Full reference video quality assessment using convolutional neural networks

Anuj Mohan Pillai*
Indian Institute of Information Technology Vadodara
201952205@iiitvadodara.ac.in

Pramit Mazumdar
Indian Institute of Information Technology Vadodara
pramit.mazumdar@iiitvadodara.ac.in

ABSTRACT

In this paper, we propose a full reference video quality assessment model, which predicts the quality score of a distorted video, comparing it with its corresponding reference/original video. The proposed model evaluates a video on the basis of various parameters such as, spatio-temporal features, structure, texture, salience, etc. The proposed model can be broadly divided into two parts, first is the feature extractor network and subsequently the regression network. For the feature extraction part, we obtain the feature level matrix from the convolutional neural network layers, which allows the model to extract all the visual information from top to bottom. Specifically, the structure and texture similarities of feature vectors extracted from all intermediate layers are calculated as the feature representation for this model. For the regression part, we use fully connected layers to obtain a patch level quality score and patch weights for each frame, which are finally spatially pooled to obtain a frame quality score. In the end, all these frame level scores for a distorted video are passed through a temporal pooling layer to obtain the resultant video quality score.

CCS CONCEPTS

• **Human-centered computing → Virtual reality**; • **Computing methodologies → Interest point and salient region detections**.

KEYWORDS

Video quality assessment, Feature extraction, Convolutional neural networks, Fully connected layers, Deep learning

ACM Reference Format:
Anuj Mohan Pillai and Pramit Mazumdar. 2018. Full reference video quality assessment using convolutional neural networks. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Audio-visual communication technologies are rapidly increasing in recent years. This includes media to human systems such as streaming over OTT platforms, and also human to human systems such as video conferencing platforms. The transmitted media also depends upon factors such as increased resolution (4K, HDR), and also high network bandwidths (4G). Therefore, there is a need to assess the perceived quality with respect to the Quality of Experience (QOE) of the end-users based on the varying set of factors. Video Quality Assessment (VQA) studies the factors that affect the perceived visual experience of a video and subsequently, proposes an objective quality metric to computationally estimate the visual quality of a video before transmission. Industries such as Facebook, YouTube, Netflix, Qualcomm, Tencent, Nokia, etc. are working in this direction and their benchmark quality metrics are made available for reuse. Generally, video/image quality measures are classified depending on the amount of information available from an original reference image – full reference (FR), reduced reference (RR) or no reference (NR). FR video quality assessment models usually measure the fidelity between the reference and distorted frames of videos as the video quality [4]. NR approach on the other hand assess quality of distorted videos without considering the source reference videos [2]. In this work, a full reference video quality assessment model is presented.

2 PROPOSED MODEL

The proposed model is illustrated in Fig. 1 Here the inputs to the model are a distorted video and its corresponding reference video, using which the quality score for the distorted video will be predicted. Firstly, the video is divided into frames, out of which, 'n' distinct frames are selected from each video and these are used to calculate the frame level features. These frame-level feature vectors from the reference and distorted video are then fused together and fed into the regressor and patch weighing module correspondingly. They output the frame-level quality scores for the distorted video, which are converted to video quality score using a temporal pooling strategy.

2.1 Feature Extraction

Frames extracted from the distorted video are passed through a convolutional neural network model, which will help the CNN model learn about the visual information present in each frame [5]. Now, after obtaining the frame level features vector, we measure the distance between the reference feature vector and the distorted feature vector. This is achieved by evaluating the feature vectors on structural and textural similarities [3]. Finally, we obtain the structure and texture similarities of feature vectors, extracted by each stage of the CNN model and all of them constitute the quality aware feature representation of our proposed model.

[*] Third Year student, B.Tech IT at IIIT Vadodara

Woodstock '18, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

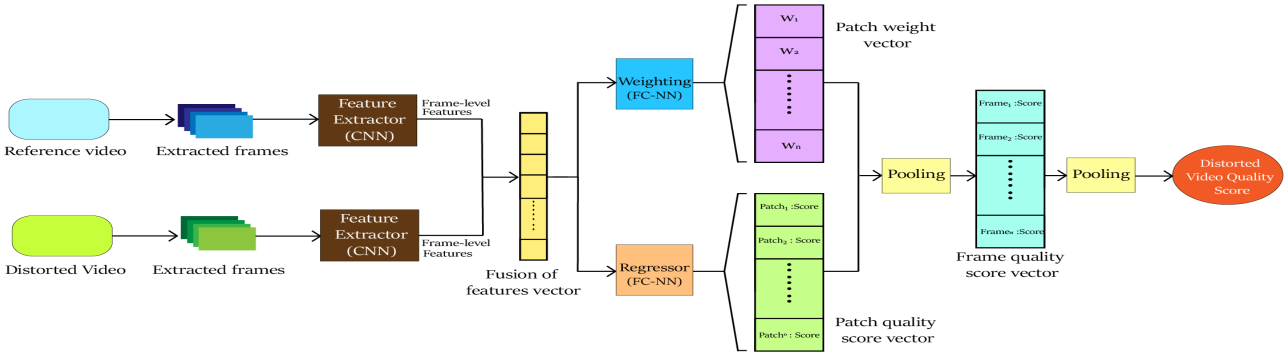

Figure 1: The proposed model framework. It takes the reference and distorted videos as input, converts the video into frames, evaluates each frame, and provides a quality score per frame. Quality scores of all the frames are then used to calculate the overall video quality score.

2.2 Regression and weighting

In order to serve as input to the regression part of the network, the extracted feature vectors are combined in a feature fusion step. Then this is fed to the regression and weighing networks. Regression layer consists of a fully connected neural network which outputs the patch-wise quality score for a single frame. The weighting layer depicts the salience of a particular patch in its corresponding frame [1].

Fig. 1 The proposed model framework. It takes the reference and distorted videos as input, converts the video into frames, evaluates each frame, and provides a quality score per frame. Quality scores of all the frames are then used to calculate the overall video quality score.

2.3 Pooling

The outcome from the regression layer and the weighing layer are pooled using a spatial pooling technique to estimate a quality score of a frame [4]. Subsequently, we give weights to different frames using a temporal pooling strategy which finally provides the video quality score predicted by the model for the input video.

3 CONCLUSION

In this paper a neural network-based video quality assessment model is proposed that allows feature learning and regression in an end-to-end framework. The model considers structural and textural similarities for the videos. This significantly helps in the network to learn the features. Also the proposed model considers weighing each frame based on salience. The pooling technique used in this approach also improves the overall assessment. The model would be evaluated on benchmark video quality assessment datasets using metrics such as SROCC, KROCC, RMSE, etc. The network architecture can be easily modified for incorporating quality assessment for other media such as 360° videos, live streams, sports videos, point clouds, light fields, etc. Video quality assessment models are necessary in today's world, where online entertainment platforms are expanding rapidly. These kinds of models help the service provider enhance the user experience and provide consumers with good quality media.

REFERENCES

[1] Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek. 2017. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on Image Processing* 27, 1 (2017), 206–219.

[2] Kamal Lamichhane, Pramit Mazumdar, Federica Battisti, and Marco Carli. 2021. A No Reference Deep Learning Based Model for Quality Assessment of UGC Videos. In *International Conference on Multimedia & Expo Workshops*. IEEE, 1–5.

[3] Wei Sun, Tao Wang, Xiongkuo Min, Fuwang Yi, and Guangtao Zhai. 2021. Deep Learning Based Full-Reference and No-Reference Quality Assessment Models for Compressed UGC Videos. In *International Conference on Multimedia & Expo Workshops*. IEEE, 1–6.

[4] Munan Xu, Junming Chen, Haiqiang Wang, Shan Liu, Ge Li, and Zhiqiang Bai. 2020. C3DVQA: Full-reference video quality assessment with 3d convolutional neural network. In *International Conference on Acoustics, Speech and Signal Processing*. IEEE, 4447–4451.

[5] Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*. Springer, 818–833.