Stereoscopic Video Quality Prediction Based on End-to-End Dual Stream Deep Neural Networks

Wei Zhou, Zhibo Chen(✉), and Weiping Li

CAS Key Laboratory of Technology in Geo-spatial
Information Processing and Application System,
Department of Electronic Engineering and Information Science,
University of Science and Technology of China, Hefei 230027, China
weichou@mail.ustc.edu.cn, {chenzhibo, wpli}@ustc.edu.cn

Abstract. In this paper, we propose a no-reference stereoscopic video quality assessment (NR-SVQA) method based on an end-to-end dual stream deep neural network (DNN), which incorporates left and right view sub-networks. The end-to-end dual stream network takes image patch pairs from left and right view pivotal frames as inputs and evaluates the perceptual quality of each image patch pair. By combining multiple convolution, max-pooling and fully-connected layers with regression in the framework, distortion related features are learned end-to-end and purely data driven. Then, a spatiotemporal pooling strategy is employed on these image patch pairs to estimate the entire stereoscopic video quality. The proposed network architecture, which we name End-to-end Dual stream deep Neural network (EDN), is trained and tested on the well-known stereoscopic video dataset divided by reference videos. Experimental results demonstrate that our proposed method outperforms state-of-the-art algorithms.

Keywords: Convolutional neural network · Stereoscopic video · No-Reference video quality assessment · Spatiotemporal pooling

1 Introduction

Stereoscopic video quality assessment (SVQA) is challenging because the left and right views of 3D/stereoscopic videos can synthetically generate depth perception, which leads to an additional perceptual dimension to be considered. Through the whole 3D media processing chain from acquisition, compression, to transmission, reconstruction, and display, etc., original stereoscopic videos undergo a variety of quality degradations. Consequently, it is important to effectively predict and optimize the quality of experience (QoE) throughout the processing chain.

According to the existence of non-distorted reference videos, SVQA algorithms can be divided into three categories: full-reference SVQA (FR-SVQA), reduced-reference SVQA (RR-SVQA), and no-reference SVQA (NR-SVQA). The main purpose of SVQA is to design an objective criterion to accurately predict

the perceptual subjective quality of 3D videos. During recent years, some 3D video quality metrics have been proposed. Early FR-SVQA models requiring pristine stereoscopic videos were studied. For example, the perceptual quality metric (PQM) used conventional 2D objective metrics to assess 3D video quality [9]. The modified PSNR, called PHVS-3D, exploited 3D discrete cosine transform (3D-DCT) to evaluate the perceptual quality of stereoscopic videos [8]. In [18], the spatial frequency dominance (SFD) model considered the phenomenon that spatial frequency determines view domination based on the human visual system (HVS). The 3D spatial-temporal structural (3D-STS) metric utilized the inter-view correlation of spatial and temporal structural information [4]. In [21], an objective metric named SJND-SVA was designed by integrating the stereoscopic visual attention (SVA) with the stereoscopic just-noticeable difference (SJND).

The drawback of these FR models is that original 3D videos are not always available in most practical situations. Thus, NR-SVQA models should be developed to assess stereoscopic video quality without needing reference 3D videos. In [26], an NR optical flow-based method was developed to predict 3D video quality. Recently, the motion feature based no reference stereo video quality metric (MNSVQM) [7] and the blind stereoscopic video quality evaluator (B-SVQE) [2] have been proposed. However, most of SVQA algorithms still extract hand-crafted features from stereoscopic videos, and then yield visual quality evaluation. In other words, one of the advantages of applying deep neural network (DNN) to SVQA is that it can directly input raw image/video and combine feature learning with quality regression in the training stage. Moreover, the DNN-based SVQA metrics are robust and can be trained end-to-end with little prior domain knowledge. Therefore, this paper focuses on studying the NR-SVQA method for 3D videos using DNN-based end-to-end learning.

Deep learning has achieved remarkable results in object detection, image classification, and recognition [3,14,11]. At the same time, the application of DNN in image/video quality assessment has also been started to be explored in several studies. However, how to effectively predict the quality of image/video using DNN is quite different from traditional computer vision tasks. Specifically, DNN based image/video quality prediction is a challenging problem, mainly due to most of the existing data augmentation as well as patch preprocessing techniques are not suitable for image/video quality assessment [13].

From the perspective of training strategies, the relevant works of applying convolutional neural network (CNN) to no-reference image quality assessment (NR-IQA) can be broadly divided into two kinds of categories, including patch-wise training and image-wise training. The patch-wise training strategy [10,17,16,27] partitions the image into patches, and then independently predicts the quality of each patch through regression. In contrast, the image-wise training strategy [1,6,12,19] obtains the image quality by aggregating and pooling patch features or predicted scores.

The datasets of 2D/3D video quality assessment (VQA) are smaller than that of IQA. Thus, existing 2D VQA models exploit extracted video features, and then feed these features into the proposed deep learning framework. In

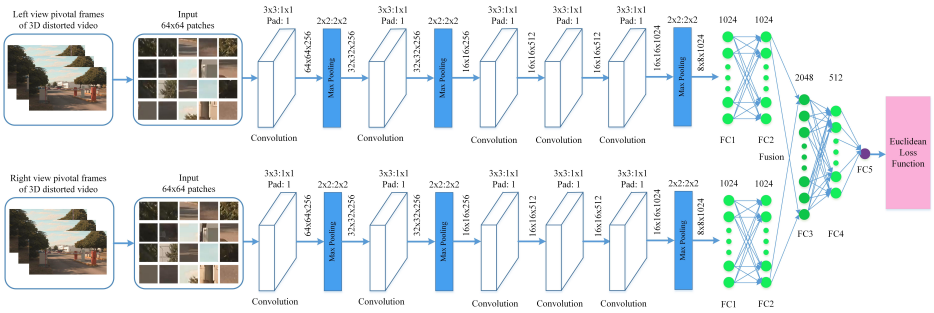


Fig. 1. Architecture of the End-to-end Dual stream deep convolutional Neural network (EDN) for quality regression of image patches from both left and right view videos. There are two sub-networks that all output $1024-dim$ vectors as representations of image patches. These two sub-networks have identical configuration. The predicted score is fed into the final Euclidean loss function for comparison with ground truth quality label. The numbers shown above each arrow give the size of the corresponding output. The numbers shown above each box ($pad = 1$) indicate the size of kernel as well as the size of stride for the corresponding layer.

[15], the 3D shearlet transform was applied to extract spatial-temporal feature from 2D videos and input the feature for 1D CNN to predict the score. In [24], several features were extracted from video stream. The unsupervised restricted Boltzmann machine (RBM) [5] was then employed to predict 2D video quality. These 2D VQA algorithms take hand-crafted features as inputs to deep learning models. However, how to directly input raw image/video to develop an end-to-end DNN architecture for 3D VQA, which integrates both feature extraction and quality regression, has not been proposed.

In this paper, to our best knowledge, this is the first study for the end-to-end learning of stereoscopic video quality. Specifically, we present a DNN-based NR-SVQA approach named End-to-end Dual stream deep Neural network (EDN) to predict the perceptual quality of stereoscopic videos. Our basic idea is that each stereoscopic video has left and right views, which motivates our proposed dual stream network. Moreover, compression distortions introduced in the existing 3D video quality dataset are usually homogeneous which is also consistent in most real application scenarios. Therefore, we can divide the stereoscopic videos into patches and then assign the score of whole stereoscopic video to cropped patches, which solves insufficient training data effectively. Note that both data augmentation and patch preprocessing are not used according to the characteristic of perceptual quality assessment. Extensive experimental results demonstrate that the proposed EDN can effectively predict stereoscopic video quality, in addition achieve highly competitive correlation with human subjective scores compared with many state-of-the-art quality prediction algorithms.

The rest of this paper is organized as follows. In Sect. 2, we present the proposed deep CNN-based NR-SVQA method step by step. Sect. 3 shows the

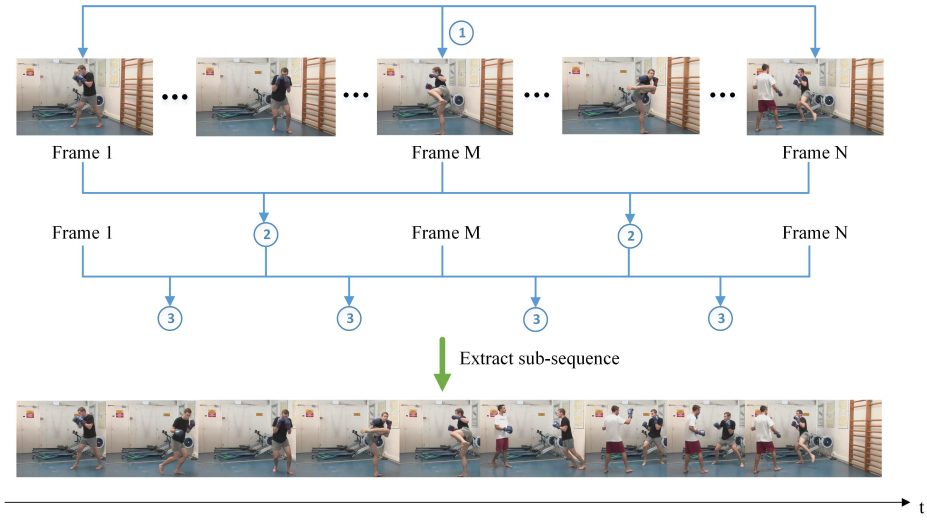


Fig. 2. Demonstration of three-step approach to extract intermediate frame as pivotal frames in temporal domain. The numbers in blue circles represent each step. Frame 1, Frame M and Frame N are the first frame, the intermediate frame and the last frame, respectively.

experimental results and analysis. The conclusion and future work are given in the final section.

2 Proposed Method

The proposed architecture of the End-to-end Dual stream deep convolutional Neural network (EDN) for quality regression is shown in Fig. 1. Due to each stereoscopic video has left and right views, this network contains two sub-networks which are two CNNs with identical configuration and shared parameters. They are designed to automatically learn image visual information through the lower level to higher level feature learning. The image patch pairs for left and right views are taken as inputs to the two sub-networks respectively. The architecture and settings of these two sub-networks are inspired by AlexNet [14] which achieves promising effects in solving image visual related tasks.

For each sub-network, the input to this sub-network is the pixel data of RGB channels of an image patch. Five convolutional layers are then applied to the input image patch. The convolutional layers use convolution kernels (with $stride = 1, pad = 1$), and reduce the size of feature maps only through max pooling. The output of each sub-network is a $1024 - dim$ feature vector (i.e. $FC2$). Then, we utilize a fusion layer for the outputs of these two sub-networks to obtain a $2048 - dim$ feature vector. In other words, we train the left and right

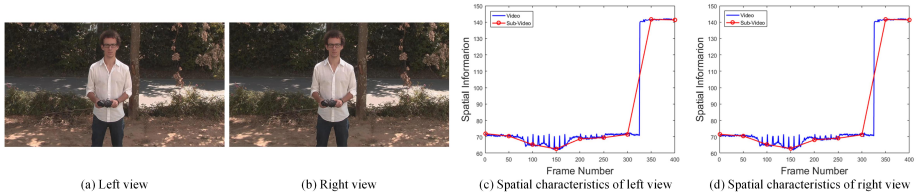


Fig. 3. Spatial characteristics of the whole video frames and the corresponding extracted video sub-sequence. (a) and (b) are the first frames of left and right views for a distorted stereoscopic video; (c) and (d) are spatial information distribution of all frames for the left and right view videos as well as the extracted video sub-sequences corresponding to (a) and (b), respectively.

views separately, and then fuse them to ensure an end-to-end learning as follows:

$$V = [V_l, V_r], \tag{1}$$

where V_l and V_r denote the output feature vectors of left view and right sub-networks, respectively. The fusion layer, which synthesizes the left and right views, may reflect depth perception as the human brain works [20]. Afterwards, two full-connection layers are applied to perform regression onto a single quality score for each input image patch pair. Finally, the predicted score is fed into the Euclidean loss function to be verified by the ground truth quality score.

2.1 Stereoscopic Video Preprocessing

In order to generate more training data to solve the problem of small stereoscopic video datasets and ensure the content diversity and category balance of training data, we preprocess stereoscopic videos as follows.

First, we conduct three-step approach to extract intermediate frames as pivotal frames in temporal domain, as illustrated in Fig. 2. From Fig. 2, we can see that the first frame and the last frame of an input video are preferential, which represent probable video content generally. Suppose for the moment that we extract m frames from the input video, we then compute the remaining frame indices of pivotal frames by:

$$index = \left\lfloor \frac{n}{m-1} \times N \right\rfloor, \tag{2}$$

where $n = 1, 2, ..., m-2$, N is the number of frame for the input video. Specifically, the first step is that the intermediate frame between the first frame and the last frame is picked. Likewise, the second step is that the intermediate frame between the first frame and previous intermediate frame is extracted. The intermediate frame between the previous intermediate frame and the last frame is also picked. Additionally, the third step is also to extract the intermediate frames,

which is similar to the second step. In other words, the frames are extracted equidistantly.

In addition, the spatial characteristics of the whole video frames and the corresponding extracted video sub-sequence can be seen in Fig. 3. In Fig. 3, (a) and (b) show the first frames of left and right views for a distorted stereoscopic videos. Meanwhile, (c) and (d) are the spatial information (SI) [22] distribution of all frames for the left and right view videos as well as the extracted video sub-sequences corresponding to (a) and (b), respectively. In general, we can find that the left and right views of a specific stereoscopic video have similar spatial information distribution. For each view, the image content of each extracted frame in the sub-sequence differs from each other. Moreover, these extracted frames constitute a video sub-sequence, which can represent the input video in a sense.

Second, the extracted pivotal frames are divided into 64x64 non-overlapped patches spatially. Moreover, the image patches are not preprocessed, such as resizing, to maintain the originally perceptual quality. Finally, we input the generated image patch pairs to our proposed EDN.

2.2 Patch Pair-wise Learning

Let a stereoscopic video be represented by left and right view videos V_l and V_r. Each video has N frames. We extract m pivotal frames from the left and right views respectively. For each extracted frame, we then divide it into 64x64 non-overlapped image patches, i.e. Pl_i and Pr_i patches, $i = 1, 2, ..., p$. The predicted quality score for each patch pair of the left and right views is given by the output of EDN with network weights w. Here, the ground truth quality label for each patch pair is assigned the same as the global subjective score of the corresponding stereoscopic video. Our learning objective is defined by mean square error (MSE) as:

$$\min_{w} ||q_i - y_i||_F^2, \tag{3}$$

where q_i, $i = 1, 2, ..., p$ denote the outputs of our proposed EDN, which are the predicted quality scores for p patch pairs in the same image frame. Moreover, y_i, $i = 1, 2, ..., p$ are the corresponding ground truth quality scores for the input image patches. Then, we apply a spatiotemporal pooling strategy to these image patch pairs to estimate the entire stereoscopic video quality. Specifically, the predicted quality score for each frame-pair of the left and right views is computed as follows:

$$Q_{\mathrm{j}} = \frac{1}{p} \sum_{i=1}^{p} q_i, \tag{4}$$

where $j = 1, 2, ..., m$ are the temporal indices of pivotal frames. Finally, the quality of the stereoscopic video is averaged by:

$$Q = \frac{1}{m} \sum_{\mathrm{j}=1}^{m} Q_j, \tag{5}$$

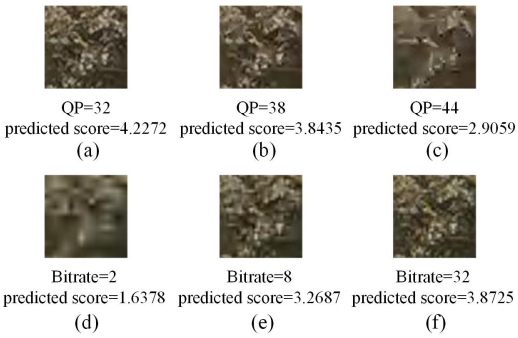


Fig. 4. Illustration of some patch examples for left view and the predicted quality scores for the corresponding image patch pairs. The unit for bitrate is Mb/s.

where Q_j represents the predicted quality for the jth frame, and $j = 1, 2, ..., m$. In our experiment, there are totally nine frames for a given video (i.e. $m = 9$). Then, the patches of extracted frames are exploited to train the EDN model.

3 Experimental Results and Analysis

3.1 Dataset Description and Evaluation Methodology

The NAMA3DS1-COSPAD1 dataset [23] is used in our experiments. This dataset has 10 reference stereoscopic video sequences with a variety of texture, structure, temporal, and depth information. The video sequences have a resolution of 1920x1080 pixels. The frame rate of these videos is 25 fps. The duration of 9 reference videos is 16 seconds, and the remaining one reference video has 13 seconds. In other words, each stereoscopic video has either 400 or 325 frames. Additionally, distortions in the dataset contain H.264 video compression artifacts and JPEG2000 still image compression artifacts. These artifacts are introduced symmetrically to left and right view videos in the NAMA3DS1-COSPAD1 dataset. The H.264 video compression artifacts are produced using JM reference software by varying the quantization parameter (QP) setting as 32, 38, and 44. Moreover, the JPEG2000 still image compression artifacts, which use 2, 8, 16, and 32 Mb/s, are applied to video frames. We can predict the stereoscopic video quality from its pivotal frames and corresponding patch pairs. The ground truth is the mean opinion score (MOS) obtained by human subjective ratings. Two commonly used criteria including Spearman rank-order correlation coefficient (SROCC) and Pearson linear correlation coefficient (PLCC) are adopted for quantitative performance comparison. SROCC is evaluated according to the rank of scores which measures the prediction monotonicity. PLCC is used to evaluate the prediction accuracy. Higher correlation coefficients means better correlation with human quality judgement.

Table 1. Comparison of 2D IQA and 3D VQA state-of-the-art metrics on NAMA3DS1-COSPAD1 stereoscopic video dataset.

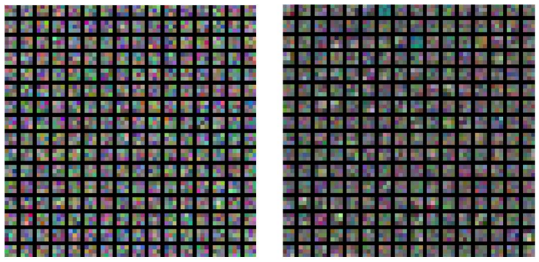
Metrics	SROCC	PLCC
PSNR	0.6470	0.6699
SSIM	0.7492	0.7664
PQM	0.6006	0.6340
PHVS-3D	0.5146	0.5480
SFD	0.5896	0.5965
3D-STS	0.6214	0.6417
SJND-SVA	0.6229	0.6503
Optical flow-based method	0.8552	0.8949
MNSVQM	0.8394	0.8611
BSVQE	0.9086	0.9239
NR CNN method	0.8570	0.8926
Proposed EDN	**0.9334**	**0.9301**

3.2 Performance Comparison

The performance of the proposed metric is conducted on NAMA3DS1-COSPAD1 dataset. Due to the lack of publicly available distorted stereoscopic video datasets, we limit our evaluation on the NAMA3DS1-COSPAD1 dataset. We report the SROCC and PLCC performance values between the obtained quality scores through different metrics and the ground truth MOS for stereoscopic videos. Meanwhile, higher SROCC and PLCC performance values indicate the better agreement with the perceptual quality scores rated by viewers.

Our proposed EDN is compared with two classic 2D IQA metrics which are peak signal to noise ratio (PSNR) and structural similarity (SSIM) [25]. In addition, several 3D VQA algorithms are performed including PQM[9], PHVS-3D[8], SFD[18], 3D-STS[4], SJND-SVA[21], and an optical flow-based method[26]. We also employ another proposed CNN architecture for no-reference image quality assessment (i.e. NR CNN method) [10] and feed left and right view image patches to train the network separately. Then, the predicted quality score can be computed by averaging left and right view scores.

In the experiment, we randomly choose 80% of the reference videos for training and the other 20% for testing. We then obtain 331040 training patch pairs and 77760 testing patch pairs by preprocessing stereoscopic videos. We conduct 100000 iterations in patch pair-wise training stage, and the batch size is 64. The results are shown in Table 1. From Table 1, we can see that our EDN framework outperforms both 2D IQA and 3D VQA state-of-the-art metrics. Furthermore, some patch examples for left view and the predicted quality scores for the corresponding image patch pairs using our proposed EDN can be seen in Fig. 4. From Fig. 4, we can find that the patches with different quality scores, i.e. figures (a-f), are distinguished well through the proposed EDN framework, both for H.264 video compression artifacts and JPEG2000 still image compression artifacts.



(a) Left view (b) Right view

Fig. 5. Visualized learned kernels in the first convolutional layer. (a) Left view; (b) Right view.

The quality scores of patch pairs for each stereoscopic video can be predicted through our trained model, which reaches around 30fps and is suitable for real-time applications. Therefore, our NR-SVQA trained model is accurate, real-time, and adaptable to new video content.

3.3 Visualize Learned Kernel

In order to analyze the characteristics of the trained model regarding to extracting features from the deep learning structure, which is generally taken as an unknown black-box, we visualize the learned kernels in the first convolutional layer.

Fig. 5 depicts the 256 pairs of each 3×3 kernel at the first convolutional layer both for the left view and right view. We can observe that the left and right views of training patches have different spatial texture characteristics since the 3D depth perception exists between these two views and the stereoscopic pairs are fused by the human eye based on spatial correlation in 3D perception [20].

4 Conclusion and Future Work

In this paper, we propose the first study of an end-to-end deep CNN based NR-SVQA framework. The lack of training data for stereoscopic videos is effectively resolved by the stereoscopic video preprocessing method and the spatiotemporal pooling strategy. In addition, the content diversity and category balance of training data are also ensured through the proposed three-step approach to extract pivotal frames from input left and right view videos. Experimental results demonstrate that the proposed method outperforms state-of-the-art approaches. In the future research, the local ground truth target generated for each training patch needs to tackle the non-stationary characteristic of perceptually spatiotemporal quality for stereoscopic videos and more HVS characteristics such as attention mechanism should be taken into consideration.

Acknowledgement. This work was supported in part by the National Key Research and Development Program of China under Grant No. 2016YFC0801001, the National Program on Key Basic Research Projects (973 Program) under Grant 2015CB351803, NSFC under Grant 61571413, 61632001, 61390514, and Intel ICRI MNC.

References

1. Bosse, S., Maniry, D., Wiegand, T., Samek, W.: A deep neural network for image quality assessment. In: Image Processing (ICIP), 2016 IEEE International Conference on. pp. 3773–3777. IEEE (2016)
2. Chen, Z., Zhou, W., Li, W.: Blind stereoscopic video quality assessment: From depth perception to overall experience. IEEE Transactions on Image Processing **27**(2), 721–734 (2018)
3. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 580–587 (2014)
4. Han, J., Jiang, T., Ma, S.: Stereoscopic video quality assessment model based on spatial-temporal structural information. In: Visual Communications and Image Processing (VCIP), 2012 IEEE. pp. 1–6. IEEE (2012)
5. Hinton, G.E., Osindero, S., Teh, Y.W.: A fast learning algorithm for deep belief nets. Neural computation **18**(7), 1527–1554 (2006)
6. Hou, W., Gao, X., Tao, D., Li, X.: Blind image quality assessment via deep learning. IEEE transactions on neural networks and learning systems **26**(6), 1275–1286 (2015)
7. Jiang, G., Liu, S., Yu, M., Shao, F., Peng, Z., Chen, F.: No reference stereo video quality assessment based on motion feature in tensor decomposition domain. Journal of Visual Communication and Image Representation **50**, 247–262 (2018)
8. Jin, L., Boev, A., Gotchev, A., Egiazarian, K.: 3D-DCT based perceptual quality assessment of stereo video. In: 2011 18th IEEE International Conference on Image Processing. pp. 2521–2524. IEEE (2011)
9. Joveluro, P., Malekmohamadi, H., Fernando, W.C., Kondoz, A.: Perceptual video quality metric for 3D video quality assessment. In: 2010 3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video. pp. 1–4. IEEE (2010)
10. Kang, L., Ye, P., Li, Y., Doermann, D.: Convolutional neural networks for no-reference image quality assessment. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1733–1740 (2014)
11. Kavukcuoglu, K., Sermanet, P., Boureau, Y.L., Gregor, K., Mathieu, M., Cun, Y.L.: Learning convolutional feature hierarchies for visual recognition. In: Advances in neural information processing systems. pp. 1090–1098 (2010)
12. Kim, J., Lee, S.: Fully deep blind image quality predictor. IEEE Journal of Selected Topics in Signal Processing **11**(1), 206–220 (2017)
13. Kim, J., Zeng, H., Ghadiyaram, D., Lee, S., Zhang, L., Bovik, A.C.: Deep convolutional neural models for picture quality prediction. IEEE Signal Processing Magazine (2017)
14. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105 (2012)

15. Li, Y., Po, L.M., Cheung, C.H., Xu, X., Feng, L., Yuan, F., Cheung, K.W.: No-reference video quality assessment with 3D shearlet transform and convolutional neural networks. IEEE Transactions on Circuits and Systems for Video Technology **26**(6), 1044–1057 (2016)
16. Li, Y., Po, L.M., Feng, L., Yuan, F.: No-reference image quality assessment with deep convolutional neural networks. In: Digital Signal Processing (DSP), 2016 IEEE International Conference on. pp. 685–689. IEEE (2016)
17. Li, Y., Po, L.M., Xu, X., Feng, L., Yuan, F., Cheung, C.H., Cheung, K.W.: No-reference image quality assessment with shearlet transform and deep neural networks. Neurocomputing **154**, 94–109 (2015)
18. Lu, F., Wang, H., Ji, X., Er, G.: Quality assessment of 3D asymmetric view coding using spatial frequency dominance model. In: 2009 3DTV Conference: The True Vision-Capture, Transmission and Display of 3D Video. pp. 1–4. IEEE (2009)
19. Lv, Y., Yu, M., Jiang, G., Shao, F., Peng, Z., Chen, F.: No-reference stereoscopic image quality assessment using binocular self-similarity and deep neural network. Signal Processing: Image Communication **47**, 346–357 (2016)
20. Parker, A.J.: Binocular depth perception and the cerebral cortex. Nature Reviews Neuroscience **8**(5), 379 (2007)
21. Qi, F., Zhao, D., Fan, X., Jiang, T.: Stereoscopic video quality assessment based on visual attention and just-noticeable difference models. Signal, Image and Video Processing **10**(4), 737–744 (2016)
22. Rec, I.: P. 910: Subjective video quality assessment methods for multimedia applications. Int. Telecomm. Union, Geneva (2008)
23. Urvoy, M., Barkowsky, M., Cousseau, R., Koudota, Y., Ricorde, V., Le Callet, P., Gutierrez, J., Garcia, N.: NAMA3DS1-COSPAD1: Subjective video quality assessment database on coding conditions introducing freely available high quality 3d stereoscopic sequences. In: Quality of Multimedia Experience (QoMEX), 2012 Fourth International Workshop on. pp. 109–114. IEEE (2012)
24. Vega, M.T., Mocanu, D.C., Famaey, J., Stavrou, S., Liotta, A.: Deep learning for quality assessment in live video streaming. IEEE Signal Processing Letters **24**(6), 736–740 (2017)
25. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
26. Yang, J., Wang, H., Lu, W., Li, B., Badiid, A., Meng, Q.: A no-reference optical flow-based quality evaluator for stereoscopic videos in curvelet domain. Information Sciences (2017)
27. Zhang, W., Qu, C., Ma, L., Guan, J., Huang, R.: Learning structure of stereoscopic image for no-reference quality assessment with convolutional neural network. Pattern Recognition **59**, 176–187 (2016)