DEEP NEURAL NETWORKS FOR NO-REFERENCE VIDEO QUALITY ASSESSMENT

Junyong You[1], Jari Korhonen[2]

1. Norwegian Research Centre (NORCE), Bergen, Norway; 2. Shenzhen University, Shenzhen, China

ABSTRACT

Video quality assessment (VQA) is a challenging task due to the complexity of modeling perceived quality characteristics in both spatial and temporal domains. A novel no-reference (NR) video quality metric (VQM) is proposed in this paper based on two deep neural networks (NN), namely 3D convolution network (3D-CNN) and a recurrent NN composed of long short-term memory (LSTM) units. 3D-CNNs are utilized to extract local spatiotemporal features from small cubic clips in video, and the features are then fed into the LSTM networks to predict the perceived video quality. Such design can elaborately tackle the issue of insufficient training data whilst also efficiently capture perceptive quality features in both spatial and temporal domains. Experimental results with respect to two publicly available video quality datasets have demonstrate that the proposed quality metric outperforms the other compared NR quality metrics.

Index Terms—3D-CNN, deep learning, LSTM, video quality assessment

1. INTRODUCTION

With the rapid evolution of digital and communication systems, video services are experiencing tremendous growth in popularity. Although modern video compression and transmission technologies evolve continuously, video signals still suffer quality degradation due to lossy compression and transmission errors in different situations. Consequently, accurate assessment of perceived video quality plays a crucial role in the chain of video services, especially in the context of providing the best Quality of Experience (QoE) to users.

Subjective quality assessment performed by human reviewers is the most reliable way to evaluate the perceived quality of distorted video [1]. However, this approach is expensive in terms of time and effort. Thus, objective quality metrics mimicking subjective quality perception has become a reasonable alternative. Based on the availability of reference video, objective video quality metrics (VQM) can be divided into three categories: full-reference (FR), reduced-reference (RR) and no-reference (NR). Most studies on video quality assessment (VQA) have focused on the FR category, which usually achieves better performance compared with the other two [2]. Considering that humans are the main users of video services, the characteristics of human visual system (HVS), e.g., visual attention mechanism, have been heavily taken into account in video compression, transmission and quality assessment [3][4]. In addition, as a characteristic clue in video presentation, motion information has also been widely employed in VQA [5].

On the other hand, NR quality metrics potentially have broader applicability than FR and RR methods, as a reference video is not always available in certain scenarios, e.g., mobile user generated content with limited transmission bandwidth. Most of the NR video metrics have focused on specific distortion types, e.g., compression or transmission artifacts [6]-[8]. In addition, several works have been proposed using machine learning based on elaborately designed video quality features. In [9], a video quality metric (V-BLIINDS) is proposed to extract spatiotemporal natural scene statistics and motion coherence feature, and then use support vector regression (SVR) for VQA. Similarly, X. Li *et al.* [10] derive spatiotemporal statistics by 3D-DCT transform for leaning based VQA. Based on local descriptors of raw-image-patches and soft-assignment coding, effective image features can be extracted in an unsupervised manner for an image quality metric (CORNIA) [11]. Consequently, VQA metric can be developed by temporally pooling the features extracted by CORNIA over frames [12]. In order to avoid specific-distortion-type issue in most NR metrics, D. Ghadiyaram and A. C. Bovik [13] propose a bag of feature-maps approach for extracting natural scene statistics from images (FRIQUEE). Recently, we have also proposed a learning-based VQM by extracting quality features hierarchically. Low complexity, motion related, features are derived for all video frames first, and then high complexity features, e.g., spatial activity, contrast, noise, sharpness, exposure, will be computed from a subset of representative frames selected in terms of the low complexity features [14].

VQA can essentially be treated as a classification or regression problem, i.e., classify video quality into a category or predict a quality score. Nowadays deep learning dominates image/video classification applications, e.g., convolution neural networks (CNN) has demonstrated outstanding performance in content recognition and object detection [15][16]. Naturally, deep learning should also be applicable in VQA. Several studies have been proposed aiming for applying CNN in image/video quality assessment [17]-[20] and QoE prediction together with other classifiers [21]. M. Giannopoulos *et al*. [17] follow a common architecture of CNN containing 3D-CNN layers, max-pooling layers and fully connected (FC) layers to build an end-to-end video quality model. The V-MEON model combines a 3D-CNN for feature extraction and a codec classifier using FC layers to predict video quality [18]. For image quality assessment, 2D-CNN has been widely employed to extract distortion related features from image patches, and then fed into the followed FC layers for predicting image quality [19][20]. However, even though CNNs have already been successfully applied in image/video content classification and recognition [22][23], their feasibility in VQA is still doubtful. There are two main issues. 1) Can CNNs appropriately capture perceived quality features from video? 2) Can an ordinary VQM dataset provide sufficient data for training a CNN model? Section 2.1 will detail our solutions to these issues.

Furthermore, although 3D-CNN can extract video features in temporal domain to some extent, a particular type of recurrent neural networks (RNN) with long short-term memory (LSTM) units, is better suited to classifying time series data [24]. Therefore, in this work we propose to use 3D-CNN plus LSTM to predict the perceived video quality. 3D-CNNs are mainly served as an extractor of quality related local spatiotemporal features, and then these features are fed into an LSTM regressor predicting the overall video quality. Fig. 1

illustrates the architecture of the proposed metric, and experimental results with respect to two publicly available video quality datasets have demonstrated promising performance of the proposed metric.

The remainder of this paper is organized as follows. Section 2 first explains the motivation of the 3D-CNN plus LSTM based video quality metric, and then presents how to build the metric. Section 3 describes implementation details of the metric and reports the experimental results with respect to two publicly available video quality datasets. Finally, concluding remarks are drawn in Section 4.

2. 3D-CNN PLUS LSTM BASED VIDEO QUALITY ASSESSMENT

2.1. Motivation of using 3D-CNN and LSTM

2D and 3D CNNs have been widely used in image and video classifications, and it has been demonstrated that CNNs can accurately extract features in the context of content classification. However, even though VQA can also be treated as a classification problem, it is definitely different from content classification. Taking a simple example, video sequences with same content can present significantly different quality levels after being compressed at different levels or transmitted through variant error-prone channels. Vice versa, video with different content genres can also represent the same quality level. Thus, it is inappropriate to adopt directly CNNs derived from content classification to VQA. On the other hand, several quality perception studies have demonstrated that the HVS is sensitive to structural changes when evaluating visual quality [25][26]. CNNs are adept at extracting structural features from images, e.g., edges. Therefore, CNNs should also be feasible in visual quality assessment after appropriate training with sufficient amount of data such that the CNN parameters can well represent quality related features.

Most of image/video content classification approaches based on CNNs first need to downscale spatial and temporal resolutions in order to reduce the number of CNN parameters. Such downscaling does not affect the distinguishing characteristics of different contents. However, downscaling can definitely affect quality perception of image/video, e.g., small distortions can become imperceptible after being downscaled. On the other hand, keeping the original video resolution when applying 3D-CNN in VQM creates tremendous amount of memory requirement, making it impossible to train the 3D-CNN model on a normal computer.

Another issue with applying deep learning models in VQA lies in the amount of training data. In video content classification scenario, different subjective annotation datasets can be easily combined together as human subjects usually follow same criteria when manually recognize video contents. However, even though many subjective video quality experiments have already been conducted separately and the produced databases are available to public [27], they cannot be assembled directly as different methodologies and guidelines have been employed. For example, a video with fair quality in a subjective experiment might be considered as a low quality video in another experiment. Thus, it is difficult to create a video quality dataset containing sufficient amount of data for training deep neural networks with large number of parameters.

In order to tackle the above situations, we artificially augment existing video quality datasets by splitting every video from the datasets into multiple small cubic clips and assuming each clip has same quality as the original video. For example, we can extract many cubic video clips of smaller resolutions (e.g., 16 frames × 224 horizontal pixels × 224 vertical pixels × 3 colour channels) from a

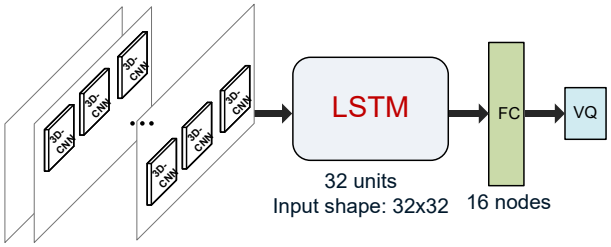

Fig. 1. Illustration of the architecture of the 3D-CNN plus LSTM based video quality metric.

300-frame long standard HD video sequence with a 1920×1080 resolution. We can choose to have or not have spatial and temporal overlaps between the extracted clips. Such an artificial augmentation approach can overcome two issues: 1) avoid downscaling video resolution as the cubic video clips can be fed into a 3D-CNN model directly without changing size; and 2) increase significantly the amount of training data. Whereas, this approach also has a potential problem whether or not it is appropriate to assume that the extracted video clips have same quality as the original video. Therefore, we propose to use a 3D-CNN model as a feature extractor expecting that the model outputs can be served as quality features. The features will be further served as input to another regressor to predict the overall video quality. In this way, even though the output of the 3D-CNN model might not be sufficiently accurate to predict video quality, the regressor will have a chance to adjust quality estimation given by 3D-CNN purely and then provide accurate prediction of the overall video quality.

VQA is actually a time series processing problem, as human perception on video quality at a late timepoint can definitely be affected by previous quality. In addition, the quality features produced by the 3D-CNN model also form time series data. Naturally, an RNN composed of LSTM units, i.e., an LSTM network, is chosen as the regressor to predict the overall video quality. As existing video quality datasets contain relatively small numbers of video sequences, a shallow architecture of the LSTM based regressor has been designed, i.e., one LSTM layer plus another FC layer with 16 nodes as illustrated in Fig. 1.

2.2. 3D-CNN plus LSTM based video quality metric

Following several works on image classification based on deep learning, the spatial input shape of the 3D-CNN is set to 224×224×3. Considering the physical GPU memory on our PC and the descriptive capability in temporal domain, the duration of a video clip is set to 16 frames. In other words, the input of the 3D-CNN is a cubic video clip with 224×224 pixels in 3 colour channels within 16 continuous frames. The selection of video clip resolution might not be universally optimal, but we have experimented with other resolution settings and found this one provides robust results.

Subsequently, four Conv blocks are constructed based on the input in a sequential order, as illustrated in Fig. 2. The first Conv block contains a 3D convolutional layer including 32 filters and the kernel size being (3×3×3) with ReLU as activation, and then followed by a 3D max-pooling layer. In order to preserve quality information in temporal domain, the size of the pooling layer is set to (1×2×2) meaning that pooling is performed across 2×2 spatial pixels while no pooling in temporal domain. This also applies to other pooling layers in the followed Conv blocks. This ensures that temporal quality information is not discarded by the pooling operation. The second Conv block also contains a 3D convolutional layer with

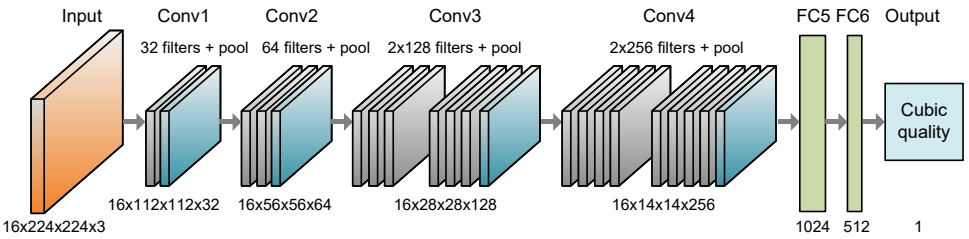

Fig. 2. Illustration of the 3D-CNN model. The numbers below the elements indicate layer output shapes.

64 filters, same kernel size, activation and max-pooling layer. Subsequently, two convolutional layers with 128 filters and same kernel size are contained in the third Conv block. This block also has the same 3D max-pooling layer. The depth (i.e., filter number) of convolutional layers in the fourth Conv block is further increased. Two convolutional layers with 256 filters are employed. In addition, we have found in our training experiment that a smaller kernel size in late phase of CNNs produce better results in VQA. This is probably because the HVS is more sensitive to smaller local spatiotemporal regions in deeper representations of video signals. Thus, the kernel size of convolution in the fourth Conv block is set to (2×2×2). Same as previous blocks, a max-pooling layer with size of (1×2×2) is added in this Conv block.

Finally, two FC layers are added to the Conv blocks. The first FC layer has 1024 nodes followed by a dropout layer with a rate of 0.5, and the second has 512 nodes followed by a similar dropout layer. As the 3D-CNN model is to predict quality of a small video clip, the output layer is set as a single node FC layer with linear activation. Fig. 2 summarizes the architecture of the 3D-CNN model, and indicates the output shape of each block/layer.

As explained in Section 2.1, an LSTM regressor is built based on the output of the 3D-CNN model. The LSTM regressor contains an LSTM layer and another FC layer. The input shape of the LSTM layer is determined by the number of video clips extracted from a video sequence. As mentioned above, it is possible to extract different numbers of video clips from an original video by adjusting the overlap intervals. In our experiments, we first trained the 3D-CNN model and then evaluated different settings of parameters of the LSTM regressor with respect to the training data from video quality datasets. We found that the parameter size combination of (32×32×32) provides the best results in our experiments. More explicitly, the first 32 denotes the unit number of the LSTM layer, and the other 32×32 (time-steps × data dimension) is the shape of the input to the LSTM layer. In other words, a fixed number of 32×32 cubic video clips should be extracted from every video sequence, and each clip has a single output of clip quality. In order to achieve the fixed number of video clips, different overlap intervals have been employed for video sequences with different spatial resolutions and frame numbers. In practice, a video sequence needs to be first divided into 32 groups of frames, and each group contains 16 frames, and the overlap interval between two adjacent groups can be determined by video frame number. Subsequently, 32 blocks in each frame in every group are extracted, and the block size is 224×224. The horizontal and vertical overlap intervals between neighbouring blocks are determined by spatial resolution of video frame. Consequently, 16 blocks from the 16 frames in a group are concatenated temporally to form a cubic video clip with size of 16×224×224(×3 colour channels).

After the LSTM layer, a FC layer with ReLU activation is added and it has been found that 16 nodes of this layer can provide the best results from our experiments. Finally, the output layer is set as a FC layer with a single node to predict the overall quality of a video sequence.

3. EXPERIMENTS

3.1. Data generation and model training

In order to train the proposed deep neural networks and evaluate the performance, two publicly available video quality datasets are used, including KonViD-1k database [28] and LIVE-Qualcomm mobile in-capture database [29]. The KonViD-1k dataset contains 1,200 video sequences originally from Flickr, with resolution of 960×564 and variant frame rates, diverse contents and distortion types. Video quality was assessed by 642 subjects, using the crowdsourcing methodology. Due to its diversity and large number of video sequences, KonViD-1k dataset is suitable for evaluating deep learning based VQMs. The LIVE-Qualcomm dataset contains 208 video sequences with resolution of 1920×1080 and frame rate of 30 frames per second, suffering from mobile in-capture distortions. Considering that the selection of the contents, distortions, and methodologies for the two datasets are different, model training and evaluation have been performed separately. In the KonViD-1k dataset, 90% of the video sequences have been randomly chosen from the dataset for training the networks; while 80% of video sequences from the LIVE-Qualcomm dataset were randomly chosen as the training set in order to assure sufficient evaluation data in this dataset.

As the LSTM networks in the proposed VQM actually depend on the outputs of the 3D-CNN model, the two networks are trained independently. The 3D-CNN model was trained first, and then followed by LSTM training. A no-spatial-overlap approach has been employed to generate the cubic video clips from each video sequence in the training sets, whilst the temporal overlap interval was set to 8. For example, a video frame from the KonViD-1k dataset can be divided into 8 blocks with 224×224 pixels, and the boundary pixels (64 in horizontal and 116 in vertical) are discarded. Assuming this video sequence has 32 frames, in total 24 (8×3) cubic clips can be extracted from this video. If the frame number of a video is not divisible by 8, the remainder frames in video beginning are discarded. In this way, a total of 224,640 cubic video clips from the KonViD-1k training set have been generated. The same approach was performed on the training set of LIVE-Qualcomm dataset and in total 292,864 cubic clips have been generated.

Subsequently, the 3D-CNN model training, including architecture and parameter tuning, have been performed on the training sets of the two dataset separately. As explained in Section 2.2, the size selection of the cubic clips, node numbers of the Conv blocks and FC layers have both been optimized based on the training sets. Finally, the best combination of model architecture and parameters obtained in the training was chosen as the 3D-CNN model and then further trained through 50 epochs to obtain the optimum parameters with the cost of 5 days training on our computer with two GPU.

Table I. Evaluation results of video quality metrics with respect to KonViD-1k and LIVE-Qualcomm evaluation sets

Methods	KonViD-1k Evaluation Set			LIVE-Qualcomm Evaluation Set		
	PCC	SRCC	RMSE	PCC	SRCC	RMSE
V-BLIINDS	0.591	0.600	0.493	0.686	0.560	9.005
V-CORNIA	0.720	0.687	0.419	0.566	0.405	10.202
FRIQUEE	0.775	0.766	0.380	0.756	0.642	8.368
Pure 3D-CNN	0.781	0.771	0.376	0.780	0.645	7.790
3D-CNN + LSTM	***0.808***	***0.800***	***0.365***	***0.792***	***0.687***	***7.740***

In addition, the LSTM training was performed on video sequence level in the two training sets. Considering that the KonViD-1k dataset contains significantly more video sequences than LIVE-Qualcomm, LSTM training is performed on the KonViD-1k training set first, and the trained weights were used as initial weights to fine tune the LSTM weights with respect to the LIVE-Qualcomm training set by transfer learning. It should be noted that the LSTM model requires fixed number of inputs, i.e., fixed number of extracted cubic clips, and therefore, different overlap intervals have been employed for video sequences with variant resolutions and frame numbers.

3.2. Experimental results

In order to evaluate the performance of the proposed metric, three other learning based video quality metrics have been included for fair comparison as benchmarks, namely V-BLIINDS [9], V-CORNIA [11][12] and FRIQUEE [13]. It is worth noting that V-CORNIA and FRIQUEE both extract features from images, and therefore average temporal pooling has been applied to the image features to derive the video features. In our earlier work [14], we found that sampling rate of 1 frame/second is sufficient for V-CORNIA and FRIQUEE. SVR was chosen as regressor for the three metrics to predict video quality, as it provides better results than other regressors (e.g., random forest) in our experiments. We used the SVR hyper-parameters fine-tuned separately for each method and dataset [14]. In addition, the 3D-CNN model can predict the quality of a cubic video clip. If the quality values of all cubic clips extracted from a video are pooled, it can also provide an estimation of the overall video quality. In our experiments, the same approach as in generation of training data has been employed to extract cubic clips from video sequences in the evaluation set. Consequently, the average of the quality values of all the cubic clips is assumed as the overall video quality. This approach is named as pure 3D-CNN in our experiments.

In order to compare different metrics fairly, the training of all the included metrics have been conducted on the same training sets as explained in Section 4.1, and the performance was then evaluated with respect to the remainder video sequences. Three numerical measures, including Pearson linear correlation coefficient (PLCC), Spearman rank order correlation coefficient (SRCC) and root mean squared error (RMSE), have been chosen as evaluation criteria.

Table I reports the results of the video quality metrics with respect to the evaluation sets of KonViD-1k and LIVE-Qualcomm datasets, respectively. According to the evaluation results, the deep learning based metrics outperform other benchmark metrics, even though the 3D-CNN model without LSTM can provide better prediction of video quality. As a comparison, we have also tested utilizing the proposed 3D-CNN model at video sequence level by downscaling video resolution. In other words, a video sequence was downscaled to the size of 16×224×224 (×3 colour channels) and then fed to the 3D-CNN model, similar to other 3D-CNN based video classifications. However, the performance was very poor, e.g. PCC=0.21. Potential reasons can be the impact of downscaling to the perceived quality and insufficient training data, as explained in Section 2.1. This confirms that appropriate handling of video data can definitely improve the applicability of deep neural networks for quality assessment.

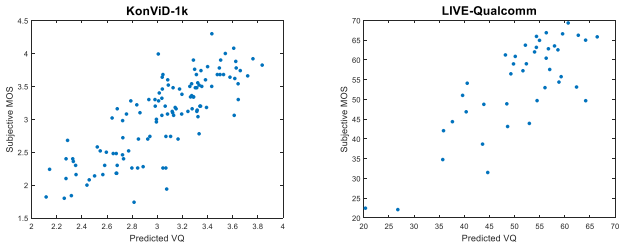

Fig. 3. Scatter plots of the proposed 3D-CNN + LSTM based video quality metric.

Furthermore, the proposed metric with LSTM network using the features extracted by 3D-CNN offers the best prediction of video quality. This is due to two reasons: 1) descriptive capability of the 3D-CNN model for video quality perception; 2) LSTM can better represents the characteristics of the quality features in time series than simply averaging them over space and time. Fig. 3 shows scatter plots of the predicted video quality values with respect to the subjective quality scores of the two evaluation sets. In addition, by comparing the pure 3D-CNN model and the proposed metric across the two datasets, it can be observed that the LSTM model improves quality assessment based on 3D-CNN model more significantly in the KonViD-1k dataset than the LIVE-Qualcomm dataset. We believe this is because the KonViD-1k dataset contains more video sequences that can better train the LSTM model.

However, there is still an issue with the proposed metric. Because the metric includes 3D-CNN and LSTM as two separate models, it is not possible to train them in an end-to-end manner. In the future work, we plan to develop a new architecture by adopting the idea of full convolution networks (FCN) to build fully 3D convolution networks, employing 1×1×1 3D convolution to replace the FC layers, and then adding the LSTM layer on top. In this way, we expect to achieve an end-to-end trainable model.

4. CONCLUSIONS

This work proposes an accurate NR video quality metric based on two deep neural networks, namely 3D-CNN and LSTM. 3D-CNN serves as a feature extractor to derive the quality values of small cubic video clips. The clip quality values are then employed as input to another LSTM model to predict the overall video quality. In this way, the proposed metric can overcome several issues in applying deep learning in VQM directly, e.g., resolution downscaling, amount of training data, time series processing. The evaluation results with respect to two publicly available video quality datasets have demonstrated that the proposed metric significantly outperforms the benchmark NR quality metrics representing the state of the art.

5. REFERENCES

[1] M. H. Pinson, L. Janowski, R. Pepion, Q. Huynh-Thu, C. Schmidmer, P. Corriveau, A. Younkin, P. Le Callet, and M. Barkowsky, "The influence of subjects and environment on audiovisual subjective tests: An international study," *IEEE J. Select. Topics Signal Process.*, vol. 6, no. 6, pp. 640–651, Oct. 2012.

[2] J. You, U. Reiter, M. M. Hannuksela, M. Gabbouj, and A. Perkis, "Perceptual-based objective quality metrics for audio-visual services - A survey," *Signal Process. Image Commun.*, vol. 25, no. 7, pp. 482-501, 2010.

[3] J. You, T. Ebrahimi, and A. Perkis, "Attention driven foveated video quality assessment," *IEEE Trans. Image Process.*, vol. 23, no. 1, pp. 200-213, Oct. 2013.

[4] M. Pinson, and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Trans. Broadcast*., vol. 50, no. 3, pp. 312-322, Sep. 2004.

[5] K. Seshadrinathan, and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE Trans. Image Process.*, vol. 19, no. 2, pp. 335-350, Feb. 2010.

[6] G. Valenzise, S. Magni, M. Tagliasacchi, and S. Tubaro, "No-reference pixel video quality monitoring of channel-induced distortion," *IEEE Trans. Circuits and Syst. for Video Tech*., vol. 22, no. 4, pp. 605-618, Apr. 2012.

[7] Z. Chen and D. Wu, "Prediction of transmission distortion for wireless video communication: Analysis," *IEEE Trans. Image Process*., vol. 21, no. 3, pp. 1123-1137, Mar. 2012.

[8] J. Søgaard, S. Forchhammer and J. Korhonen, "No-reference video quality assessment using codec analysis*," IEEE Trans. Circ. and Syst. for Video Tech.*, vol. 25, no. 10, pp. 1637-1650, Oct. 2015.

[9] M. A. Saad, and A. C. Bovik, "Blind prediction of natural video quality." *IEEE Trans. Image Proc*., vol. 23, no. 3, pp. 1352-1365, Mar. 2014.

[10] X. Li, Q. Guo, and X. Lu, "Spatiotemporal statistics for video quality assessment," *IEEE Trans. Image Proc*., vol. 25, no. 7, pp. 3329-3342, May 2016.

[11] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, Providence, RI, USA, 2012.

[12] J. Xu, P. Ye, Y. Liu, and D. Doermann, "No-reference video quality assessment via feature learning," in *Proc. IEEE Int. Conf. Image Proc. (ICIP)*, Paris, France, 2014.

[13] D. Ghadiyaram and A. C. Bovik, "Perceptual quality prediction on authentically distorted images using a bag of features approach," *J. of Vision*, vol. 17, no. 1, pp. Jan. 2017.

[14] J. Korhonen, "Hierarchical approach for no-Rrference consumer video quality assessment," submitted to *IEEE Trans. Image Proc.*, under review, 2018. Preprint available online: https://github.com/jarikorhonen/nr-vqa-consumervideo/blob/master/TIP_hiviqum_github.pdf.

[15] Y. LeCun, Y. Benjio, and G. Hinton, "Deep Learning," *Nature*, vol. 521, pp. 436-444, May 2015.

[16] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and F-F. Li, "Large-scale Video Classification with Convolutional Neural Networks," in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, Jun. 2014, Columbus, OH, USA.

[17] M. Giannopoulos, G. Tsagkatakis, S. Blasi, F. Toutounchi, A. Mouchtaris, P. Tsakalides, M. Mrak, and E. Izquierdo, "Convolutional neural networks for video quality assessment," Submitted to *Signal Process. Image Commun.,* Sep. 2018, Preprint available online: https://arxiv.org/abs/1809.10117.

[18] W. Liu, Z. Duanmu, and Z. Wang, "End-to-end blind quality assessment of compressed videos using deep neural networks," in *Proc. ACM Int. Conf. Multimedia*, Oct. 2018, New York, USA.

[19] S. Bosse, D. Maniry, K-R. Müller, T. Wiegand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Trans. Image Proc*., vol. 27, no. 1, pp. 206-219, Oct. 2017.

[20] S. Bianco, L. Celona, P. Napoletano, and R. Schettini, "On the use of deep learning for blind image quality assessment," *Signal, Image and Video Processing*, vol. 12, no. 2, pp. 355-362, Feb. 2018.

[21] M. Lopez-Martin, B. Carro, J. Lloret, S. Egea, and A. Sanchez-Esguevillas, "Deep learning model for multimedia Quality of Experience prediction based on network flow packets," in *IEEE Communication Mag*., vol. 56, no. 9, pp. 110-117, Sep. 2018.

[22] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet?" in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, 2018.

[23] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, pp. 4489-4491, Dec. 2015.

[24] J. Y-H. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 4694-4702, 2015.

[25] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Proc*., vol. 13, no. 4, pp. 600-612, Apr. 2004.

[26] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Trans. Image Proc*., vol. 20, no. 8, pp. 2378-2386, Jan. 2011.

[27] Image and Video Quality Resources, available online: https://stefan.winkler.site/resources.html.

[28] V. Hosu, F. Hahn, M. Jenadeleh, H. Lin, H. Me,T. Szirányi, S. Li, and D. Saupe, "The Konstanz natural video database (KoNViD-1k)," in *Proc. 9th Int. Conf. Quality of Multimedia Experience (QoMEX)*, May 2017.

[29] D. Ghadiyaram, J. Pan, A. C. Bovik, A. K. Moorthy, P. Panda, and K. C. Yang, "In-capture mobile video distortions: A study of subjective behavior and objective algorithms," *IEEE Trans. Circ. and Syst. for Video Tech.*, vol. 28, no. 9, Sep. 2018.