No-reference video quality measurement: added value of machine learning

No-reference Video Quality Measurement: The Added Value of Machine Learning

Decebal Constantin Mocanu[a], Jeevan Pokhrel[b], Juan Pablo Garella[c], Janne Seppänen[d], Eirini Liotou[e], Manish Narwaria[f]

[a] ECO group, Department of Electrical Engineering, Eindhoven University of Technology, Eindhoven, the Netherlands, e-mail: (d.c.mocanu@tue.nl)

[b] Montimage, Paris, France, e-mail: (jeevanpokhrel@gmail.com)

[c] Facultad de Ingeniería, Universidad de la República, Montevideo, Uruguay, e-mail: (jpgarella@fing.edu.uy)

[d] Network Performance Team, VTT Technical Research Centre of Finland Ltd., Finland, e-mail: (janne.seppanen@vtt.fi)

[e] GAIN group, National & Kapodistrian University of Athens, Greece, e-mail: (eliotou@di.uoa.gr)

[f] DA-IICT, Gandhinagar, India, e-mail: (manish_narwaria@daiict.ac.in)

Abstract. Video quality measurement is an important component in the end-to-end video delivery chain. Video quality is, however, subjective and thus there will always be inter-observer differences in the subjective opinion about the visual quality of the same video. Despite this, most existing works on objective quality measurement typically focus only on predicting a single score, and evaluate their prediction accuracies based on how close it is to the mean opinion scores (or similar average based ratings). Clearly, such an approach ignores the underlying diversities in the subjective scoring process, and as a result, does not allow further analysis on how reliable the objective prediction is in terms of subjective variability. Consequently, the aim of this paper is to analyze this issue and present a machine learning based solution to address it. We demonstrate the utility of our ideas by considering the practical scenario of video broadcast transmissions with focus on Digital Terrestrial Television (DTT), and proposing a no-reference objective video quality estimator for such application. We conducted meaningful verification studies on different video content (including video clips recorded from real DTT Broadcast transmissions) in order to verify the performance of the proposed solution.

Keywords: no-reference video quality assessment, deep learning, subjective studies, objective studies, quality of experience.

Address all correspondence to: Decebal Constantin Mocanu, Eindhoven University of Technology, ECO group, Department of Electrical Engineering, Den Dolech 2, Eindhoven, the Netherlands, 5612AZ; Tel: +31 40-247 5394; E-mail: d.c.mocanu@tue.nl

1 Introduction

With the ever increasing demand for video services and applications, real-time video processing is one of the central issues in multimedia processing systems. Given the practical limitations in terms of resources (bandwidth, computational power, memory etc.), video signals need to be appropriately processed (e.g. compressed) to make them more suitable for transmission, storage and subsequent rendering. However, most of the mentioned processing will degrade the visual quality to varying extents. As a consequence, the end user may view a significantly modified video signal in comparison to the original source content. It is, therefore, important to measure the quality of the processed video signal and benchmark the performance of different video processing algorithms in terms of video quality assessment. Video quality is essentially a component of the larger concept of Quality of Experience (QoE). It is therefore an intrinsically subjective measure and can depend on multiple factors including degree of annoyance (related to artifact visibility), aesthetics, emotions, past experience etc.[1] Thus, subjective viewing tests remain the most reliable and accurate methods, given appropriate laboratory conditions and a sufficiently large subject panel. However, subjective assessment may not be feasible in certain situations (e.g. real-time video compression, transmission), and an objective approach is more suitable in such scenarios. While the performance of objective approaches may not accurately mimic the subjective opinion, it can still potentially provide approximate and relative estimates of video quality, in a given application.

Objective quality estimation can be classified into three groups, i.e. Full-Reference (FR), Reduced-Reference (RR) and No-Reference (NR)[2] , as detailed in Table 1. Among them, NR estimation is more challenging since it relies only on the processed signal. As a result, it is more related to detection and quantification of certain irregularities or absence of specific features which would be typically found in the reference video. It can also exploit application-specific features (e.g. bit rate) from the video bit stream in order to quantify quality, and there are existing works to this end, as discussed in the next section. Subjective estimation of video quality, on the other hand, involves a number of human observers rating the video quality on a fixed pre-defined scale, typically in controlled laboratory conditions. Excellent treatment of the various factors in video quality assessment is readily available in the form of standards and recommended practices.[3]

Pre-printed version.

Please cite as: *Mocanu D.C., Pokhrel J., Garella J.P., Seppänen J., Liotou E., Narwaria M.; No-reference video quality measurement: added value of machine learning. J. Electron. Imaging. 0001;24(6):061208. doi:10.1117/1.JEI.24.6.061208.*

Table 1 Description of video QoE objective estimation categories.

	Full-Reference (FR)	**Reduced-Reference (RR)**	**No-Reference (NR)**
Reference video	The reference video is available.	Only some information (e.g. metrics) extracted from the reference video are required.	No information from the reference video is required.
Methodology	The quality is estimated based on a comparison between the reference and a processed video.	The quality is estimated based on the information extracted from the reference video and a processed video.	The quality is estimated based just on some information extracted from a processed video.
Accuracy (in general)	Higher than RR and NR.	Higher than NR. Lower than FR.	Lower than FR and RR.

An important aspect of any subjective study is the underlying variability in the collected ratings. This happens because the same stimuli typically do not receive the same rating by all the observers. This is of course expected since the notion of video quality is highly subjective, and this injects certain variability or inter-observer differences in the stimuli rating. While these are generally reported in subjective studies (in the form of standard deviations, confidence intervals etc.), a survey of literature reveals that they are not typically accounted for in objective quality prediction. As a result, a majority of works on objective quality estimation focus only on predicting a single score that may represent an average of all the ratings per stimuli. Further, the prediction accuracies of objective methods are generally based on how close the objective scores are to the averaged subjective ratings (this is generally quantified by correlation coefficients, mean square error, scatter plots, etc.). However, the inherent subjective variability and its impact are not directly taken into account. This may potentially reduce the reliability of the objective estimates especially when there is larger disagreement (high variability) among subjects on the quality of a certain stimuli. Therefore, the aim of this paper is to analyze this issue in more details, and subsequently present a NR video quality assessment method based on that. The presented approach is based on defining a reasonable measure of subjective data diversity and modeling it through the paradigm of machine learning.

The remainder paper is organized as follows. Section II first provides a brief review of machine learning based NR video quality measurement methods, and also outlines their limitations. We also present our contributions in this section. Analysis of the importance of diversity in subjective rating process is presented in Section III. The proposed method and its application within a practical scenario is explained in Section IV while its experimental verification has been reported in Section V. The next section presents relevant discussion about the results while section VII concludes the paper.

2 Background and Motivation

2.1 Previous work

Even the research in NR video quality assessment is more than a decade old, we are still far from a general purpose NR quality indicator that can accurately predict video quality in all situations. The authors in[4] presented one of the first comprehensive method for estimating video quality based on neural networks. In this work, a methodology using Circular Back Propagation (CBP) neural networks is used for the objective quality assessment of motion picture expert group (MPEG) video streams. The work in[5] employed Convolutional Neural Networks (CNN) in order to estimate video quality. It differs from conventional neural network approach since it relies on the use of CNNs that allows a continuous time scoring of the video. A NR method was presented in,[6] which is based on mapping frame level features into a spatial quality score followed by temporal pooling. The method developed in[7] is based on features extracted from the analysis of discrete cosine transform (DCT) coefficients of each decoded frame in a video sequence, and objective video quality was predicted using a neural network. Another NR video quality estimator was presented in,[8] where symbolic regression based framework was trained on a set of features extracted from the received video bit-stream. Another recent method in[9] works on the similar principle of analyzing several features. These are based on distinguishing the type of codec used (MPEG or H.264/AVC), analysis of DCT coefficients, estimation of the level of quantization used in the I-frames etc. The next step is to apply Support Vector Regression (SVR) to predict video quality in NR fashion. The NR method proposed in[10] was based on polynomial regression model, where the independent variables (or features) were based on spatial and temporal quantities derived from video spatio-temporal complexity, bit rate and packet loss measurements. The works mentioned here by no means constitute the entire list of contributions on the topic of NR video quality measurement but merely represent the most recent and relevant for the purpose of this paper. The reader is encouraged to refer to survey papers, for example.[11]

2.2 Limitations of existing methods

As mentioned, there has already been significant research work on NR video quality estimation especially for video compression applications. However, most of these methods share three common limitations related to their design and validation as enlisted below:

- Most of these methods rely only on mean opinion scores (MOS) or degradation MOS (DMOS) both for training and validation. This, to our mind is problematic since the MOS or DMOS (obtained by averaging raw scores for each observer) tend to neglect the variability inherently present in the subjective rating process.
- Most of these methods have been validated only on limited set of videos and lacked a comprehensive method evaluation from the viewpoint of its robustness to untrained content.
- Lastly, a majority of existing work focus only on video compression. Thus, they would be limited in their applicability to other applications (e.g. video transmission) where the fully decoded video content may not be available and so quality must be predicted only from the bit stream information.

2.3 Our contributions

In this paper, we aim to address the limitations mentioned above. Thus, our main contribution is to perform statistical analysis on the performance of various machine learning methods (e.g. linear regression,[12] decision trees for regression,[12] random neural networks,[13] deep belief networks[14]) in predicting video quality on a real-world database.[15] More specifically, in contrast to most of the existing works on NR video quality estimation, we focus on three aspects that have been largely ignored.

First, we model the diversity that inevitably exists in any subjective rating process, and we analyze statistically its relation with MOS. Thus, we attempt to take into account inter-observer differences since it will help in a better interpretation of how reliable the objective quality score is and what it conveys about the user satisfaction levels. Such an approach also adds significant value from a business perspective when it comes to telecom operators or internet service providers (ISPs), as will be further analyzed in the next section. Thus, in the proposed approach, we do not just train our method in an effort to maximize correlations with the average ground truth, but simultaneously allow the algorithm to learn the associated data variability. To our knowledge, this is the first work towards the design of an application-specific NR video quality estimator, which can provide additional output that can help to understand the meaning of the objective score under a given application scenario. The presented analysis will be therefore of interest to the QoE community, which has largely focused only on MOS as the indicator of subjective video quality.

Secondly, we exploit the promising deep learning framework in our method and demonstrate its suitability for the said task, while we assess its prediction performance against three widely used machine learning algorithms and two statistical methods. Specifically, deep networks can benefit from unsupervised learning thus requiring less training data in comparison to the traditional learning methods. An analysis pertaining to the training of the deep networks weights is also presented to provide insights into the training process.

Finally, we focus on meaningful verification of the proposed method on several challenging video clips within the practical framework of DTT, which help to evaluate the proposed method against diverse content and distortion severities. We highlight that half of the video clips used for experiments (i.e. 200) come from a real-world video delivery chain with impairments produced by a real video transmission system and not produced by noise added artificially, thus representing a realistic scenario.

3 Exploring Diversity in Subjective Viewing Tests

It can be seen that a vast majority of objective studies rely only on the mean or average (MOS or DMOS) of the individual observer ratings. As we know, simple arithmetic mean is a measure of the central tendency but it tends to ignore the dispersion of the data. Expectedly, simple averaged based ratings have been contested in literature as they result in an information loss of how opinions of subjective assessment participants deviate from each other. The authors of[16] argue against averaging subjective judgments and suggest that taking into account the diversity of subjective views increases the information extracted from the given dataset. Authors of[17] apply this principle in their QoE study, where in addition to MOS a standard deviation of opinion scores (SOS) is studied. The mathematical relation between MOS and SOS is defined, and several databases for various applications are analyzed using SOS in addition to average user ratings.

3.1 Scattering of subjective opinions

The subjective tests remain the most reliable approach to assess human factors such as degree of enjoyment (video quality). Still, expectedly some amount of inherent subjectivity will always be injected into the data collected from such studies. This can be attributed to several factors including the viewing strategy (some observers make decisions instinctively based on abstract video features while others may base their decision on more detailed scene analysis), emotions, past experience etc. For video quality evaluation, this means that while the individual observer ratings may indicate a general trend about perceived quality, they may still differ/disagree on the magnitude of such an agreement. Such diversity can provide valuable information that can be exploited for specific applications. However, before that, it is necessary to quantify the said diversity (scattering) meaningfully and not merely rely on averaged measures such as MOS.

The deviation of individual ratings from the mean can for instance provide a measure of the spread i.e. standard deviation (SOS). Another related measure is the confidence interval which is derived from standard deviation and also depends on the number of observers. These have been often reported in subjective studies involving video quality measurement. But using these measures to supplement for objective quality prediction is not always interpretable in a stand alone fashion. For example, simply providing a standard deviation along with a predicted objective score does not allow a clear interpretation of what it may mean in the context of an application. This is partly due to the mathematical relation between MOS and standard deviation (high or low MOS always results in small deviation), and also because standard deviation does not indicate skewness of opinions scattered around the average value. Hence, it may be desirable to devise a more interpretable measure of quantifying the diversity of subjective opinion and more importantly what it may mean in the context of a particular application.

3.2 A new measure to quantify subjective uncertainty

It is known that low MOS for a given service indicates bad quality and therefore disappointment to the service, but even if MOS is high, we cannot know from this single value how many users are actually dissatisfied with the service. Moreover, not only do negative experiences affect customers more than positive experiences, but customers are also prone to share their negative experiences more likely than positive ones. Therefore we could see a negative experience of a single user to have a risk of avalanche where the negative experience is spread to several other current/potential customers who will see the service in a more negative light than before, without actually having bad experience with the service. As already highlighted, a majority of objective methods simply ignore the diversity of user opinions, and instead focus only on average ratings as their target. To overcome this, we first need to define a plausible way in order to exploit data uncertainty so that it adds value to the objective quality prediction. To that end, we studied various methods for expressing the diversity, and considering a business-oriented application, we found that an appropriate measure of profitability (which is of course the key goal of any business) can be derived from the answer to question "how many users are unsatisfied with the service". From service management and business point, satisfied users are less interesting than dissatisfied users. This is due to the fact that, from quality perspective, satisfied users require no quality management for their service (although this is not to say that satisfied users should not be considered at all in overall service marketing).

MOS is a straightforward indicator for expressing the opinion of a majority of users, but as discussed, this is hardly enough if we want to maintain service reputation and hold on to the current customer base. Therefore we introduce a new indicator along with MOS - Percentage of Dissatisfied Users (PDU) against MOS. It indicates the percentage of users who would give an opinion score less than certain threshold given a certain MOS score, i.e.

$$PDU = \frac{\#(OS < th)}{N} \times 100 \tag{1}$$

where OS denotes the opinion score from an individual observer, th is the user-defined threshold, N is the total number of observers evaluating the given condition (service quality).

As an example, let us consider that 3 independent and random observers evaluated a sample (video stimulus) and gave scores 2, 5, and 5 (on a scale from 1 to 5, 5 denotes excellent quality). We can quickly calculate the MOS for this sample as 4, which is a fairly good score considering the defined scale of evaluation in this case. But we note that one individual gave a score of 2, which is very poor. Consequently, we can conclude that 33% of users were not satisfied (i.e. $PDU = 33\%$) with this sample, despite the MOS being high. It is therefore easy to realize the limitation

Assuming these tests are conducted in proper viewing conditions (controlled lighting, well defined viewing distance/angles etc. for the considered application) and with a sufficiently large subject panel.

Table 2 MOS scores provided by 25 observers to a particular video clip.

2	5	4	4	4
4	3	3	2	2
3	2	4	2	4
3	5	3	5	4
4	3	5	2	5

of average based ratings (even with this somewhat limited example) where the MOS would conceal the fact that not all users were happy with the sample (despite a high MOS). We can also observe such effects on real subjective data shown in Table 2. It represents the individual subjective opinion scores of 25 observers (this was as part of a subjective study conducted in our lab) for a processed video. We note that the mean of these individual ratings is 3.48 which is in the higher range (the scale of rating was from 1 to 5), and may lead to conclude that the video quality would be generally at least acceptable. Still, we note that $PDU = 24\%$ (when mean is considered as th) meaning that almost one-fourth of the customers/observers were dissatisfied with the video quality. This information should then be used to devise corrective actions. It can also be seen that the definition of PDU depends on the free parameter th, and hence it can be set by the service provider. This would depend on what quality level is considered intolerable and the actions required to avoid customer churn. In this paper, we selected a value of 3, i.e. $th = 3$ (assuming a scale from 1 to 5), but especially for commercial applications where customers pay a monthly fee or pay per view, this number could be even higher. Hence, it can be customized.

Before we conclude this section, it is important to mention that the proposed measure PDU may not always be a function of MOS nor it may be directly related to standard deviation of the individual subjective ratings. So one cannot assume that a higher MOS will imply lower PDU or a lower MOS always implies a larger PDU. The reason is that different quality degradations may have different impacts on the consistency of user opinions. We can easily understand this with our previous example, where scores 2, 5, and 5 lead to a MOS of 4. However, we may have the same MOS in another situation. For instance if the scores were 4, 4, and 4, the resultant MOS would still be 4 but $PDU = 0$ in this case. Also, standard deviation may not be a substitute for PDU for two reasons. First, as already stated the former may not be interpretable in a stand alone manner. Second, standard deviation can be similar for two very different MOSs in which case it does not provide any information on possible corrective measures. In contrast, similar PDU for two different MOSs may indicate a course correction (if PDU is high) irrespective of the MOS.

4 Application in NR Video Quality Estimation

In this section we demonstrate the practical utility of the proposed method in a NR scenario, within the framework of Digital Terrestrial Television. The proposed method follows similar design philosophy as some of the existing methods but there are some important differences that add value to our proposal. First, we exploit the framework of deep learning methods, which to our knowledge has not been exploited towards NR video quality measurement. Specifically, in the considered application, it is assumed that source video data is not available and quality needs to be predicted only from coded stream information. Secondly, our method is trained to provide PDU values in addition to objective quality. This allows the user to better interpret the reliability of the objective prediction especially from the viewpoint of satisfied/dissatisfied user percentage.

A block diagram of the proposed approach is shown in Figure 1. Note that in the DTT scenario there can be multiple TV channels broadcasting signals over the air and these signals are pre-processed (source and channel coded) before transmission. Also the wireless channel (air) is ideally not transparent and hence will introduce errors in the relayed bitstream. All these will show up as spatio-temporal artifacts in the video that will be rendered to the end user. In order to model what the end user perceives regarding the quality of the rendered videos, we first extract features from channel streams and then develop a model based on machine learning, in order to provide objective scores as well as PDU. However, such system development will first require training data to set the model parameters. Therefore, we developed a simulated video database in which video quality was rated by human observers. In order to train the proposed method for a wide range of situations, video clips with different content, encoding settings and simulation of transmission errors were included in the said database. We also used videos captured from ISDB-T broadcast transmissions to validate and benchmark the proposed model. Hence, the model can be built from simulated data and applied in practice by extracting features from the code stream and obtain predicted MOS (i.e. objective quality score) as well as predicted PDU (i.e. % of dissatisfied users as predicted by the objective model).

We now describe the video database, features employed and the machine learning techniques employed for feature pooling.

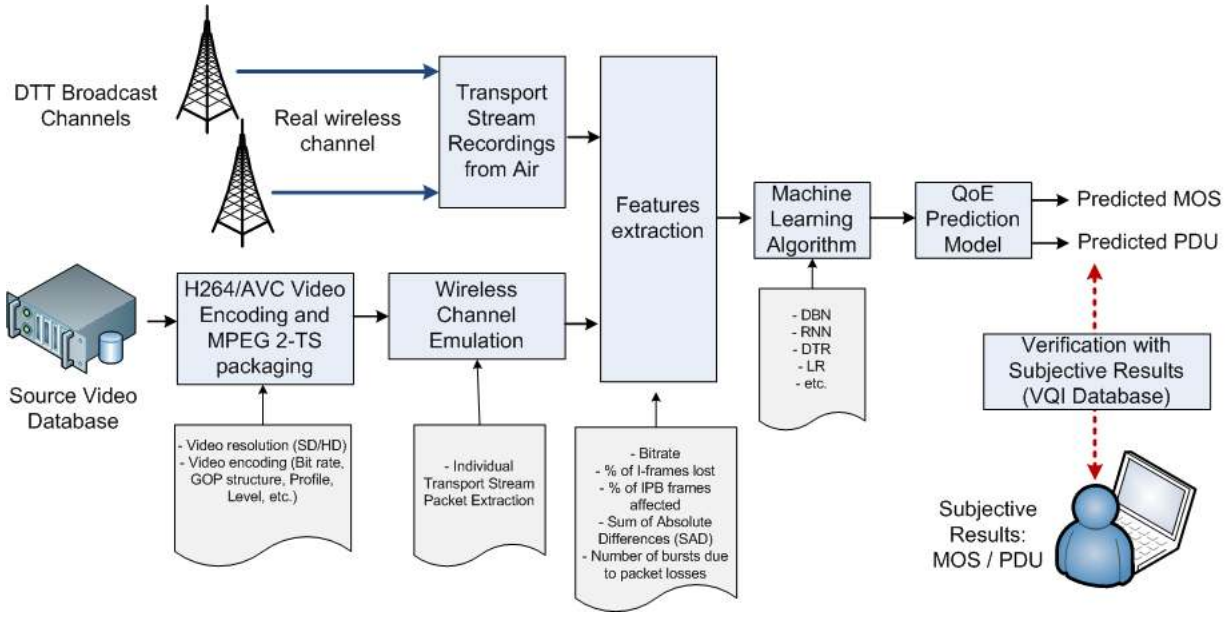

Fig 1 An overview of the proposed idea in practical video transmission network.

4.1 Datasets

We used a recently published database with video clips and raw subjective scores of subjective video quality within the context of DTT. The database is extracted from[15] and is suitable to train, verify and validate video quality objective models in multimedia broadband and broadcasting operations, under the ISDB-T standard. Specifically under the Brazilian version of the standard, known as ISDB-Tb that uses H.264/AVC for video compression, Advanced Audio Coding (AAC) for audio compression and MPEG-2 Transport Stream (TS) for packaging and multiplexing video, audio and data signals in the digital broadcasting system. The subjective tests in this database were conducted following the recommendation ITU-R BT.500-13 [3] using the Absolute Category Rating with Hidden Reference (ACR-HR) method. Subjective score collection was automated by employing a software based system.[18] The database includes two datasets with video clips that are 9 to 12 seconds in duration.

The first dataset consists of videos distorted by simulation of the video delivery chain. For this dataset, five High Definition (HD, resolution being 1920×1080) source (reference) sequences were used, namely "Concert", "Football", "Golf", "Foxbird" and "Voile". Each source video has undergone an encoding process with different encoding settings according to the ISDB-Tb standard using H.264/AVC and MPEG 2-TS for packaging. Then, a process of individual TS packet extraction was performed in order to simulate transmission errors. A total of 20 encoding and packet loss pattern conditions were generated for each source sequence providing $5 \times 20 = 100$ HD distorted video sequences. Since resolution is an important aspect in video quality, the same process was applied to down-sampled source video sequences, thus providing another 100 SD resolution (720×576) distorted sequences. Thus, the first dataset has 200 (100 HD and 100 SD) distorted video sequences. The encoding settings that have been imposed on the videos are: for SD (HD) videos: Profile = Main (High), Level = 3.1 (4.1), GOP length = 33, frame rate = 50fps and bit rate from 0.7 to 4 Mbps (3.5 to 14 Mbps). As for the different packet loss patterns, it was used 0% (no losses), 0.3% of losses with uniform distribution and 0.1% or 10% of packet losses within zero, one, two or three burst errors. For more details on the creation of this dataset, the interested reader can refer to.[15]

The second dataset, generated for validation purposes, includes only real recorded video clips from air from two different DTT Broadcast Channels. In this dataset, different encoding impairments and real packet losses patterns can be found in both HD and SD resolution (thus, there are 200 sequences, 100 HD and 100 SD). Each of the 200 video versions were evaluated by a human panel consisting of at least 18 viewers (27 for any HD video and 18 for any SD video) in a controlled environment. The MOS scale was used for these evaluations. All results were recorded in the database of[15] that is used here as well.

In this paper, both datasets were used, i.e. a total of 400 video sequences distorted by encoding impairments and transmission errors. Also note that the content types (i.e. source sequences) in both datasets were different.

these were taken from http://www.cdvl.org and IRCCyN IVC 1080i Video Quality Database[19]

4.2 Feature set

In DTT the video signal is typically coded in H264/AVC or MPEG-2 and packetized in small packets of 188 bytes (TS packets) prior to being modulated and transmitted. In MPEG-2 compression the compressed video frames are grouped into Group of Pictures (GoP). Each GoP usually uses three types of frames, named: I-intra , P-predictive, and B-bidirectional. I frames are encoded with Intra-frame compression techniques while P and B frames use motion estimation and compensation techniques. I frames are used as reference frames for the prediction of P and B frames. The GoP size is given by the number of frames existing between two I frames. In the case of H264/AVC each frame can be split into multiple slices: I, P or B. Both compression techniques can be packaged in Transport Stream (TS) packets. Each TS packet contains 4 bytes of header and 184 of payload. The header contains, among other fields, a 4-bit long Continuity Counter that can be used to count the amount of packet losses in the received bit stream.

Our approach to select the features was based on previous no-reference methods such as the one described in.[20] For our method, the selected features are the following:

- **Bit rate:** The obtained video bit rate due to the encoding process (H.264/AVC) and the MPEG-2 TS packaging.
- **Percentage of I-frames lost:** The I-frames carry the most reliable and important information, compared to P and B frames. Also I frames help decode non I frames, therefore their partial or total loss due to transmission errors is a key quality degrading factor.
- **Percentage of I,P,B frames lost:** In addition to the most crucial I frames, we also use this metric to account for P and B frames directly hit by transmission errors (without any further distinctions though).
- **SAD (Sum of Absolute Differences):** The SAD of Residual Blocks is a spatio-temporal metric that for instance addresses the degree of complexity of a sequence of images to be compressed.
- **Number of bursts:** Transmission errors normally affect groups of frames. The amount of bursts was selected in order to quantify the number of sequential frames directly hit by transmission errors in a video transmission (e.g. first a IIBPP frames are directly hit by transmission errors and then a PBPIPIBBB), we employ the number of bursts as a factor for objective quality prediction.

These features are used as input to the ML algorithm, as depicted in Figure 1. Otherwise put, they constitute the key QoE influence factors that we have identified, which will be used to build the ML-based QoE prediction model. Once a QoE model is built and put into practice, these features will be extracted from data streams and used as input for the QoE prediction. Of course, additional or different features can be used and hence the described method is scalable in terms of feature selection.

4.3 Feature pooling

We employed a number of feature pooling methods. These include both linear and non-linear models namely Linear Regression (LR), Decision Tree based Regression (DTR), Artificial Neural Networks (ANNs), and Deep Belief Networks (DBN).

4.3.1 Random Neural Networks (RNN)

The first model under scrutiny is Random Neural Network (RNN), which combines classical ANNs with queuing networks. Similar to ANN, RNN is composed of different layers of interconnected processing elements (i.e. neurons/nodes) that cooperate to resolve a specific problem by instantaneously exchanging signals between each other and from/to the environment. RNN is well adapted for QoS/QoE learning[13] since it takes short training time as compared to ANN, is less sensitive to selection of hidden nodes as compared to ANN and can capture QoS/QoE mapping functions in a more robust and accurate way. The success of the use of RNN for learning is suggested in a number of works.[13,21–26]

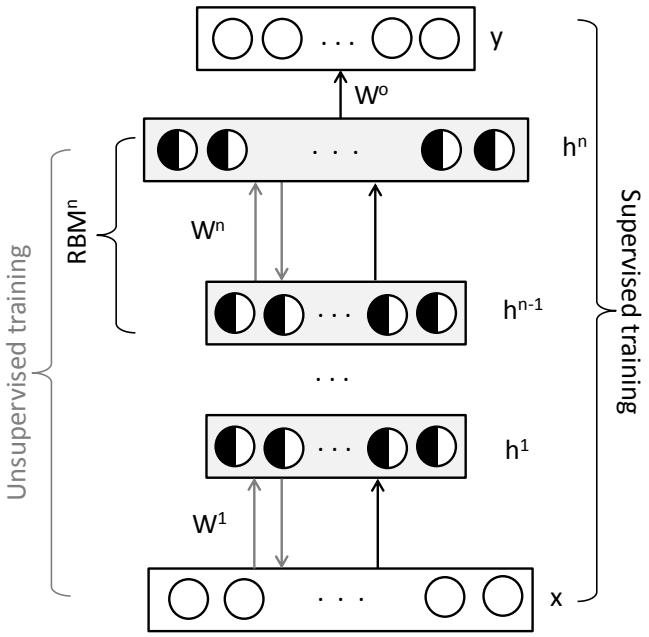

Fig 2 General architecture of DBN.

4.3.2 Deep Belief Networks (DBN)

The second model studied in this paper is inspired from Deep Learning (DL),[14] which makes small steps towards the mimicking of the human brain.[27] Technically, DL can be seen as the natural evolution of ANN.[28] Besides that, DL methods achieve very good results outperforming state-of-the-art algorithms, including classical ANN models (e.g. Multi Layer Perceptron), in different real-world problems such as multi-class classification,[29] collaborative filtering,[30] transfer learning,[31] people detection,[32] information retrieval,[33] activity recognition[34] and so on. Hence, our goal was to investigate to what extent DL can be applied to the problem of NR video quality prediction. While some prior work of applying DL for image quality evaluation exists,[35–38] a study of its effectiveness for NR video quality estimation especially in a multi-output scenario, as considered in this paper, has not been reported in literature.

Specifically, in this paper, we employed Deep Belief Networks (DBN) which are stochastic neural networks with more hidden layers and high generalization capabilities. They are composed by many, much simpler, two-layers stochastic neural networks, namely Restricted Boltzmann Machines (RBMs)[39] which are stacked one above the other in a deep architecture as depicted in Figure 2. More precisely, a DBN consists of an input layer with real values (i.e. $\mathbf{x}$), a number of n hidden binary layers (i.e $\mathbf{h^1}$,...,$\mathbf{h^n}$), and an output layer (i.e. $\mathbf{y}$) with real-values. The neurons from different layers are connected by weights (i.e. $\mathbf{W^1}$, ..., $\mathbf{W^n}$, $\mathbf{W^o}$). Formally, a DBN models the joint distribution between the input layer $\mathbf{x}$ and the n hidden layers, as it is shown next:

$$P(\mathbf{x}, \mathbf{h^1}, ..., \mathbf{h^n}) = \Big(P(\mathbf{x}|\mathbf{h^1}) \prod_{k=1}^{n-2} P(\mathbf{h^k}|\mathbf{h^{k+1}})\Big) P(\mathbf{h^{n-1}}, \mathbf{h^n}) \tag{2}$$

, where $P(\mathbf{h^k}|\mathbf{h^{k+1}})$ is a conditional distribution of the input units conditioned on the hidden units of the RBM^{k+1}, $\forall 1 \leq k < n-1$, given by:

$$P(\mathbf{h^k}|\mathbf{h^{k+1}}) = \prod_j P(h_j^k|\mathbf{h^{k+1}}) \tag{3}$$

$$P(h_j^k = 1|\mathbf{h^{k+1}}) = \frac{1}{1 + e^{-\sum_l W_{jl}^{k+1} h_l^{k+1}}} \tag{4}$$

, and $P(\mathbf{h^{n-1}}, \mathbf{h^n})$ is the joint distribution of the two layers composing RBM^n, computed as:

$$P(\mathbf{h^{n-1}}, \mathbf{h^n}) = \frac{1}{\mathcal{Z}(\mathbf{W^n})} e^{\sum_{j,l} W_{jl}^n h_j^{n-1} h_l^n} \tag{5}$$

, with $\mathcal{Z}(\mathbf{W^n})$ being the partition function of RBM^n. For RBM^1, $P(\mathbf{x}|\mathbf{h^1})$ can be computed in a similar manner with $P(\mathbf{h^k}|\mathbf{h^{k+1}})$.

The learning of DBNs parameters (e.g. $\mathbf{W^k}$) is made in two phases, as described in.[40] The first one is the *unsupervised training* phase. Herein, the weights $\mathbf{W^1}$, ..., $\mathbf{W^n}$ are considered to be bidirectional and the model is trained in an unsupervised way to learn to reconstruct probabilistically the inputs as well as possible, by using just the input data. As it is shown in Figure 2, in this phase just the neurons from the input and the hidden layers are involved. After this training phase, the hidden layers may perform automatically features extraction on inputs (i.e. the neurons which compose the hidden layers turn on or off when some specific values in a subset of the input neurons set occur). The second phase is the *supervised training* and the neurons from all the layers are involved in it. Herein, the model learns to perform classification or regression. More exactly, the previous learned DBN model is transformed in a directed neural network from bottom to top. The weights $\mathbf{W^1}$, ..., $\mathbf{W^n}$ are initialized with the previous learned values, while $\mathbf{W^o}$ are randomly initialized. After that, the DBN model is trained to fit pairs of input and output data points, as best as possible, by using a standard neural network training algorithm, such as back-propagation.[41] However, the above represents just a high level description of the DBNs formalism with the scope of providing to the non-specialist reader an intuition about the mechanisms behind DBNs. The overview of the deep learning complete mathematical details do not constitute one of the goals of this paper and the interested reader is referred to[14] for a thorough discussion.

5 Experimental Results and Analysis

This section presents experimental evaluation, and related analysis of the results obtained.

5.1 Test method setup

To assess the performance of our proposed method, we have considered two scenarios. First, we performed content-independent within dataset cross validation using the first video dataset (recall there are two datasets used in this study as discussed in the previous section). Since there are 5 different types of content, we performed a 5 fold cross-validation, where each fold represents one video type. In total, we repeated the experiments five times, each time choosing a different video to test the models, and the other four to train them. In the second scenario, we employed cross dataset validation: one dataset was used as training set and the other one as testing set. Hence we ensured that in both scenarios, train and test sets were content independent. In both scenarios, for all the machine learning algorithms analyzed, the inputs consist of features described in Section 4.2.

A distinct advantage that DBN offers over other competing methods is that they can be effectively initialized with unlabeled data in the unsupervised learning phase, and the second phase involves labeled data. As a result, they would require much less labeled training data to achieve similar or better prediction performance. Clearly, this is desirable in the context of video quality estimation where the availability of labeled data (i.e. subjective video quality ratings) is limited for obvious reasons. Thus, we have used two DBN models which employed less labeled training data (i.e. pairs inputs-outputs) in the *supervised learning* phase, while in the *unsupervised learning* phase they were trained with all the data but without the need of the corresponding label. Besides that, we have analyzed DBN and RNN models with one output (i.e. the model is specialized to predict just MOS or just PDU) or with two outputs (i.e. the model is capable to predict both, MOS and PDU). More specifically, in all sets of experiments performed, we have used the following DBN and RNN models: DBN^1_{100} (it used 100% of the labeled training data and it had 1 output), DBN^2_{100} (it used 100% of the labeled training data and it had 2 outputs), DBN^1_{40} (it used 40% of the labeled training data chosen randomly and it had 1 output), DBN^2_{40} (it used 40% of the labeled training data chosen randomly and it had 2 outputs), DBN^1_{10} (it used 10% of the labeled training data chosen randomly and it had 1 output), DBN^2_{10} (it used 10% of the labeled training data chosen randomly and it had 2 outputs), RNN^1 (it had 1 output), and RNN^2 (it had 2 outputs).

For the DBN models, we used 3 hidden layers with 10 hidden neurons on each of them. The learning rate (i.e. the factor which applies a greater or lesser portion of the weights adjustments computed in a specific epoch to the older weights computed in the previous epochs) was set to 10^{-3}, momentum (i.e. the factor which allows to the weights adjustments made in a specific epoch to persist for a number of epochs with the final goal to increase the learning speed) to 0.5, the weight decay (i.e. the factor which reduces overfitting to the training data, and shrinks the useless weights) to 0.0002, and the weights were initialized with $\mathcal{N}(0, 0.01)$ (i.e. Gaussian distribution). The number of training epochs in the *unsupervised training* phase was set to 200, while the number of training epochs in the *supervised training* phase using back-propagation was set to 1600. To ensure a smooth training, the data have been normalized to have zero mean

Table 3 Performance evaluation with 5-fold cross-validation.

Videos	Concert				Foot				Golf				Ntia				Voile				Average			
Metrics	MOS		PDU		MOS		PDU		MOS		PDU		MOS		PDU		MOS		PDU		MOS		PDU	
	RMSE	PCC	RMSE	PCC	RMSE	PCC	RMSE	PCC	RMSE	PCC	RMSE	PCC	RMSE	PCC	RMSE	PCC	RMSE	PCC	RMSE	PCC	RMSE	PCC	RMSE	PCC
PSNR	0.30	0.93	n/a	n/a	0.42	0.91	n/a	n/a	0.53	0.87	n/a	n/a	0.47	0.90	n/a	n/a	0.47	0.86	n/a	n/a	0.44±0.08	0.89±0.03	n/a	n/a
LR	0.60	0.76	0.26	0.70	0.60	0.77	0.25	0.71	0.55	0.82	0.22	0.80	0.70	0.79	0.24	0.78	0.70	0.71	0.26	0.71	0.63±0.08	0.77±0.04	0.25±0.01	0.71±0.04
DTR	0.61	0.79	0.34	0.57	0.54	0.81	0.27	0.69	0.77	0.71	0.31	0.67	0.68	0.80	0.25	0.78	0.60	0.79	0.34	0.67	0.64±0.08	0.78±0.03	0.30±0.04	0.67±0.07
RNN^1	0.4	0.85	0.23	0.79	0.52	0.83	0.23	0.76	0.45	0.86	0.20	0.84	0.57	0.88	0.17	0.86	0.59	0.81	0.21	0.80	0.51±0.08	0.85±0.02	0.21±0.03	0.81±0.04
RNN^2	0.54	0.84	0.23	0.78	0.54	0.82	0.23	0.75	0.49	0.87	0.19	0.84	0.61	0.88	0.17	0.86	0.61	0.79	0.21	0.80	0.56±0.06	0.84±0.04	0.21±0.03	0.80±0.05
DBN^1_{100}	0.49	0.83	0.21	0.79	0.52	0.82	0.21	0.78	0.54	0.83	0.21	0.82	0.60	0.84	0.19	0.84	0.66	0.78	0.22	0.77	0.56±0.06	0.82±0.02	0.21±0.01	0.80±0.02
DBN^2_{100}	0.47	0.85	0.20	0.82	0.54	0.82	0.22	0.77	0.50	0.85	0.20	0.83	0.64	0.81	0.21	0.81	0.61	0.80	0.22	0.79	0.55±0.06	0.83±0.02	0.21±0.01	0.80±0.02
DBN^1_{40}	0.44	0.83	0.20	0.81	0.54	0.83	0.22	0.78	0.53	0.83	0.21	0.82	0.58	0.85	0.19	0.84	0.63	0.80	0.22	0.78	0.54±0.06	0.83±0.02	0.21±0.01	0.81±0.02
DBN^2_{40}	0.52	0.82	0.23	0.78	0.52	0.83	0.22	0.78	0.55	0.84	0.21	0.81	0.59	0.84	0.20	0.82	0.57	0.82	0.19	0.81	0.55±0.02	0.83±0.01	0.21±0.01	0.80±0.02
DBN^1_{10}	0.49	0.82	0.23	0.80	0.56	0.83	0.22	0.78	0.59	0.82	0.20	0.83	0.61	0.84	0.22	0.84	0.80	0.79	0.19	0.78	0.61±0.10	0.82±0.02	0.21±0.01	0.81±0.02
DBN^2_{10}	0.46	0.86	0.22	0.82	0.66	0.82	0.24	0.75	0.57	0.81	0.26	0.72	0.68	0.80	0.23	0.80	0.70	0.80	0.26	0.77	0.61±0.08	0.82±0.03	0.24±0.01	0.78±0.03
FixSig	n/a	n/a	0.20	0.83	n/a	n.a	0.28	0.72	n/a	n/a	0.28	0.78	n/a	n/a	0.22	0.84	n/a	n/a	0.39	0.73	n/a	n/a	0.27±0.06	0.78±0.05
FitSig	n/a	n/a	0.19	0.84	n/a	n.a	0.27	0.73	n/a	n/a	0.28	0.77	n/a	n/a	0.22	0.84	n/a	n/a	0.38	0.75	n/a	n/a	0.26±0.06	0.79±0.04

and one unit variance as discussed in.[42] For the RNN models we used the implementation offered by Changlin Liu and Luca Muscariello. For the LR and DTR implementations we have used the scikit-learn library.[43]

Besides that, to assess the quality of the PDU predictions using the various machine learning techniques under scrutiny (which are applied directly on the features extracted from the videos), we tried to estimate also the PDU values by using two simpler statistical approaches in which we have exploited the sigmoid-like relation between MOS and PDU. Formally, for each video i from the testing set, we have estimated its PDU value, $\widehat{PDU_i}$, from a Gaussian probability density function, as follows:

$$\widehat{PDU_i} = P(1 \leq x \leq th) = \int_1^{th} \frac{1}{\sigma_i\sqrt{2\pi}} e^{\frac{-(x-\mu_i)^2}{2\sigma_i^2}} \tag{6}$$

, where th represents the selected threshold for PDU, μ_i represents the MOS for the video i, and σ_i means the standard deviation of all individual subjective scores associated with video i. However, due to the fact that in a real video service it is impossible to obtain μ_i and σ_i in real-time, in our experiments we set μ_i to the MOS value predicted for the video i by the best performer among the machine learning techniques used. At the same time, we have estimated σ_i considering two cases: (1) a fixed value given by the mean value of all standard deviations, computed each of them on the individual subjective scores associated with each video from the training set (method dubbed further FixSig); (2) a variable value given by a Gaussian curve fitted on the MOS values of the videos from the training set and their corresponding standard deviation (method dubbed further FitSig) and the previous discussed μ_i.

The performance was assessed using Pearson (PCC) and Spearman (SRCC) correlation coefficients, and the root mean squared error (RMSE) values. Note that we employed the mentioned performance measures for both MOS and PDU prediction accuracies. To serve as a benchmark, we also computed the results using peak-signal-to-noise (PSNR), which is still a popular FR method. The results (correlations, RMSE) for PSNR were computed after the non-linear transformation recommended in.[44] The reader will however recall that in the considered application, decoded video data is assumed to be unavailable, and hence objective methods that require pixel data cannot be employed in practice.

5.2 Test results

The results for the first scenario i.e. 5-fold cross validation are presented in Table 3, in which we have reported the RMSE and correlation values for each fold as well as the average over the 5 folds. We can observe that while all the methods achieve statistically similar performances for MOS prediction accuracies, DBNs perform better in predicting PDU. To obtain further insights, we have plotted in Figure 3 the outcomes of DBN^2_{10} on two content types namely "Concert" and "Voile". In these plots, the blue dots show the locations of subjective MOS vs the predicted MOS (obviously they will lie on the 45° in case of perfect prediction) while the error bars represent PDU. We have shown the results only for DBN^2_{10} due to the fact that it is probably the most interesting model because it uses only 10% labeled training data and hence is practically more robust against the amount of labeled training data available. Moreover, recall that DBN^2_{10} outputs both MOS and PDU simultaneously from single training unlike other models which need to be trained twice on subjective MOS and actual PDU. Hence, it is able to predict both values at the same time. It can be observed in both plots that the blue dots lie close to the main diagonals (which represent the perfect predictions for the MOS values). Moreover, predicted PDU is close to the actual PDU, although the accuracy is less in case of "Voile" sequence at higher subjective MOS.

https://code.google.com/p/qoe-rnn/, Accessed on March 7th, 2015.

Please recall that in this paper th is set to 3.

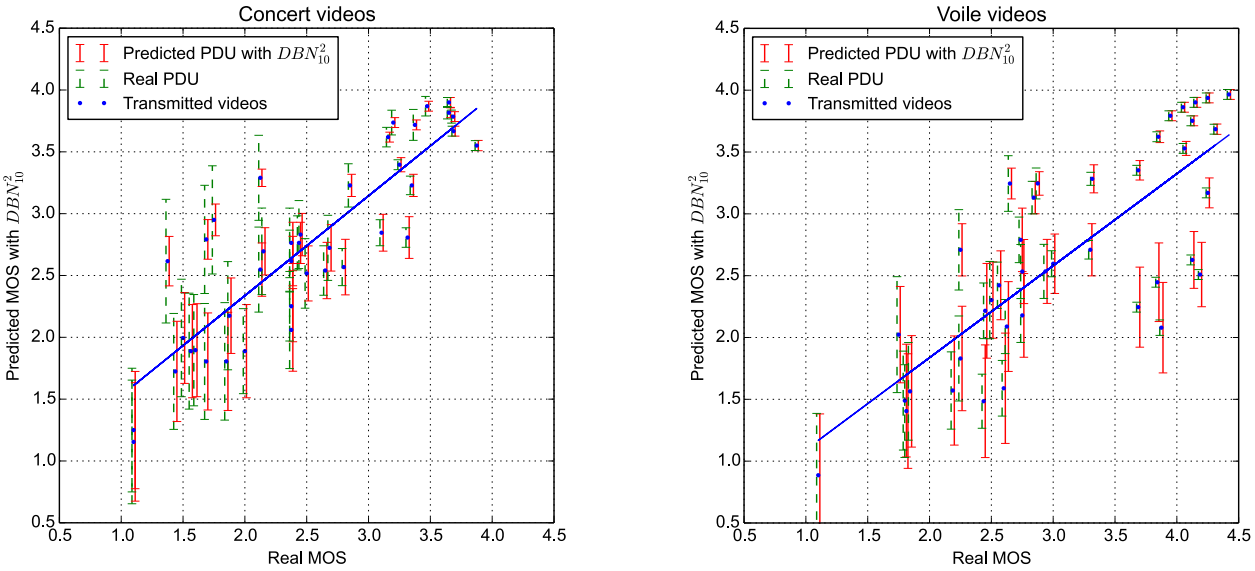

Fig 3 Cross-validation results snapshot. The real MOS and PDU values plotted against the predicted MOS and PDU values using DBN_{10}^2 on the best performers (i.e "Concert" videos) and on the worst performers (i.e. "Voile" videos). Each point represents an impaired video.

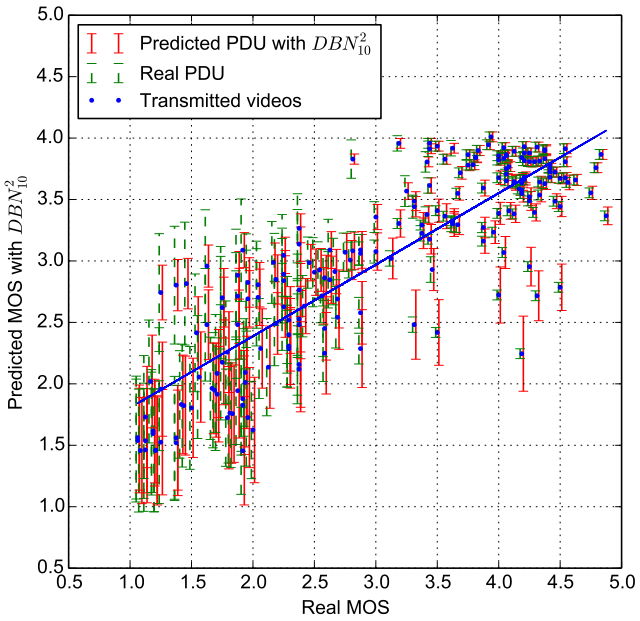

Fig 4 Results for the second dataset (note that the system was trained using only the first dataset). The real MOS and PDU values plotted against the predicted MOS and PDU values using DBN_{10}^2. Each point represents an impaired video.

The results for the second test scenario (cross-dataset validation) are presented in Tables 4 and 5. One can again see that DBNs tend to perform better considering both MOS and PDU predictions. Note that PSNR results cannot be computed in case of Table 4 because the videos were registered from air and hence the source (reference) video is unavailable. Hence, these results are relevant for a practical end-to-end video delivery chain where FR methods cannot be employed. Finally, the MOS-PDU plot for the scenario considered in Table 4 is shown in Figure 4 (for DBN_{10}^2). This allows the reader to judge the scatter around the diagonal as well as compare the actual and predicted PDU values.

In both test scenarios, we may observe that DBNs perform better for PDU predictions that any other methods in terms of the all evaluation metrics. Besides that, it is interesting to note that even the two simpler statistical methods performs quite well, being able to predict PDUs with a good correlation factors, but having some flaws in the case of the RMSE metric. Moreover, we would like to highlight that in our experiments FitSig proven to be more robust

Table 4 Cross dataset validation. The system was trained with sequences from the first dataset (200 sequences, 100 HD and 100 SD), and the test set consisted of 200 videos taken from air (the second dataset).

Metrics	MOS			PDU		
	RMSE	PCC	SRCC	RMSE	PCC	SRCC
LR	0.70	0.82	0.81	0.24	0.81	0.82
DTR	0.62	0.83	0.80	0.24	0.80	0.81
RNN^2	0.77	0.84	0.85	0.18	0.90	0.84
DBN^2_{100}	0.58	0.87	0.85	0.20	0.87	0.83
DBN^2_{40}	0.61	0.88	0.83	0.19	0.90	0.84
DBN^2_{10}	0.60	0.86	0.82	0.19	0.88	0.83
FixSig	n/a	n/a	n/a	0.27	0.83	0.81
FitSig	n/a	n/a	n/a	0.28	0.84	0.81

Table 5 Cross dataset validation. The system was trained with 200 videos taken from air (second dataset), and the test set consisted of 200 sequences (100 HD and 100 SD) from the first dataset.

Metrics	MOS			PDU		
	RMSE	PCC	SRCC	RMSE	PCC	SRCC
LR	0.75	0.75	0.77	0.27	0.72	0.77
DTR	0.77	0.69	0.65	0.31	0.66	0.68
RNN^2	1.25	0.78	0.76	0.24	0.77	0.77
DBN^2_{100}	0.60	0.81	0.81	0.23	0.76	0.80
DBN^2_{40}	0.63	0.79	0.78	0.25	0.74	0.77
DBN^2_{10}	0.65	0.80	0.78	0.24	0.76	0.79
FixSig	n/a	n/a	n/a	0.29	0.62	0.71
FitSig	n/a	n/a	n/a	0.29	0.75	0.77

than its counterpart FixSig, especially when the subjective studies came from different datasets, due to its better representational power given by a better fitted standard deviation σ_i. For a better insight into the differences between DBNs and the statistical approaches in Figure 5 we plot the results of DBN^2_{10} and FitSig in the case of the cross dataset validation scenario. Herein, it is interesting to see that at small MOS values FitSig performs better than DBN^2_{10}, while at MOS values usually higher than 2.5, DBNs perform much better. Similarly, we have observed the same behavior also for the other DBN models on one side and FixSig and FitSig on the other side in both test scenarios, the 5-fold cross validation and the cross dataset validation. These, corroborated with the fact that FixSig and FitSig still need an external prediction method to estimate μ_i, make DBNs the most suitable method to predict PDU.

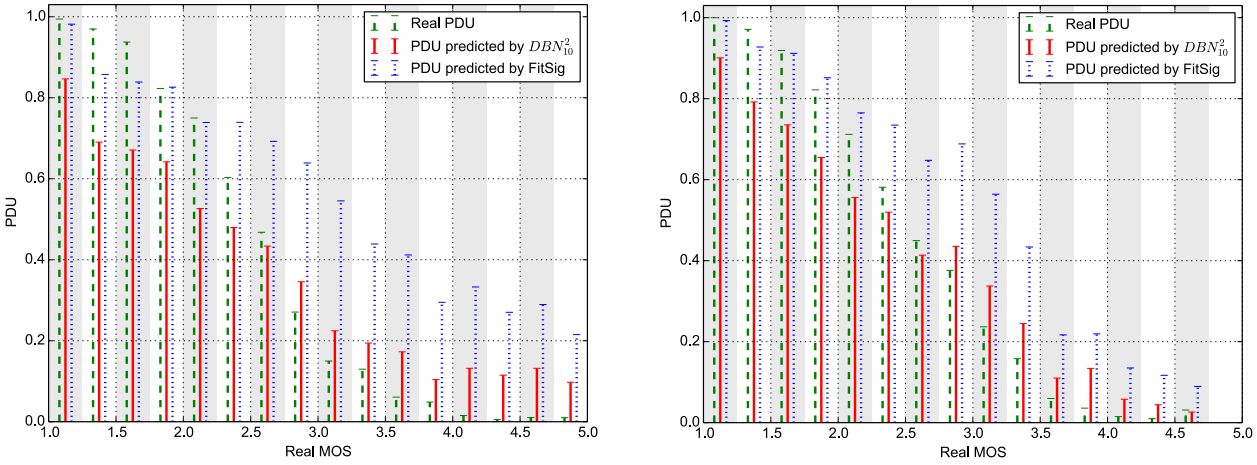

Fig 5 Comparison of the real PDU with the predictions made by DBN^2_{10} and FitSig. Each PDU bar represents the mean values of the PDUs situated in the light gray or in the white areas, respectively. In the left plot, the system was trained with the 200 sequences (100 HD and 100 SD) from the first dataset, and the test set consisted of the 200 videos taken from air (second dataset), while in the right plot the training and the testing sets were reversed.

Table 6 Analytic study of the relations between the DBNs weights in different learning phases. The assessment metrics are computed between the weights of the DBN under scrutiny after the *supervised learning* phase and their corresponding values obtained after *unsupervised learning* phase and before the *supervised* one.

Training Set	Model	$\mathbf{W^1}$			$\mathbf{W^2}$			$\mathbf{W^3}$		
		RMSE	PCC	SRCC	RMSE	PCC	SRCC	RMSE	PCC	SRCC
First Dataset	DBN^2_{100}	0.17	0.98	0.96	0.49	0.93	0.93	0.30	0.97	0.98
	DBN^2_{40}	0.20	0.97	0.95	0.54	0.92	0.91	0.36	0.96	0.97
	DBN^2_{10}	0.22	0.96	0.94	0.56	0.91	0.90	0.38	0.96	0.96
Second Dataset	DBN^2_{100}	0.11	0.99	0.99	0.23	0.99	0.98	0.26	0.98	0.98
	DBN^2_{40}	0.15	0.99	0.98	0.26	0.98	0.98	0.29	0.98	0.97
	DBN^2_{10}	0.14	0.99	0.98	0.26	0.98	0.97	0.27	0.98	0.98

5.3 Learning of weights in DBN

To understand better how deep learning works, in Figure 6, the behavior of DBN^2_{10} during the training on the first video dataset is plotted. It can be observed that in the *unsupervised learning* phase the model learns to reconstruct the inputs well after approximately 50 training epochs, and after roughly 100 training epochs it reconstructs them very precisely, independently of the RBM under scrutiny (RBM^1,RBM^2,RBM^3). More than that, the same plot suggests a clear correlation between the three performance metrics used over the training epochs to assess the learning process, such that when the averaged RMSE and P-value tends to get closer to zero, the averaged PCC value tends to get closer to one, showing overall a perfect correlation between them. Further on, in the *supervised learning* phase, DBN^2_{10} learns with back-propagation to predict the training outputs with a very small error after about 800 training epochs. We would like to highlight, that all the DBN models discussed in this paper, independently on the scenario, had a similar behavior as the one described previously for DBN^2_{10}.

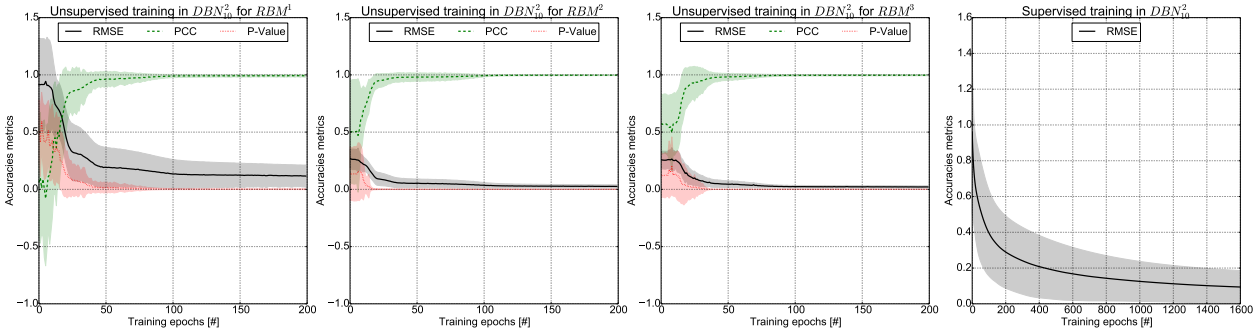

Fig 6 The behavior of DBN^2_{10} during the training on the first video dataset. The first three plots depict the *unsupervised training* phase for each RBM belonging to the DBN^2_{10}, while the last one presents the *supervised training* phase in which the DBN^2_{10} was trained using back-propagation. The straight lines represent the mean and the shaded areas reflect the standard deviation computed for all the data points.

Furthermore, we have analyzed the most important free parameters (i.e. the weights $\mathbf{W^1}$, $\mathbf{W^2}$, $\mathbf{W^3}$, and $\mathbf{W^o}$) of the DBN models used in this experiment. The relations between these parameters are exemplified visually in Figure 7, and presented in Table 6. In both, it can be observed, that practically the weights learned during the *unsupervised training* phase do not change too much after the *supervised training* phase, independently if we study DBN^2_{100}, DBN^2_{40}, or DBN^2_{10}. This probably explains why in the literature, the latter one is called "fine tuning". At the same time, the fact that the weights of the three fine tuned DBNs end up in a region very close to the one discovered by the initial unsupervised learning procedure reflects also why a DBN which uses just 10% of the labeled data for the back propagation training has a similar performance with one which uses 100% of the labeled data. Besides that, the sparsity patterns of the weights reflect which input neurons contribute more to any hidden neurons. As an example, we can observe that the neuron number 8 from $\mathbf{h^1}$ is affected just by neurons 3 (i.e. % of total frames lost) and 5 (i.e. # of bursts) from $\mathbf{x}$, or in other words the DBNs models automatically find a correlation between % of total frames lost and # of bursts. Similarly, we can deduce that the 10^{th} hidden neuron from $\mathbf{h^1}$ represents a relation between all the 5 input features used. It is worth highlighting, that using similar cascade deductions, one might discover why the neuron number 9 from $\mathbf{h^3}$ has such a strong impact on both neurons (i.e. MOS and PDU) from the output layer $\mathbf{y}$.

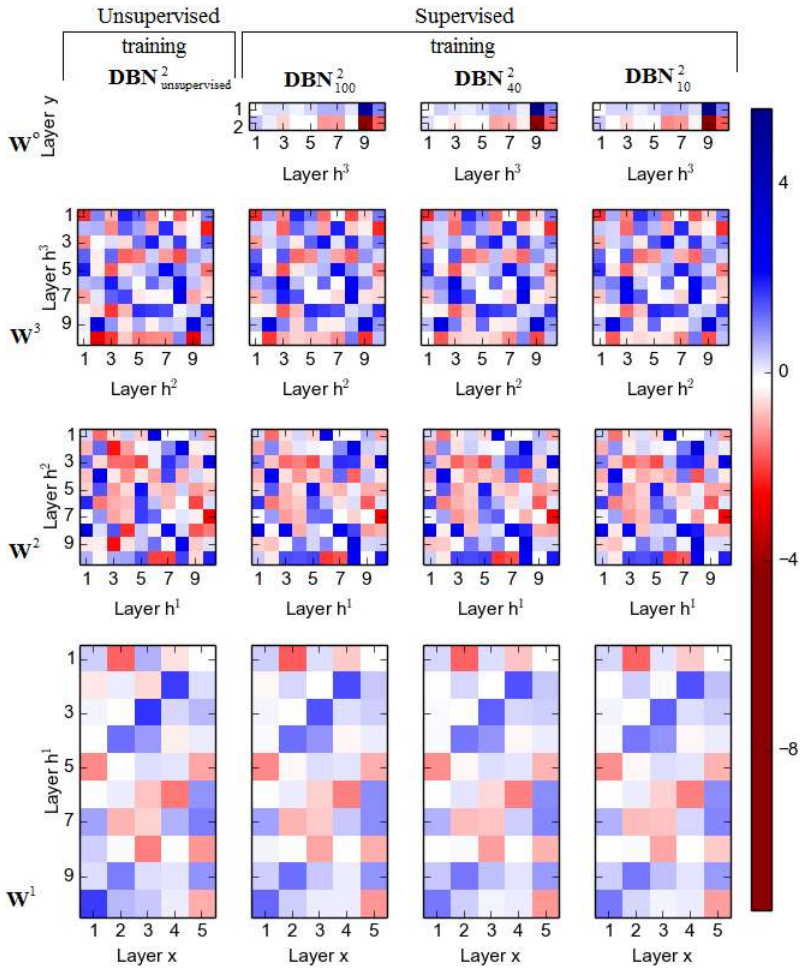

Fig 7 The values of the weights in the DBNs models when the training was done on the first data set. The values on the x-axis and y-axis represent the index of neurons from that specific layer. The neurons from the input layer $\mathbf{x}$ represent the following features: real bit rates (i.e. 1^{st} neuron), % of I-frames lost (i.e. 2^{nd} neuron), % of total frames lost (i.e. 3^{rd} neuron), SAD (i.e. 4^{th} neuron), # of bursts (i.e. 5^{th} neuron). The first column reflects the weights of the $\text{DBN}^2_{unsupervised}$ obtained after the *unsupervised training* phase, while the last three columns represent the weights of the DBNs obtained after the *supervised training* phase. Moreover, on rows, the bottom one represents the DBNs inputs, while the top one represents the DBNs outputs. The dark red in the heat maps represents weights values closer to -11, while the white depicts weights values around 0 and the dark blue shows weights values towards 6.

6 Discussion

One of the main aims of this paper has been to demonstrate how objective quality prediction can be augmented by considering variability of subjective data. Particularly, we have shown how machine learning can add value to objective video quality estimation by considering a two-output DBN model. Hence, we train the model not only to predict MOS but also to put subjective variability into observation. Consequently, we are able to deepen our understanding of the service in question from two perspectives: overall service quality and the satisfaction of the customer base. Utilizing percentage below threshold instead of standard deviation or other typical mathematical scattering indicators unveils the answer to the question "how many users are not happy with the service" instead of "are users on average happy with the service". These two perspectives have a profound difference when it comes to quality management, as quality does not translate directly into business success: slightly bad quality does not mean slightly decreased market share. In some cases it can be the differentiating factor between success and failure. Meeting the needs of all customers and detecting and dealing with customer dissatisfaction are key components in service quality management, especially when we consider things of high abstraction level such as quality of experience.

During the course of the study we learned that there is a rough sigmoid-like correlation between MOS and uncertainty of MOS for this dataset. This observation cannot be generalized for all datasets and selected features, but it

is nonetheless notable that when MOS drops around the selected threshold of satisfaction, the number of dissatisfied users increases the fastest. Different features may pose different kind of relations depending on how opinions of subjects vary due to particular feature. This phenomenon becomes more apparent if participants of subjective assessment are selected from different regions, age groups, cultures and backgrounds. This was noted for example in[45] where authors studied website aesthetics and discovered a major difference between Asian and non-Asian users in perception of website visual appeal. We propose that this may also apply to certain quality aspects where some user groups perceive some quality degradation as much worse than other users.

User dissatisfaction information can be utilized in many ways in practice. Traditional management mechanisms such as traffic shaping, admission control or handovers can be further enhanced to also include a "risk threshold" for user dissatisfaction in addition to MOS threshold. For instance, let us assume a QoE managed service where a provider is able to automatically monitor the service per-user level. The provider uses a machine-learning model which outputs two values, objective MOS and probability that the user is not satisfied. The management mechanism can step in to improve the user experience if either the estimated MOS drops below a certain threshold, or if the estimated dissatisfaction level rises above a certain value (for example, MOS is required to remain above 3 and risk that the user opinion is below 3 must be less than 5%).

But what may be even more useful for the service provider is the overall MOS and dissatisfaction percentage throughout the service. This also helps providers to reflect how the service is doing competition-wise and if they can expect user churn in the near future. Holistic, real-time monitoring may also help to indicate serious faults and problems either with the service or the transfer network and help to act accordingly. Operators can therefore react to user dissatisfaction before customers either terminate their service subscription or burden customer service.

7 Concluding Thoughts

While the problem of objective video quality assessment has received considerable research attention, most existing works tend to focus only on averaged ratings. As a result, valuable information generated as a result of inter-observer differences (i.e. subjective variability) is simply lost in objective quality prediction. This paper attempted to introduce and analyze one such instance of how the scattering of subjective opinions can be exploited for business-oriented video broadcasting applications. This was accomplished by first analyzing and formulating interpretable measure of user dissatisfaction which may not always be reflected in averaged scores. To put the idea into practice, we then explored the deep learning framework and jointly modeled the averaged scores and user dissatisfaction levels so that the predicted objective video quality score is supplemented by user satisfaction information. At the same time we showed that by using deep belief networks the amount of subjective studies required to learn to make accurate predictions, which outperform clearly the other machine learning models considered for comparison in this paper (i.e. linear regression, regression trees, and random neural networks), may be reduced up to 90%. This will be useful in a typical video broadcasting system where customer (user) churn needs to be continuously monitored. We also demonstrated a practical implementation of our ideas in the context of video transmission. We designed the system so that video quality and user dissatisfaction can be predicted from data bit stream with out the need of the fully decoded signal. This greatly facilitates real time video quality monitoring since objective quality can be predicted from the code stream.

Acknowledgments

The work was supported in part by funding from Qualinet (COST IC 1003) which is gratefully acknowledged. Eirini Liotou's work was supported by the European Commission under the auspices of the FP7-PEOPLE MITN-CROSSFIRE project (grant 317126). Janne Seppänen's work was supported by Tekes, the Finnish Funding Agency for Technology and Innovation, under QuEEN (Quality of Experience Estimators in Networks) and NOTTS (Next generation Over-The-Top multimedia Services) projects. Juan Pablo Garella's work was partially funded by Comisión Académica de Posgrado, Universidad de la República, Uruguay.

References

1 "Qualinet white paper on definitions of quality of experience," European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003) (2013).

2 S. Chikkerur, V. Sundaram, M. Reisslein, and L. Karam, "Objective video quality assessment methods: A classification, review, and performance comparison," *Broadcasting, IEEE Transactions on* **57**, 165–182 (2011).

3 "Methodology for the subjective assessment of the quality of television pictures," Recommendation ITU-R BT.500-13 (2012).
4 P. Gastaldo, S. Rovetta, and R. Zunino, "Objective quality assessment of mpeg-2 video streams by using cbp neural networks," *Neural Networks, IEEE Transactions on* **13**, 939–947 (2002).
5 P. Le Callet, C. Viard-Gaudin, and D. Barba, "A convolutional neural network approach for objective video quality assessment," *Neural Networks, IEEE Transactions on* **17**, 1316–1327 (2006).
6 J. Xu, P. Ye, Y. Liu, and D. Doermann, "No-reference video quality assessment via feature learning," in *Image Processing (ICIP), 2014 IEEE International Conference on*, 491–495 (2014).
7 K. Zhu, C. Li, V. Asari, and D. Saupe, "No-reference video quality assessment based on artifact measurement and statistical analysis," *Circuits and Systems for Video Technology, IEEE Transactions on* **PP**(99), 1–1 (2014).
8 N. Staelens, D. Deschrijver, E. Vladislavleva, B. Vermeulen, T. Dhaene, and P. Demeester, "Constructing a no-reference h.264/avc bitstream-based video quality metric using genetic programming-based symbolic regression," *Circuits and Systems for Video Technology, IEEE Transactions on* **23**, 1322–1333 (2013).
9 J. Sogaard, S. Forchhammer, and J. Korhonen, "No-reference video quality assessment using codec analysis," *Circuits and Systems for Video Technology, IEEE Transactions on* **PP**(99), 1–1 (2015).
10 B. Konuk, E. Zerman, G. Nur, and G. Akar, "A spatiotemporal no-reference video quality assessment model," in *Image Processing (ICIP), 2013 20th IEEE International Conference on*, 54–58 (2013).
11 M. Shahid, A. Rossholm, B. Lovstrom, and H.-J. Zepernick, "No-reference image and video quality assessment: a classification and review of recent approaches," *EURASIP Journal on Image and Video Processing* **2014**(1), 40 (2014).
12 T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer Series in Statistics, Springer New York Inc., New York, NY, USA (2001).
13 S. Mohamed and G. Rubino, "A study of real-time packet video quality using random neural networks," *Circuits and Systems for Video Technology, IEEE Transactions on* **12**, 1071–1083 (2002).
14 Y. Bengio, "Learning deep architectures for ai," *Found. Trends Mach. Learn.* **2**, 1–127 (2009).
15 J. P. Garella, J. Joskowicz, R. Sotelo, M. Juayek, and D. Durán, "Subjective video quality test: Methodology, database and experience," in *IEEE International Symposium on Broadband Multimedia Systems and Broadcasting*, (2015).
16 E. Karapanos, J.-B. Martens, and M. Hassenzahl, "Accounting for diversity in subjective judgments," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, *CHI '09*, 639–648, ACM, (New York, NY, USA) (2009).
17 T. Hobfeld, R. Schatz, and S. Egger, "Sos: The mos is not enough!," in *Quality of Multimedia Experience (QoMEX), 2011 Third International Workshop on*, 131–136 (2011).
18 J. Joskowicz, R. Sotelo, M. Juayek, D. Durán, and J. P. Garella, "Automation of subjective video quality measurements," in *Proceedings of the Latin America Networking Conference on LANC 2014*, 7:1–7:5, ACM (2014).
19 S. Péchard, R. Pépion, and P. Le Callet, "Suitable methodology in subjective video quality assessment: a resolution dependent paradigm," in *International Workshop on Image Media Quality and its Applications, IMQA2008*, 6 (2008).
20 J. Joskowicz and R. Sotelo, "A Model for Video Quality Assessment Considering Packet Loss for Broadcast Digital Television Coded in H.264," *International Journal of Digital Multimedia Broadcasting* **2014**(5786), 11 (2014).
21 C. Cramer, E. Gelenbe, and P. Gelenbe, "Image and video compression," *Potentials, IEEE* **17**, 29–33 (1998).
22 H. Bakrcolu and T. Koak, "Survey of random neural network applications," *European Journal of Operational Research* **126**(2), 319 – 330 (2000).
23 E. Gelenbe, "Stability of the random neural network model," *Neural Computation* **2**, 239–247 (1990).
24 K. Singh and G. Rubino, "Quality of experience estimation using frame loss pattern and video encoding characteristics in dvb-h networks," in *Packet Video Workshop (PV), 2010 18th International*, 150–157 (2010).
25 K. Singh, Y. Hadjadj-Aoul, and G. Rubino, "Quality of experience estimation for adaptive http/tcp video streaming using h.264/avc," in *Consumer Communications and Networking Conference (CCNC), 2012 IEEE*, 127–131 (2012).
26 E. Aguiar, A. Riker, A. Abelem, E. Cerqueira, and M. Mu, "Video quality estimator for wireless mesh networks," in *Quality of Service (IWQoS), 2012 IEEE 20th International Workshop on*, 1–9 (2012).
27 N. Jones, "Computer science: The learning machines," *Nature* **505**(7482), 146–148 (2014).

28 J. Laserson, "From neural networks to deep learning: Zeroing in on the human brain," *XRDS* **18**, 29–34 (2011).
29 H. Larochelle and Y. Bengio, "Classification using discriminative restricted boltzmann machines," in *Proceedings of the 25th International Conference on Machine Learning*, *ICML '08*, 536–543, ACM, (New York, NY, USA) (2008).
30 R. Salakhutdinov, A. Mnih, and G. Hinton, "Restricted boltzmann machines for collaborative filtering," in *Proceedings of the 24th International Conference on Machine Learning*, *ICML '07*, 791–798, ACM, (New York, NY, USA) (2007).
31 H. Ammar, D. Mocanu, M. Taylor, K. Driessens, K. Tuyls, and G. Weiss, "Automatically mapped transfer between reinforcement learning tasks via three-way restricted boltzmann machines," in *Machine Learning and Knowledge Discovery in Databases*, *Lecture Notes in Computer Science* **8189**, 449–464, Springer Berlin Heidelberg (2013).
32 E. Mocanu, D. Mocanu, H. Ammar, Z. Zivkovic, A. Liotta, and E. Smirnov, "Inexpensive user tracking using boltzmann machines," in *Systems, Man and Cybernetics (SMC), 2014 IEEE International Conference on*, 1–6 (2014).
33 P. V. Gehler, A. D. Holub, and M. Welling, "The rate adapting poisson model for information retrieval and object recognition," in *Proceedings of the 23rd International Conference on Machine Learning*, *ICML '06*, 337–344, ACM, (New York, NY, USA) (2006).
34 D. C. Mocanu, H. B. Ammar, D. Lowet, K. Driessens, A. Liotta, G. Weiss, and K. Tuyls, "Factored four way conditional restricted boltzmann machines for activity recognition," *Pattern Recognition Letters* **66**, 100 – 108 (2015). Pattern Recognition in Human Computer Interaction.
35 D. Mocanu, G. Exarchakos, and A. Liotta, "Deep learning for objective quality assessment of 3d images," in *Image Processing (ICIP), 2014 IEEE International Conference on*, 758–762 (2014).
36 D. Mocanu, G. Exarchakos, H. Ammar, and A. Liotta, "Reduced reference image quality assessment via boltzmann machines," in *Integrated Network Management (IM), 2015 IFIP/IEEE International Symposium on*, 1278–1281 (2015).
37 H. Tang, N. Joshi, and A. Kapoor, "Blind image quality assessment using semi-supervised rectifier networks," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, 2877–2884 (2014).
38 D. Ghadiyaram and A. Bovik, "Blind image quality assessment on real distorted images using deep belief nets," in *Signal and Information Processing (GlobalSIP), 2014 IEEE Global Conference on*, 946–950 (2014).
39 P. Smolensky, "Parallel distributed processing: Explorations in the microstructure of cognition, vol. 1," ch. Information Processing in Dynamical Systems: Foundations of Harmony Theory, 194–281, MIT Press, Cambridge, MA, USA (1986).
40 G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.* **18**, 1527–1554 (2006).
41 D. Rumelhart, G. Hintont, and R. Williams, "Learning representations by back-propagating errors," *Nature* **323**(6088), 533–536 (1986).
42 G. E. Hinton, "A practical guide to training restricted boltzmann machines," in *Neural Networks: Tricks of the Trade*, *Lecture Notes in Computer Science* **7700**, 599–619, Springer Berlin Heidelberg (2012).
43 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research* **12**, 2825–2830 (2011).
44 VQEG, "Final report from the video quality experts group on the validation of objective models of video quality assessment," (2003).
45 M. Varela, T. Maki, L. Skorin-Kapov, and T. Hossfeld, "Towards an understanding of visual appeal in website design," in *Quality of Multimedia Experience (QoMEX), 2013 Fifth International Workshop on*, 70–75 (2013).

Decebal Constantin Mocanu received the B.Eng. degree in Computer Science from Polytechnic University of Bucharest, Romania, in 2010, and the M.Sc. degree in Artificial Intelligence from Maastricht University, Netherlands, in 2013. At the same time, from 2001 until 2013, he has worked as a software engineer in various companies. Since 2013, he is a Ph.D. student at Eindhoven University of Technology, Netherlands. His research interests include, among others, artificial intelligence, machine learning, and computer vision.

Jeevan Pokhrel is currently working as a research engineer at Montimage, France. He completed his PhD from Institute Mines Telecom, France in 2014. He has been contributing in some of the European and French research projects. His research is focused on network performance evaluation and security issues. His topics of interest cover performance evaluation, multimedia Quality of Experience (QoE), wireless networks, machine learning etc.

Juan Pablo Garella received the degree of Electrical Engineer from Universidad de la Republica, Uruguay (UdelaR) in 2011. He is currently a candidate to obtain the Master Degree in Electrical Engineer at UdelaR. He participated in research projects on digital television with emphasis in perceived video quality estimation and QoS monitoring. Currently he is working as a consultant at the Digital TV Lab, LATU, Uruguay. His research interests include: Perceived Video Quality; QoE; Digital TV; ISDB-Tb.

Janne Seppänen (M.Sc) is working at VTT Technical Research Centre of Finland Ltd. as a research scientist, covering topics such as QoE-driven network management, QoE assessment, network traffic measurement, space communication, and traffic identification. He also has experience in neural networks research.

Eirini Liotou received the Diploma in Electrical & Computer Engineering from the National Technical University of Athens and the MSc in Communications & Signal Processing from the Imperial College of London. She has worked as a Senior Software Engineer in Siemens Enterprise Communications within the R&D department. Since 2013 she is a PhD candidate in the Department of Informatics & Telecommunications in the University of Athens, working on QoE provisioning in 4G/5G cellular networks.

Manish Narwaria obtained a Ph.D. degree in Computer engineering from Nanyang Technological University, Singapore in 2012. After that, he worked as researcher at IRCCyN-IVC lab, France before joining DA-IICT, India as an Assistant professor in Dec. 2015. His major research interests are in the area of multimedia signal processing with focus on perceptual aspects towards content capture, processing and transmission.