovik and G. De Veciana, An information fidelity criterion for image quality assessment using natural scene statistics, *IEEE Transactions on Image Processing*, **14**:12 (2005), 2117–2128.

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

[9] Z. Wang and Q. Li, Information content weighting for perceptual image quality assessment, *IEEE Transactions on Image Processing*, **20**:5 (2011), 1185–1198.

[10] L. Zhang, L. Zhang, X. Mou and D. Zhang, FSIM: a feature similarity index for image quality assessment, *IEEE Transactions on Image Processing*, **20**:8 (2011), 2378–2386.

[11] L. Zhang, Y. Shen and H. Li, VSI: a visual saliency-induced index for perceptual image quality assessment, *IEEE Transactions on Image Processing*, **23**:10 (2014), 4270–4281.

[12] H.R. Sheikh, M.F. Sabir and A.C. Bovik, A statistical evaluation of recent full reference image quality assessment algorithms, *IEEE Transactions on Image Processing*, **15**:11 (2006), 3440–3451.

[13] N. Ponomarenko, O. Ieremeiev, V. Lukin, K. Egiazarian, L. Jin, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti and C.C.J. Kuo, Color image database tid2013: Peculiarities and preliminary results, Proceedings of the 4th European Workshop on Visual Information Processing, pages 106–111, June 2013.

[14] R.C. Gonzalez, Digital image processing, Pearson Education India, 2009.

[15] M.C. Morrone and R.A. Owens, Feature detection from local energy, *Pattern Recognition Letters*, **6**:5 (1987), 303–313.

[16] D.R. Martin, C.C. Fowlkes and J. Malik, Learning to detect natural image boundaries using local brightness, color, and texture cues, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **26**:5 (2004), 530–549.

[17] X. Ren, Multi-scale improves boundary detection in natural images, European Conference on Computer Vision, pages 533–545, Springer, 2008.

[18] P. Arbelaez, M. Maire, C. Fowlkes and J. Malik, Contour detection and hierarchical image segmentation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **33**:5 (2011), 898–916.

[19] X. Ren and L. Bo, Discriminatively trained sparse code gradients for contour detection, Advances in Neural Information Processing Systems, pages 584–592, 2012.

[20] N. Yokoya and M.D. Levine, Range image segmentation based on differential geometry: a hybrid approach, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **11**:6 (1989), 643–649.

[21] P.F. Felzenszwalb and D.P. Huttenlocher, Efficient graph-based image segmentation, *International Journal of Computer Vision*, **59**:2 (2004), 167–181.

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

[23] C. Li and A.C. Bovik, Three-component weighted structural similarity index, IS&T/SPIE Electronic Imaging, pages 72420Q–72420Q, International Society for Optics and Photonics, 2009.

[24] A. Pessoa, A. Falcao, R. Nishihara, A. Silva and R. Lotufo, Video quality assessment using objective parameters based on image segmentation, *Society of Motion Picture and Television Engineers*, **108**:12 (1999), 865–872.

[25] Series J: Cable Networks and Transmission of Television, Sound Programme and Other Multimedia Signals, Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference? ITU-T Rec, 2004.

[26] C.E. Guo, S.C. Zhu and Y.N. Wu, Primal sketch: Integrating structure and texture, *Computer Vision and Image Understanding*, **106**:1 (2007), 5–19.

[27] Y.C. Pati, R. Rezaiifar and P. Krishnaprasad, Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition, Signals, Systems and Computers, 1993. Conference Record of The Twenty-Seventh Asilomar Conference on, pages 40–44, IEEE, 1993.

[28] M. Aharon, M. Elad and A. Bruckstein, K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation, *IEEE Transactions on Image Processing*, **54**:11 (2006), 4311.

[29] S. Belongie, G. Mori and J. Malik, Matching with shape contexts, Statistics and Analysis of Shapes, pages 81–105, Springer, 2006.

[30] P.E. Greenwood and M.S. Nikulin, A Guide to Chi-Squared Testing, volume 280, John Wiley & Sons, 1996.

[31] D.G. Lowe, Distinctive image features from scale-invariant keypoints, *International Journal of Computer Vision*, **60**:2 (2004), 91–110.

[32] A. Bosch, A. Zisserman and X. Munoz, Image classification using random forests and ferns, 2007 IEEE 11th International Conference on Computer Vision, pages 1–8, IEEE, 2007.

[33] J. Sivic and A. Zisserman, Efficient visual search of videos cast as text retrieval, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **31**:4 (2009), 591–606.

[34] C.H. Papadimitriou and K. Steiglitz, Combinatorial optimization: algorithms and complexity, Courier Corporation, 1998.

[35] J.D. Gibbons and S. Chakraborti, Nonparametric statistical inference, Springer, 2011.

[36] E.C. Larson and D.M. Chandler, Most apparent distortion: full-reference image quality assessment and the role of strategy, *Journal of Electronic Imaging*, **19**:1 (2010), 011006–011006.

[37] L. Zhang, L. Zhang and X. Mou, RFSIM: A feature based image quality assessment metric using Riesz transforms, 2010 IEEE International Conference on Image Processing, pages 321–324, IEEE, 2010.

[38] G. Wang and M. Jiang, Axiomatic characterization of nonlinear homomorphic means, *Journal of Mathematical Analysis and Applications*, **303**:1 (2005), 350 – 363.

[39] M. Jiang, G. Wang and X.M. Ma, A general axiomatic system for image resolution quantification, *Journal of Mathematical Analysis and Applications*, **315**:2 (2006), 462 – 473.

[40] J.A. Osullivan, M. Jiang, X.M. Ma and G. Wang, Axiomatic quantification of multidimensional image resolution, *IEEE Signal Processing Letters*, **9**:4 (2002), 120–122.