Pattern Recognition 76 (2018) 242–255

Contents lists available at ScienceDirect

Pattern Recognition

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

Learning a referenceless stereopair quality engine with deep nonnegativity constrained sparse autoencoder

Qiuping Jiang[a,b], Feng Shao[a,*], Weisi Lin[b], Gangyi Jiang[a]

[a] Faculty of Information Science and Engineering, Ningbo University, Ningbo 315211, China
[b] School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798, Singapore

ARTICLE INFO

Article history:
Received 15 March 2017
Revised 2 October 2017
Accepted 4 November 2017
Available online 7 November 2017

Keywords:
Image quality assessment
No-reference/referenceless
Stereoscopic 3D image
Deep learning
Nonnegativity constrained
Sparse autoencoder

ABSTRACT

This paper proposes a no-reference (NR)/referenceless quality evaluation method for stereoscopic three-dimensional (S3D) images based on deep nonnegativity constrained sparse autoencoder (DNCSAE). To address the quality issue of stereopairs whose perceived quality is not only determined by the individual left and right image qualities but also their interactions, a three-column DNCSAE framework is customized with individual DNCSAE module coping with the left image, the right image, and the cyclopean image, respectively. In the proposed framework, each individual DNCSAE module shares the same network architecture consisting of multiple stacked NCSAE layers and one Softmax regression layer at the end. The contribution of our model is that hierarchical feature evolution and nonlinear feature mapping are jointly optimized in a unified and perceptual-aware deep network (DNCSAE), which well resembles several important visual properties, i.e., hierarchy, sparsity, and non-negativity. To be more specific, for each DNCSAE, by taking a set of handcrafted natural scene statistic (NSS) features as inputs in the visible layer, the features in hidden layers are successively evolved to deeper levels producing increasingly discriminative quality-aware features (QAFs). Then, QAFs in the last NCSAE layer are summarized to their corresponding quality score by Softmax regression. Finally, three individual yet complementary quality scores estimated by each DNCSAE model are combined based on a Bayesian framework to obtain an overall 3D quality score. Experiments on three benchmark databases demonstrate the superiority of our method in terms of both prediction accuracy and generalization capability.

1. Introduction

Multimedia services in the format of stereoscopic three-dimensional (S3D) have become increasingly popular due to the provided immersive depth perception and enjoyable visual experience. The tremendous explosion of stereo shooting and visualization devices has led to a considerable increase in the amount of deployed stereoscopic data over networks [1,2]. However, the quality issues of S3D visual contents, including image distortion, visual discomfort, and visual presence, may confuse the human visual system (HVS) and negatively impact the 3D viewing experience [3–6]. To quantify the impact of such quality-related factors on 3D viewing experience, it is of vital importance to design effective and robust S3D image quality evaluation engines. As the interactions between different quality factors are extremely complex and have not been fully understood at the current stage, in this study we mainly focus on one most important quality factor, i.e., image distortion.

1.1. Related work

A fact needs to be aware is that human eyes are the ultimate receivers of visual signals in most applications. Thus, subjective evaluation that straightforwardly reflects human visual perception is considered as the most reliable way to assess image quality. However, subjective image quality assessment (IQA) is criticized to be time consuming and tedious as it typically requires many observers to participate into the subjective experiments. To address this problem, objective IQA which aims to computationally predict image quality in line with subjective judgment has attracted much attention. According to the participation of the original image, objective IQA can be roughly categorized into three types: i.e., full-reference (FR) [7,8], reduced-reference (RR) [9,10], and no-reference (NR) [11–19] models. For FR/RR ones [3–6], information of the reference images is fully/partially utilized for quality evaluation. Given that the participation of reference image greatly facilitates accurate quality prediction, FR/RR models may lose their

[*] Corresponding author.
E-mail address: shaofeng@nbu.edu.cn (F. Shao).

https://doi.org/10.1016/j.patcog.2017.11.001
0031-3203/© 2017 Elsevier Ltd. All rights reserved.

power in the cases where the reference images are not available. In such cases, NR models, which estimates the image quality merely based on the information extracted from the distorted image itself, would be the only applicable solution. Actually, great efforts have been made on the development of NR-IQA, evidenced by a plenty of 2D NR-IQA models achieving high correlation with subjective scores recorded in the existing 2D IQA databases [11–19].

The perceptual issues related to stereo/3D image quality assessment (SIQA) are more complex than its 2D counterparts because additional binocular visual characteristics also impact the stereo quality perception. Most SIQA metrics fall into the FR [20–26] and RR [27–29] categories. As the new challenges in SIQA mainly stem from binocular interactions, the design of SIQA models may be particularly beneficial from the simulation of binocular visual characteristics, such as binocular fusion, binocular suppression, and binocular rivalry. For example, Bensalma et al. [21] advanced a novel quality metric for stereopairs called binocular energy quality metric (BEQM) which estimated the stereo image quality by computing the binocular energy difference between the original and distorted stereopairs. In BEQM, the basic idea was to estimate binocular energy by simulating the properties of both simple and complex cells in the visual cortex. As human brain inherently merges the visual information from the left and right eyes into a binocular vision, a merged view known as the cyclopean image was synthesized from the stereopair and its corresponding disparity map to characterize the underlying binocular perception mechanism in [22]. Then, by applying traditional 2D quality metrics on this cyclopean image, a final 3D quality score was obtained. Lin and Wu [23] simulated the critical binocular integration behaviors and utilized them as the bases for binocular combination when transferring the existing 2D-IQA metrics into 3D domain. Similarly, Zhang and Chandler [25] extended the traditional 2D most apparent distortion (MAD) method to a 3D version 3D-MAD which decomposed the problem of 3D quality assessment into two modules: 1) monocular quality estimation and 2) cyclopean quality estimation. A new binocular brightness and contrast perception measure is introduced as the bases for binocular combination in 3D-MAD. Given the critical role of primary visual cortex in binocular perception, Shao et al. [24] developed a novel FR-SIQA method from the perspective of learning binocular receptive field properties using sparse representation which has been demonstrated well resemble the properties of simple cells in primary visual cortex. To deal with the new challenges in quality evaluation of asymmetrically distorted stereopairs, Wang et al. [26] devised a binocular rivalry inspired multi-scale model and utilized it as the basis for binocular combination.

The problem of NR-SIQA is less investigated than FR- and RR-SIQA. Although challenging, recent years have more or less witnessed some progresses in this field. Ryu and Sohn [30] proposed a NR-SIQA metric by utilizing the local blockiness, blurriness, and visual saliency information to estimate single-view blockiness and blurriness score. Then, single-view scores were fused by a binocular perception model. Chen et al. [31] proposed a NR-SIQA metric which extracts both 2D and 3D natural scene statistics (NSS) features from the synthesized cyclopean image for stereopair quality analyzes. Zhou et al. [32] proposed a NR-SIQA metric form the perspective of simulating the critical binocular combination and rivalry properties of the HVS. Similar with the application of NSS features in 2D NR-IQA, Appina et al. [33] utilized a bivariate generalized Gaussian distribution (BGGD) model to fit the distribution of luminance and disparity coefficients and used the fitting parameters as the final 3D NSS-based quality-aware features (QAFs). It is worth noting that the previously reviewed NR-SIQA methods differ in the way of extracting QAFs, but share the same structure for feature mapping, i.e., using shallow learning algorithms to create the nonlinear mapping from handcrafted QAFs to subjective quality scores. With shallow learning structures, the performance of these NR-SIQA metrics vastly depends on the discriminability of handcrafted QAFs. As it is known, extracting highly discriminative features from stereopairs remains a challenging problem due to the strong dependency of domain knowledge in stereo vision.

Most recently, deep learning techniques have been utilized to overcome the drawbacks of shallow learning structures in addressing NR-SIQA. We review some representative works here. Zhang et al. [34] proposed a novel NR-SIQA method by learning the underlying structures of stereo images based on convolutional neural network (CNN). Shao et al. [35] devised a 3D deep blind quality evaluator (3D-DQE) for stereopairs by taking the critical monocular and binocular interactions into account. By regarding 3D quality prediction as the result of monocular and binocular interactions, a novel 3D quality prediction framework is formulated. In that framework, monocular and binocular quality scores are first estimated by the pre-trained 2D- deep neural network (DNN) and 3D-DNN models, and then integrated based on a Bayesian inference-based framework. Lv et al. [36] employed the DNN to predict patch-level quality of stereoscopic images. Then, a variance energy-based gain control model is used to combine the left and right-view image qualities to produce a binocular integration (BI) index. Besides the BI index, they also defined a binocular self-similarity (BS) index based on the depth image-based rendering (DIBR) technique. The final 3D quality is obtained by a linear combination of BI and BS. With the employment of deep learning structures, more discriminative features can be obtained.

From all the above reviewed works, we claim that the main problems to be solved in NR-SIQA are twofold: 1) discriminative feature representation for monocular quality prediction; 2) adaptive binocular combination of monocular quality scores. Existing NR-SIQA metrics differ in both the way and the order that they consider the above two aspects. For example, Chen's [31] and Zhou's [32] metrics firstly simulate the process of critical binocular combination to create a merged-image based on which the quality-aware features are extracted. By contrast, Ryu's [30], Zhang's [34], Shao's [35] and Lv's [36] metrics firstly extract quality-aware features from each single-view image for monocular quality prediction using either shallow or deep learning architectures. Then, some critical visual characteristics are considered and modeled for adaptive binocular combination of the previously estimated monocular quality scores.

1.2. Motivation and methodology overview

It is an intuitive idea to devise appropriate computational models that well resemble the visual perception and cognitive properties to facilitate the research of IQA. Extensive studies from physiological and cognitive sciences have discovered three important characteristics in the process of visual perception and scene understanding. These critical visual characteristics include hierarchy, sparsity, and non-negativity. First, the hierarchy refers to the fact that the input visual signal is organized and processed in a hierarchical structure where the high-level semantic features in top layers are successively evolved from the basic visual primitives in low layers [37]. Second, the sparsity refers to the fact that the patterns of neuron activity provoked by an input visual signal are inherently sparse, i.e., the input stimulus activates only a relatively small number of neurons [38]. Third, the non-negativity refers to the fact that the input visual signal is intrinsically decomposed in a part-based representation fashion where the weights tend to be non-negative [39,40]. Hence, the efforts toward learning features accounting for the above characteristics will greatly contribute to the design of advanced quality metrics.

In the literature, there have been several attempts that consider parts of the above characteristics in training perceptual-aware neu-