Neurocomputing 000 (2017) 1–12

Contents lists available at ScienceDirect

Neurocomputing

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

Spatiotemporal visual saliency guided perceptual high efficiency video coding with neural network

Shiping Zhu*, Ziyao Xu

Department of Measurement Control and Information Technology, School of Instrumentation Science and Opto-electronics Engineering, Beihang University, Beijing 100191, China

ARTICLE INFO

Article history:
Received 21 January 2017
Revised 28 June 2017
Accepted 15 August 2017
Available online xxx

Communicated by Prof. Yicong Zhou.

Keywords:
Perception
HD video
Saliency
Video compression
HEVC

ABSTRACT

The perceptual video coding systems for optimization have been developed on the basis of different attributes of the human visual system. The attention-based coding system is considered as an important part of it. The saliency map method representing the region-of-interest (ROI) from the video signal has become a reliable method due to advances in the computer performance and the visual algorithms. In the present study, we propose a hybrid compression algorithm that uses the deep convolutional neural network to compute the spatial saliency followed by extraction of the temporal saliency from the compressed-domain motion information. The level of uncertainty is calculated to combine to form the video's saliency map. Afterwards, the QP search range is dynamically adjusted in HEVC, and a rate distortion calculation method is proposed to choose the pattern and guide the allocation of bits during the video compression process. Empirical reporting results proved the superiority of the proposed method over the state-of-the-art perceptual coding algorithms in terms of saliency detection and perceptual compression quality.

1. Introduction

Constant advances have been made towards a deeper understanding of the perceptual features of the human visual system and a higher-performance computation model. This makes it possible to maximize perceptual quality by eliminating the people's perceptual redundancy during video compression [1].

Studies on the people's perceptual system are of great significance to the processing of video signal. Recently, much attention has been focused to develop the idea of enhancing video perceptual quality by introducing the human perceptual computation model to the video coding system [2]. Generally, the following three challenges need to be addressed while designing a perceptual coding system:

- To determine the kind of perceptual models for finding the parts of video data susceptible to coding distortion.
- To determine the method to integrate corporate perceptual model in encoding.
- To validate the effectiveness of the algorithm.

The attention-based perceptual coding system is one of the important disciplines. Typically, this type of methods begins with calculating the video frame's saliency feature map. A latest study on the eye-tracking [3] for variations of inter-observer saliency in task-free viewing of natural images has revealed that images with known salient regions create extremely correlated eye-fixation maps for the different viewers. Moreover, with regular advances in the saliency models has proved that the visual saliency regions that attract human attention coincide more with the regions of human interest [4]. In the saliency-based video compression method, more number of bits is allocated to ROI to decrease the distortion, and less number of bits is allocated to non-ROI parts to optimize the compression ratio. In this way, the method provides better perceptual quality without increasing the total number of bits.

Some saliency-based methods attempt to enhance the compression effectiveness from different perspectives [5–18]. The bottom-up model was introduced to the perceptual models. For example, Itti et al. focused on color, glint, brightness, direction and motion, extracting low-level features and combining saliency in a non-linear manner. Finally, they proposed a model that center-surround the response characteristics of the low-level visual neurons in the primate animal's brain [5]. Li et al. [6] proposed the 3 feature maps consisting of center-surround multi scale ratio of dissimilarity, color spatial variance, and pulse DCT. The up-bottom models were used in some works to determine the boundary of ROI according to the face [8–10] or the moving objects [11,12]. The authors of the study [13] proposed a method that aggregates spatial

* Corresponding author.
E-mail address: spzhu@163.com (S. Zhu).

http://dx.doi.org/10.1016/j.neucom.2017.08.054
0925-2312/

and temporal saliency map. Most of the work on the video compression frameworks focused on optimization by taking saliency into account. In the most commonly used methods, the quantization parameters are adapted to different macro blocks [13–16] to ensure that the perceptually important macro blocks have higher coding quality using small QP values. In [17], different coding parameters (e.g., ME search range) were allocated to ROI and non-ROI parts. In [7,18], rate distortion was optimized by adapting the distortion degree computation methods to different patterns.

The evolution of video coding standards leads to more and more perceptual coding algorithms. For example, with the compatible decoder, the HEVC-based scheme is used to allocate bits by adjusting CU partition. This improvement is impossible in the H.264-based methods and other old standards.

In the present study, we propose a hybrid perceptual coding algorithm on the basis of the latest HEVC framework. The proposed method involves saliency detection, coding strategy optimization and objective evaluation.

We propose a spatial-temporal saliency combining algorithm as a perceptual model. As we know, the video's spatial saliency is very similar to the image's saliency, and the effectiveness of the convolutional neural network has been proven in recent years. Hence, we design a convolutional neural network to measure the video's spatial saliency. The research on a feasible compression-domain motion detection technique has been done for a long time. We can compute the spatial saliency of the video by processing the motion detection results obtained during the coding process. We calculate the uncertainty of the saliency map in spatial and temporal domains, and then combine their uncertainties to form the video's saliency feature as a motivation of the work proposed in [19].

We also present a video compression algorithm centered on dynamic adjustment of rate distortion. We can provide more backup patterns in HEVC than in H.264 through block partition, inter- and intra-frame search. Manually designing a customized strategy for each part is not as reliable as H.264, underscoring the need to formulate a uniform criterion for choosing patterns. A rate distortion optimization algorithm is suggested in the paper to address the current problem. In fact, the mode selection is considered as an optimization problem for reducing the value of distortion D for a given rate. It can be defined as per (1) described below:

$$min\{D\} subject\ to\ R \le R_c \tag{1}$$

Where, R and D represent rate and distortion respectively. The above cited problem with constraints for optimization can be rewritten to an unconstrained form by the Lagrange multiplier method as described in (2) described below:

$$min\{J\},\ \ where\ J = D + \lambda R \tag{2}$$

Where, J and λ represents the Lagrange cost function and Lagrange multiplier respectively. Generally, λ is obtained through empirical results and typical rate-distortion models [20]. Study in [40] shows that λ could be adapted based on inter-frame dependency for better R-D performance. In order to gain better perceptual R-D performance, we propose a saliency-based algorithm to calculate D, where perceptual redundancy is eliminated by changing the video's bit allocation strategy.

2. Saliency estimation in video

2.1. Spatial saliency

When we talk about the video's spatial saliency, we are actually talking about the saliency of each frame in the video. Various computational models of saliency detection for images have been suggested [2,4,5] initialized by Itti et al. [5]. The saliency algorithms based on CNN have proved their superiority over other methods in recent years.

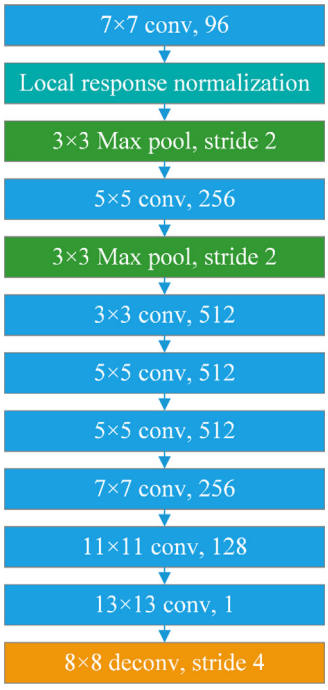

Fig. 1. Architecture of the convolutional network.

Convnets are common architectures in the field of deep learning. These architectures have been deeply researched for visual pattern recognition with its applications from a global scale image classification [21] to a more local object detection [22] or semantic segmentation [23]. The structure of the layers comprising convnets are the inspiration of by biological models. Even, a relationship between the activity of certain areas in the brain with hierarchy of layers in the convnets [24,25] have been highlighted in the literature. The convnets have shown better results and is generally outperforming other hand-crafted methods for a huge amount of training data [26,27].

In the proposed study, we adopted a multi-layer convolutional neural network as depicted in Fig. 1. The proposed network consists of 10 weight layers having 25.8 million parameters in total. The architecture of the first 3 weight layers is made compatible with that of the VGG network as described in the study [28]. These layers can be categorized as below:

(1) Convolutional layer. The convolutional layer aims to detect local feature patterns and local conjunctions from the previous layer, where neurons are organized in feature maps. For the *j*th channel at the *l-1th* layer, the convolutional operation can be written as:

$$u_j^l = \sum_{i \in M_j} x_i^{l-1} \cdot k_{i,j}^l + b_j^l \tag{3}$$

Rectified linear unit nonlinearity (ReLU) follows each convolutional layer.

(2) Local Response Normalization (LRN) layer. The layer enables to enhance the model's generalization ability. In the *l*th LRN layer, the response normalized output is given by the expression

$$u_{i,x,y}^l = \frac{x_{i,x,y}^{l-1}}{\left(k + \alpha \sum_{j=max\left(0,i-\frac{n}{2}\right)}^{min\left(N-1,i+\frac{n}{2}\right)} \left(x_{j,x,y}^{l-1}\right)^2\right)^{\beta}} \tag{4}$$

Please cite this article as: S. Zhu, Z. Xu, Spatiotemporal visual saliency guided perceptual high efficiency video coding with neural network, Neurocomputing (2017), http://dx.doi.org/10.1016/j.neucom.2017.08.054