Journal of Visual Communication and Image Representation 50 (2018) 247–262

Contents lists available at ScienceDirect

Journal of Visual Communication and Image Representation

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

No reference stereo video quality assessment based on motion feature in tensor decomposition domain

Gangyi Jiang[a,b,*], Shanshan Liu[a], Mei Yu[a,b], Feng Shao[a], Zongju Peng[a], Fen Chen[a]

[a] Faculty of Information Science and Engineering, Ningbo University, Ningbo, China
[b] National Key Lab of Software New Technology, Nanjing University, Nanjing, China

ARTICLE INFO

Keywords:
No reference stereo video quality assessment
Tensor decomposition
Motion feature
Entropy
Random forest

ABSTRACT

A no reference stereo video quality assessment method based on motion features extracted in tensor decomposition domain is proposed. Tensor decomposition is used to reduce dimension of color, view and time of stereo video, and motion information maps containing time-varying information of inter-views and intra-views are obtained. Statistical features such as generalized Gaussian distribution (GGD), asymmetric GGD, spatial entropy, spectral entropy associated with two views, and spectral entropy related to depth perception of stereo video, are extracted. Random forest is utilized to establish relationship between stereo video quality and the extracted features. Experimental results on NAMA3DS1-COSPAD1 database demonstrate that the proposed method achieves good performance on JP2K, resolution reduction, sharpening and their combination distortions, Pearson linear correlation coefficient (PLCC) values of these types of distortions are higher than 0.97, while for H.264 distortion the PLCC value is 0.8850, which means that the proposed metric is consistent with human visual perception.

1. Introduction

Stereo video has board application prospect due to its excellent immersive experience [1]. However, stereo video technologies have the same problem as that for 2-dimensional (2D) videos, that is, the distortions which cannot be avoided during the process of the video [2,3]. In order to assess process performance during the processes of acquisition, processing, compression, storage and transmission of videos, and provide satisfied display for the user of 3-dimensional (3D) video system, objective video quality assessment (VQA) is necessary for the user experience [4,5].

According to the availability of original video as the reference, objective VQA methods can be roughly classified into three categories: full reference (FR), reduced reference (RR), and no reference (NR). FR methods need full original video as the reference, however in most cases this is impossible in practice. By contrast, RR methods only need representative features of the original video, while NR methods do not require any information of reference video. Therefore, from a practical perspective, NR methods are most expected for VQA.

Generally, stereo video has two views, and obviously the quality of each view is related to the overall quality of the stereo video. However, compared with 2D video, stereo video further provides depth information which enhances the viewing experience. Thus, depth information is also significant to the overall quality of stereo video.

Yu et al. proposed a stereo VQA method, which considered the influence of temporal characteristics of video and binocular perception in human visual system (HVS) [6]. Galkandage et al. presented a FR method based on an extended HVS model including the phenomena of binocular suppression and recurrent excitation, etc [7]. But these two methods need to know all or partial information of the original video, which results in limitation when applied to practical applications. Zhao et al. thought human eyes being more sensitive to moving object and edge information [8], and put forward a NR 3D video metric on the basis of visual attention and edge difference. Nevertheless, the accuracy of disparity has large influence on this objective metric. Han et al. considered the correlation between network packet loss and perceptual video quality for different bit-rate video sequences [9]. Moreover, they modeled the impact of network packet loss at different bit-rates and frame rates on the perceived quality of stereo video to make the video quality metric more generic [10]. But the metric proposed by them is only suitable for network delivery effects. For a specific network, parameters need to be computed. In addition, in the case of bit stream

* Corresponding author at: Faculty of Information Science and Engineering, Ningbo University, Ningbo, China.
E-mail address: jianggangyi@126.com (G. Jiang).

https://doi.org/10.1016/j.jvcir.2017.12.001
Received 14 April 2017; Received in revised form 1 November 2017; Accepted 9 December 2017
Available online 11 December 2017

unavailable due to the fact that it is encrypted or processed by the third part decoders, this kind of bit stream information based metric is invalid [11].

Natural scene statistics (NSS) models have been researched extensively and achieve good performance in image quality assessment (IQA). Based on statistical characteristics including generalized Gaussian distribution (GGD), asymmetric GGD (AGGD), entropy, etc., many NR methods have been proposed [12,13]. But the performance of this kind of metrics for VQA is not as well as for IQA, because some of these statistical features do not have the ability to distinguish the distortion degree of the video. In the past few years, many researchers have focused on employing NSS feature to measure video quality. Saad et al. studied the NSS features of frame difference in DCT domain [14]. Soundararajan et al. researched the frame difference of wavelet coefficients [15]. These researches confirmed that the frame difference of original videos has certain distribution regularity, and frame difference can represent the structure of the motion edge. Therefore, the frame difference can be used to obtain temporal information which is important for VQA.

Scalars and vectors are commonly used in traditional data processing. But the real-world data are usually multi-dimensional. For example, a gray image is 2-dimensional, a color image is 3-dimensional, color video is 4-dimensional, and color 3D video with two views composes of 5-dimensional data. Therefore, scalars and vectors can not reflect the complex structure of the real world data. By contrast, tensor decomposition is suitable for multi-dimensional data processing. Over the past few decades, tensor decomposition based methods have been widely adopted in medical imaging, surveillance, machine learning, etc [16]. The CANDECOMP/PARAFAC (PC) [17] and Tucker [18] families are the mainly used classes of tensor decomposition, and many other tensor decomposition methods are derived from them.

In this paper, we propose a NR stereo video quality metric called motion feature based no reference stereo video quality metric (MNSVQM). The color stereo video, which can be represented as 5-dimensional tensor, is processed by Tucker decomposition implemented through N-mode singular value decomposition (SVD) [18]. By analyzing the principal component of the N-dimensional data, the time-varying information of the video is obtained to construct the motion information map, and four kinds of features are then extracted from the motion information map in tensor domain based on statistical models such as spatial entropy and spectral entropy, GGD and AGGD model. Benefiting from their obvious statistical regularity, these features are used to distinguish distortion type and degree of the stereo video. Finally, random forest is adopted to model human visual perception based on these features so as to predict the quality of the stereo video.

The remaining parts are organized as follows. Section 2 describes the proposed NR stereo VQA metric in details. In Section 3, experimental results and discussions are presented. Conclusions are drawn in Section 4.

2. The proposed motion feature based no reference stereo video quality metric

In this paper, we use tensor decomposition to extract main motion information from the video, and propose a NR stereo video quality assessment method, the diagram of which is shown in Fig. 1. The proposed method consists of three parts: motion information map acquisition, feature extraction, random forest model training and video quality prediction.

It is clear that time-varying information is important to the quality of a video. In this paper, motion information map is acquired with tensor decomposition. The tensor decomposition can be implemented through N-mode SVD, which can realize the principal component analysis (PCA) of high-order data. PCA process can obtain main information of the signal and achieve dimensionality reduction, it can compress the data without loss of data as much as possible, and obtain the linearity of the original variable combination [19]. PCA is good at one-dimensional and two-dimensional data processing, but has some problems when used for high-order data, for example, the monocular gray-scale video which contains three dimensions such as space and time, that is, it is the 3rd-order tensor. PCA processing will ignore the temporal and spatial relations of the video. By contrast, tensor decomposition can take into account the temporal and spatial information of the video [20], which is quite important for VQA. In this paper, the purpose of feature extraction in tensor decomposition domain is to obtain the features related to stereo video quality on the basis of time-varying information as well as depth information. Then, these features are utilized for the subsequent random forest model training and video quality prediction.

Let the frame size of stereo video be $W \times H$, K be the color dimension, S denote the time dimension. For each view of the stereo video, makes every S frames to form a group so that a 4th-order tensor $\chi \in \mathbb{R}^{W\times H\times K\times S}$ can be obtained, and the tensor is called group of frame (GOF) tensor hereinafter. The first two dimensionalities (i.e. mode-1 and mode-2) of χ represent the spatial information, the third dimensionality (i.e. mode-3) represents the RGB color information, and the mode-4 represents the time information, respectively. Then the left and right views of the stereo video can be represented by sets $\{\chi_1^L, \chi_2^L, \cdots, \chi_T^L\}$ and $\{\chi_1^R, \chi_2^R, \cdots, \chi_T^R\}$, respectively, where T is the number of GOF tensor in one view of the stereo video. The mode-3 matrix of χ is processed with SVD decomposition, so as to reduce the color dimension from three-dimensional to one-dimensional. And then, same processing is performed in time dimension to obtain the motion information map. At the same time, to obtain depth information of the stereo video, for each pair of GOFs, for example, the t-th GOFs of the left and right views ($1 \le t \le T$), the frames can also form a new 4th-order tensor set $B_t\{y_1, \ldots, y_S\}$, $y_i \in \mathbb{R}^{W\times H\times K\times V}$, where the first three dimensionalities (or modes) of y_i have the same meanings as that of χ_t^L and χ_t^R, while the fourth dimensionality V represents the view of the stereo video. After Tucker decomposition on the view and color dimensions of y_i, the $W \times H \times 3 \times 2$ tensor y_i is transformed to a $W \times H$ matrix that contains the main view and color information of y_i, then the S matrices corresponding to $B_t\{y_1, \ldots, y_S\}$ forms a new 3rd-order tensor $Z \in \mathbb{R}^{W\times H\times S}$. The dimension of the mode-3 matrix of Z can be reduced to obtain the motion information map which contains the time-varying information in the views as well as depth information between views, because the mode-3 matrix of Z is derived from the last two dimensionalities of y which represent the color and view dimensions respectively. After that, some statistical features are extracted in tensor decomposition domain and further pooled to obtain the overall video features. Finally, the relationship between the features and the subjective evaluation scores is modeled with random forest, and an objective metric used for predicting the quality of stereo video is obtained.

2.1. Tucker tensor decomposition

Tucker tensor decomposition decomposes a tensor into a set of matrices and a core tensor, and the Tucker family includes the Tucker1, Tucker2 and Tucker3 models. Tucker tensor decomposition can be described by

$$\min_{C,\mathrm{A}^{(1)},\cdots,\mathrm{A}^{(N)}} \|\chi - C \times_1 \mathrm{A}^{(1)} \times_2 \mathrm{A}^{(2)} \cdots \times_N \mathrm{A}^{(N)}\|_F$$

$$\text{s.t. } C \in \mathbb{R}^{R_1\times R_2\cdots\times R_N},\ \mathrm{A}^{(n)} \in \mathbb{R}^{I_N\times R_n},\quad n = 1, \cdots, N \tag{1}$$

where the symbol $\chi \in \mathbb{R}^{I_1\times I_2\cdots\times I_N}$ is a tensor and C is the core of the tensor. A is 2D matrices. $||\bullet||_F$ is the Frobenius norm of the matrix. The n-mode (matrix) product of a tensor is denoted by $\times_n$. The matricization (also known as unfolding or flattening) of a tensor is the process of reordering the elements of an N-order array into a matrix $\mathrm{X}_{(n)}$, whose dimension is $I_n \times \prod_{j=1,j\neq n}^{N} I_j$. For example, let the elements of $\chi \in \mathbb{R}^{2\times 2\times 2}$ be $x_{111} = 1$, $x_{121} = 2$, $x_{211} = 3$, $x_{221} = 4$, $x_{112} = 5$,