Signal Processing: Image Communication 61 (2018) 54–72

Contents lists available at ScienceDirect

Signal Processing: Image Communication

journal homepage: www.elsevier.com/locate/image

Perceptual quality evaluation of synthetic pictures distorted by compression and transmission

Debarati Kundu *,[1], Lark Kwon Choi [1], Alan C. Bovik [1], Brian L. Evans [1]

Department of Electrical and Computer Engineering, Wireless Networking and Communications Group, The University of Texas at Austin, TX, United States

A R T I C L E I N F O

Keywords:
Image quality assessment
Natural scene statistics
Subjective study
Synthetic scenes

A B S T R A C T

Measuring visual quality, as perceived by human observers, is becoming increasingly important in a large number of applications where humans are the ultimate consumers of visual information. Many natural image databases have been developed that contain human subjective ratings of the images. Subjective quality evaluation data is less available for synthetic images, such as those commonly encountered in graphics novels, online games or internet ads. A wide variety of powerful full-reference, reduced-reference and no-reference Image Quality Assessment (IQA) algorithms have been proposed for natural images, but their performance has not been evaluated on synthetic images. In this paper we (1) conduct a series of subjective tests on a new publicly available Embedded Signal Processing Laboratory (ESPL) Synthetic Image Database, which contains 500 distorted images (20 distorted images for each of the 25 original images) in 1920 × 1080 resolution, and (2) evaluate the performance of more than 50 publicly available IQA algorithms on the new database. The synthetic images in the database were processed by post acquisition distortions, including those arising from compression and transmission. We collected 26,000 individual ratings from 64 human subjects which can be used to evaluate full-reference, reduced-reference, and no-reference IQA algorithm performance. We find that IQA models based on scene statistics models can successfully predict the perceptual quality of synthetic scenes. The database is available at: http://signal.ece.utexas.edu/%7Ebevans/synthetic/.

1. Introduction

Recent years have seen tremendous growth in the acquisition, transmission, and storage of both natural and synthetic digital pictures [3]. In addition to pictures captured by optical cameras, picture traffic also often includes synthetic scenes, such as animations, cartoons, comics, games, and internet ads. In all these cases, humans are the final consumers of the visual data and the ultimate goal is to provide a satisfactory quality-of-experience (QoE) [4]. The visual quality of synthetic scenes can be degraded both by the rendering process (e.g. video gaming on standalone devices) and by post acquisition processes such as wireless transmission. Methods of evaluating visual quality play important roles in perceptually optimized design of display devices, rendering engines, and compression standards as well as for maintaining a satisfactory QoE in streaming applications under challenging network conditions.

Although a subjective study with human observers is the most reliable way to gauge perceptual quality of pictures, human studies are time consuming and rarely feasible. The ground-truth data obtained from human observers can be used to benchmark objective IQA algorithms that aim to automate the process of visual quality assessment. Some of the largest natural image databases are the LIVE Image Quality Database [5], the Tampere Image Database 2013 [6], LIVE Challenge Database [7], the Categorical Image Quality Database [8] and EPFL JPEG XR codec [9]. Recently Cadík et al. [10] developed a synthetic image database of computer graphics generated imagery afflicted by distortions such as noise, aliasing, changes in brightness, light leakage and tone mapping artifacts. Some IQA approaches for synthetic images have also been proposed [11,12].

To automate perceptual quality evaluation, two broad categories of objective IQA algorithms are available: reference and blind or no-reference methods, based on the availability (or not) of a reference image. Reference methods may have access to either the complete reference image or some statistical features extracted from it. The

* Corresponding author.
E-mail address: debarati@utexas.edu (D. Kundu).

[1] This paper is an expanded version of [1] and [2]. D. Kundu and L.K. Choi conducted this research as Ph.D. students at UT Austin and now both of them work at Qualcomm. B.L. Evans is also with the Embedded Signal Processing Laboratory (ESPL) at UT Austin. A.C. Bovik is also with the Laboratory for Image and Video Engineering (LIVE) at UT Austin.

https://doi.org/10.1016/j.image.2017.11.004
Received 2 May 2017; Received in revised form 19 November 2017; Accepted 19 November 2017
Available online 23 November 2017
0923-5965/

former defines full-reference (FR) IQA algorithms, while the latter defines reduced-reference (RR) IQA algorithms. The performance of several publicly available state-of-the-art FR-IQA algorithms has been evaluated on popular natural image databases [13–15]. Cadík et al. [10] evaluated the performance of six FR-IQA algorithms and demonstrated that they were sensitive to brightness and contrast changes, could not distinguish between plausible and implausible shading, and failed to localize distortions precisely.

When information about the reference image is not available, no-reference (NR) IQA models are more suitable. Many NR metrics rely on machine learning approaches using features expressive of statistical regularities possessed by pristine images, commonly called natural scene statistics (NSS) models [16,17]. NSS models of good quality natural images hold reliably well irrespective of image content. In NR-IQA model design, it is often assumed that distortions tend to deviate from these statistical regularities. NR-IQA algorithms have not yet been studied in the context of images generated using computer graphics. Herzog et al. [18] propose an NR-IQA metric for quantifying rendering distortions based on machine learning. Their features were chosen heuristically, instead of being based on properties of pristine synthetic images.

With the advent of more powerful Graphics Processing Units, the degree of realism of graphical images [19] has vastly narrowed between natural scenes and high quality synthetic scenes. This does not imply that they share identical statistical properties. In our earlier work [20], we created a database of photorealistic synthetic images and modeled the distribution of mean-subtracted-contrast-normalized (MSCN) pixels [21] obtained from the image intensities using Generalized Gaussian and Symmetric α-Stable distributions. Irrespective of the content, we discovered that the scene statistics of the photorealistic graphics images show substantial similarity to those of natural images.

Here we present the results of a series of subjective tests conducted on distorted synthetic images [2]. The study included 25 high definition reference images, from which 500 images were created by adding controlled amounts of different levels of five commonly encountered artifacts: interpolation, blur and additive noise (processing artifacts), JPEG blocking (compression artifact) and fast-fading effects (transmission artifacts). Every image was evaluated by 64 observers under controlled laboratory conditions in a single stimulus experiment, where the observers rated visual quality on a continuous quality scale. Differential Mean Opinion Scores (DMOS) are supplied which augment the ESPL Synthetic Image Database [20] containing unannotated pristine and distorted images.

When creating the new database, we considered processing, transmission and compression artifacts on synthetic images that have not been considered in any previous subjective study to the best of our knowledge. In [10,18], Cadík et al. mainly focused on computer graphics generated artifacts. However, with the advent of mobile cloud gaming and animations, compression and transmission artifacts commonly occur on synthetic scenes in addition to processing and rendering artifacts.

We also evaluate the performance of more than 50 state-of-the-art FR, RR and NR IQA algorithms on the synthetic scenes and compare them to the subjective test results. The performances of the algorithms are compared and the leading models are subjected to a statistical significance analysis. We hypothesize that with some modifications, NSS based NR-IQA metrics could be successfully applied to photorealistic graphics images. Here we take a first step towards evaluating scene statistics based NR-IQA methods on synthetic scenes, expressed both in the spatial as well as transform domains. Top performing NSS-based NR-IQA algorithms show a high degree of correlation with human perception of distorted image quality on synthetic scenes, which is a promising development in regards to the successful automatic prediction of the perceptual quality of computer graphics generated imagery for which no 'ground truth' information is available.

The remainder of this paper is organized as follows. Section 2 describes the subjective study: methods employed in generating the synthetic scenes and the subjective testing framework. Section 3 outlines the statistics of the pristine and distorted synthetic scenes considered in the database. The quantitative performances of many objective IQA performance on the ESPL Synthetic Image Database are detailed in Section 4, and the results are discussed in Section 5. Section 6 concludes the paper.

2. Human subjective study

2.1. The image database

2.1.1. Source images

A total of 25 synthetic images were chosen. These high quality color images from the Internet are 1920×1080 pixels in size. Some of the images are from multiplayer role playing games (such as War of Warcraft), first person shooter games (such as Counter Strike), motorcycle and car racing games, and games with more realistic content (such as FIFA). Single frames were also collected from animated movies: The Lion King, the Tinkerbell series, Avatar, Beauty and the Beast, Monster series, Ratatouille, the Cars series, etc.[1] We incorporated natural and non-photorealistic renderings of human figures and human-made objects, renderings of fantasy figures such as fairies and monsters, close-up shots, wide angle shots, images containing both high and low degrees of color saturation, and background textures without a foreground object.

Fig. 1 shows the 25 reference images. The source complexities of the images were analyzed using the quantitative metrics of scene complexity and colorfulness in [22]. Fig. 2 shows a scatter plot of spatial information vs. colorfulness computed on the images in the ESPL Synthetic Image Database and three other publicly available image quality assessment databases (Cadík's [10], LIVE [5] and TID [6] databases). The scatter plots from the ESPL database, shown in Fig. 2(a) indicate that spatial information and colorfulness in ESPL span a similar range of scene complexity as in the other natural image databases as shown in Fig. 2(c) and (d). In Fig. 2(b), Cadík's Synthetic Image database shows a larger range but sparsely covers the range.

2.1.2. Distortion simulations

Distortions of synthetic images are often more varied than those affecting natural images. This is because distortions of synthetic images arise from two sources: firstly, the image might have artifacts from the rendering process, display and other processing steps, such as tone mapping and contrast amplification, and secondly, distortions might be introduced due to encoding at a low bit-rate or transmission over a network, such as JPEG block artifacts and transmission noise. Other distortions may arise, such as unnaturalness of shading, which can be evaluated only given access to both the rendered 2D scene, and the information provided by the 3D depth buffer. The ESPL database does not contain these other kinds of distortions, focusing instead on transmission and compression artifacts. Since we did not have access to the proprietary 3D models and the lighting information that were used to render the scenes, we introduced distortions on the rendered image themselves.

Three categories of processing artifacts are considered: interpolation (which arises frequently in texture maps, and causes jaggedness of crisp edges), blurring and additive Gaussian noise. Blur artifacts often appear in synthetic images when simulating objects in high motion or depth-of-field. Depth-of-field blur can be synthesized by placing sharper foreground objects on a uniformly blurred background. Evaluation of images with blur is an important component of image quality prediction. Although blur in computer graphics generated imagery is often intentionally introduced, it is not always, and may arise from other sources, such as a loss of resolution during transmission or inadequate resolution

[2] All images are copyright of their rightful owners, and the authors do not claim ownership. No copyright infringement is intended. The database is to be used strictly for non-profit educational purposes.