Multimed Tools Appl (2017) 76:22303–22327
DOI 10.1007/s11042-017-4831-6

Unsupervised deep learning for real-time assessment of video streaming services

Maria Torres Vega[1] · **Decebal Constantin Mocanu**[1] · **Antonio Liotta**[1]

Received: 11 November 2016 / Revised: 12 May 2017 / Accepted: 16 May 2017 /
Published online: 31 May 2017

Abstract Evaluating quality of experience in video streaming services requires a quality metric that works in real time and for a broad range of video types and network conditions. This means that, subjective video quality assessment studies, or complex objective video quality assessment metrics, which would be best suited from the accuracy perspective, cannot be used for this tasks (due to their high requirements in terms of time and complexity, in addition to their lack of scalability). In this paper we propose a light-weight No Reference (NR) method that, by means of unsupervised machine learning techniques and measurements on the client side is able to assess quality in real-time, accurately and in an adaptable and scalable manner. Our method makes use of the excellent density estimation capabilities of the unsupervised deep learning techniques, the restricted Boltzmann machines, and light-weight video features computed just on the impaired video to provide a delta of quality degradation. We have tested our approach in two network impaired video sets, the LIMP and the ReTRiEVED video quality databases, benchmarking the results of our method against the well-known full reference metric VQM. We have obtained levels of accuracy of at least 85% in both datasets using all possible cases.

Keywords Quality of experience · No-reference video quality assessment · Unsupervised machine learning · Deep learning

✉ Maria Torres Vega
m.torres.vega@tue.nl

Decebal Constantin Mocanu
d.c.mocanu@tue.nl

Antonio Liotta
a.liotta@tue.nl

[1] Department of Electrical Engineering, Eindhoven University of Technology, P.O. Box 513, 5600MB, Eindhoven, Netherlands

1 Introduction

Video content delivery clients require having their instantaneous needs fulfilled in an ever growing, ever-changing global network. Moreover, the exponential increase of video streaming types, the appearance of new high quality compression standards, and the broad variety of video content demand the use of scalable and general methods. This situation is requiring service providers to perform video quality assessment not only in real-time [3, 18, 29] but also in an adaptable and scalable manner. However, it is not enough anymore to assess the quality just on the network level, i.e. through Quality of Service (QoS) parameters. Packet losses, jitters or delays, while providing with a statistical representation of the network behavior, cannot accurately assess how the unpredictable network fluctuations may affect the perception of the final beneficiary of these services, i.e. the user's Quality of Experience (QoE) [1, 29, 30].

Video Quality Assessment (VQA) methods are drawn from human QoE [2, 21] and perception. Thus, VQA has traditionally been done by means of subjective studies and objective Full-Reference metrics (FR). However, the high requirements in terms of computation and time needed by both subjective and accurate FR metrics, make these unfeasible for deployment in real-time scenarios, such as in the mobile or the network management context. In these cases, Reduced-Reference (RR) and No-Reference (NR) metrics are the best suited metrics, due to the fact that they assess quality by means of certain features extracted from the received videos and the network conditions [34].

NR metrics aim to assess the quality of multimedia just by means of the received impaired material and measurements of external factors. This is a highly difficult task; thus, most NR metrics focus their attention on the specific behavior of certain distortions to make their assessment. Examples of these are the frame freezing approach of Huynh-Thu and Ghanbari [16], the blur tolerance analysis of Ferzli and Karam [11] or the generalized local binary pattern approach for image quality assessment of Zhang et al. [50]. Due to their aim to assess particular distortions, the accuracy of these metrics fails when other type of distortions affect user perception of QoEs. RR metrics have appeared to provide a compromise between FR and NR metrics.

Image or video statistics modeling have been considered for developing RR and NR quality metrics [20]. In [49] temporal motion smoothness of a video sequence was proposed to examine the temporal variations of the local phase structures in the complex wavelet transform domain [48]. In [48] both interframe and intraframe RR features are calculated based on statistical modeling of natural videos, which together with a robust watermarking approach, conform a very strong and accurate RR metric. Other approaches have focused on trying to model the distortion based on the encoding of the video sequences. Such examples are the MPEG-2 spatial and temporal features extraction of Wolf and Pinson [44] or the DCT measurement of Yang et al. [47], also for MPEG-2. Ma et al. presented in [20] a method which, combining aspects of both the spatial and the temporal perspectives, reaches high levels of accuracy for degradation derived from video compression in MPEG-2 and H.264. Also focused on the compression, specifically for H.264, Oelbaum and Diepold [26] proposed a method which, combining artifacts such as blur and blocking and making use of learning techniques (multivariate data analysis), achieves better results than PSNR. These last two methods although showing good results for degradations derived from compression, fail to provide good accuracy when dealing with videos impaired by real networks, which is the purpose of our method.

Prediction has been proposed to improve the accuracy of the assessment, without increasing the complexity in the client (the training part of the predictive process takes place in the

server provider in an offline manner). Promising examples of cognitive approaches are the Adaboost approach for assessing artifacts levels in videos, by Vink and de Haan [43]; the bitstream based artificial neural network, by Shahid et al. [35]; the artificial neural network for jerkiness evaluation, by Xue et al. [46]; and the regression framework for estimating the objective quality index (SSIM or PSNR), by Shanableh [36]. However, these approaches are usually based on supervised learning techniques, thus requiring labelled data to perform the offline training. Thus, they cannot be used when it comes to tackling the aforementioned combined requirements of real timeliness, scalability and adaptability.

In this work, we present a novel method based on unsupervised machine learning for real-time video quality assessment, which is sufficiently lightweight for mobile computing and general enough to deal with varied video types streamed through a broad range of network conditions. Making use of the outstanding density estimation characteristics of unsupervised deep learning methods, i.e. the restricted Boltzmann machines, in combination with light-weight NR features measured on the client side, our method is able to achieve accuracy levels close to the FR benchmarks.

Our method also provides a flexible and adaptable solution to assess relative quality degradation. It was tested on two large video-sets of network impaired videos, the LIMP Video Quality Database [41, 42] and the ReTRiEVED Video Quality Database [4, 29]. These two videosets complement each other. While the ReTRiEVED video set provides analysis on a broad range of conditions in four different categories (delay, jitter, throughput and packet loss) in a lower scale (184 MPEG2 videos), the LIMP database focuses on the combined effect of packet-loss constrained and bit/bandwidth compression, on 960 MPEG4 videos. To benchmark our solution, we selected the Video Quality Metric (VQM) [32], given its demonstrated good performance as a quality degradation assessment and its high correlation to the human visual system [8]. We have obtained overall correlations higher than 85% in both datasets.

The remainder of this paper is organized as follows. Section 2, provides background information about unsupervised learning, deep learning and the technicalities of the restricted Boltzmann machines. In Section 3, the proposed unsupervised-based video quality measurement method is presented. The two datasets under scrutiny are characterized in terms of NR and FR metrics in Section 4. Evaluations are presented in Sections 5 and 6 for the ReTRiEVED and the LIMP Video Quality Databases, respectively. Finally, Section 7 draws conclusions, highlighting our key contributions.

2 Unsupervised learning, deep learning and restricted Boltzmann machines

Not only accuracy, but adaptability, scalability and real-timeliness are crucial characteristics when video service provider decides among different quality assessment methods. Fast adaptability of the model when new videos are made available is fundamental. If the model used would belong to the supervised learning type (e.g. artificial neural networks, regression models), any newly released video sample would need to be labeled (its ground truth obtained) before inclusion in the model. This action would slow down the process and the adaptability feature would not be achieved. For this reason, in our work, we turned to focus on unsupervised learning (UL) methods. Second, to master the sheer scale of the problem, we selected Deep Learning (DL) techniques. Within this type of techniques, the Restricted Boltzmann Machines (RBMs) have demonstrated outstanding performance as density estimators [25]. This characteristic made us chose them for still images quality

estimation [22, 24]. In this work, we bring this notion to the highly complex video content delivery arena.

UL is the Machine Learning task of inferring a function to describe the hidden structure from unlabeled data [5]. Due to the difficulty of the task at hand, enhancing the predictive characteristics of these type of models has been a challenge addressed by multiple researchers. Among them, Deep Learning [6] is actively used, especially to address situations in which scalability is crucial. Deep learning methods attempt to model high-level abstractions in data by using a deep graph with multiple processing layers, composed of multiple linear and non-linear transformations. The estimation characteristics of the Restricted Boltzmann Machines [37] make them the perfect combination between unsupervised and deep learning (UDL).

RBMs are generative stochastic artificial neural networks that can learn a probability distribution over its set of inputs by means of only interlayer connections. It distributes its neurons in two layers: the visible ($\mathbf{v} = [v_1, v_2, .., v_{n_v}]$), which corresponds to the input features; and the hidden ($\mathbf{h} = [h_1, h_2, .., h_{n_h}]$), in which the hidden features are automatically extracted by the RBM model from the input data. Each visible neuron (i) is thus connected to any hidden neuron (j) by a weight ($W_{i,j}$), which is modeled according to the input (visible features). Both, the visible and the hidden neurons have associated a bias, a_i for the visible, and b_j for the hidden. Biases (both visible and hidden) together with the interlayer weights conform the RBM model $\Omega = \{\mathbf{W}, \mathbf{a}, \mathbf{b}\}$.

$$E(v, h) = -\sum_{i=1}^{n_v}\sum_{j=1}^{n_h} v_i h_j W_{ij} - \sum_{i=1}^{n_v} v_i a_i - \sum_{j=1}^{n_h} h_j b_j \tag{1}$$

To formalize an RBM, three main ingredients are required: (1) an energy function providing scalar values for a given network configuration, which can be computed by the sum over all possible interactions between neurons, weights and biases (1); (2) the probabilistic inference, which aims to determine the conditional distribution of the visible $\left(p(h_j = 1|\mathbf{v}, \Omega) = 1/\left[1 + e^{-\left(b_j + \sum_{i=1}^{n_v} v_i w_{ij}\right)}\right]\right)$ and, the hidden neurons $\left(p(v_i = 1|\mathbf{h}, \Omega) = 1/\left[1 + e^{-\left(a_i + \sum_{j=1}^{n_h} h_j w_{ij}\right)}\right]\right)$; and (3) the learning rules required for fitting the free parameters. Extensive research has been done on learning rules fit for RBMs [10, 39, 40]. However, almost all of them derive from the Contrastive Divergence (CD) method proposed by Hinton in [13]. Thus, in this research we make use of the CD method which is an approximation of the maximum likelihood learning. In order to set the learning rules, the update number, learning rate, momentum, and weights decay need to be set as thoroughly discussed in [14].

In the case of the video service provider, the characteristics of the original server video content are to act as visible neurons. Based on the video characteristics, the RBM is trained in the server (offline process). From this training, Ω is generated and transferred to the client. When the streaming session starts, the client can use the trained RBM model Ω to be used as the real-time degradation estimator, as it will be explained in the next section.

3 Unsupervised learning-based video quality assessment method

In this section we present our UDL-based method. Figure 1 shows the processes taking place both on the server (offline) and the client (in real-time).

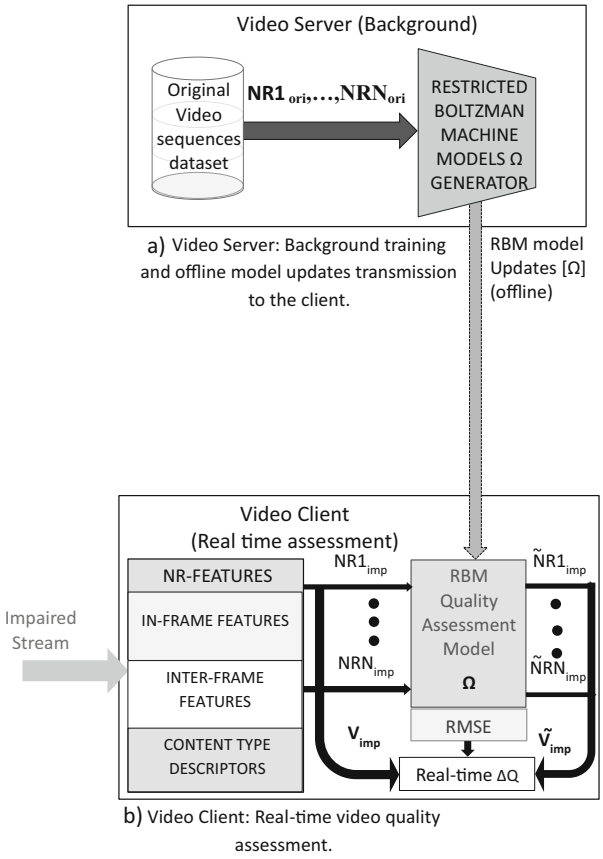


Fig. 1 Real-time UDL-based video quality assessment method

As for any prediction-based method, ours requires a training phase which takes place in the server side in an offline manner. In it, the server trains an RBM model (Ω) with the original video sequences available in the content delivery service. The training samples are video specific sets of NR features ($V_{ori} = [NR1_{ori}, ..., NRN_{ori}]$). These sets are composed by in-frame, inter-frame and content type descriptors (Table 1). This model (Ω) is transmitted to the client device, to be then used when video sequences are streamed to the client. adapted to the original available content in the content provider. On client session start, the model is transmitted to the client device and used when videos are being streamed. When a new video is made available in the content provider, the features of the video are extracted, the model is retrained (adapted to this new video) and an update is sent to the client. This is a process that requires very little overhead in the transmissions to the client, as is completely de-coupled from the online transmissions. This characteristic makes the method fully adaptable. Also, given the fact that the overhead required for training and updating/sending a new model is very small, our methodology envisions the possibility of individual video-trained models or even video and network condition trained models. This possibility will be evaluated in Sections 5 and 6.

On the other end of the transmission chain, when a new video sequence is received, the client performs a real-time extraction of the NR features ($[NR1_{imp}, ..., NR10_{imp}]$) required by the RBM model (Ω, the set of the free parameters of the model). These features

Table 1 Name, acronym and description of the ten NR features used in this method

Type	Name	Acronym	Descr.
INTRA	Blur Mean	BUM	Average level of blur per video frame Calculated following procedure of [9]
	Blur Ratio	BUR	Ratio of blur per video frame. Calculated following procedure of [9]
	Noise Mean	NOM	Average level of noise per video frame Calculated following the procedure of [9]
	Noise Ratio	NOR	Ratio of noise per video frame Calculated following the procedure of [9]
	Blockiness	BLO	Level of blocks per video frame Calculated following the procedure of [31]
	Jerkiness	JER	Video level of Jerkiness: temporal variations in the video display. Calculated following the procedure of [7]
INTER	Motion Intensity	MOI	Variation of intensity between adjecent frames. Calculated following the procedure of [7]
	Bitrate	BIT	Received bitstream bitrate Obtained directly from the ffmpeg client
CONTENT TYPE		SPI	In-frame Spatial Information
	Spatial Information	COX	Calculated following the procedure of [29] Spatial complexity: level of detail or intricacy contained within an image or frame $C = \frac{Bits_I}{2*10^6*0.91^{QP_I}}$ Where $Bits_I$ are bits of coded Intra (I) frames and QP_I represent the average I-Frames quantization parameter. Values obtained from the ffmpeg client [19]
		TEI	Inter-frame Temporal Information
	Temporal Information	MOT	Calculated following the procedure of [29] Video Motion: Amount of movement in the video $M = \frac{Bits_P}{2*10^6*0.87^{QP_P}}$ Where $Bits_P$ are bits of coded Inter (P) frames and QP_P represent the average P-Frames quantization parameter. Values obtained directly from the ffmpeg client [19]

They are divided in three different categories: intraframe, inter-frame and content type descriptors

are set as input to the RBM model. The model outputs the estimated values corresponding to the trained model, i.e. the estimated values for the impaired version of the video ($\tilde{V}_{imp} = [N\tilde{R}1_{imp}, ..., N\tilde{R}N_{imp}]$). Finally the quality degradation (ΔQ) is calculated as the Root Mean Squared Error (RMSE) [17] of the impaired measured values (V_{imp}) and the RBM reconstructed values ($\tilde{V}_{imp}$). Through this procedure, our method provides a measurement of the delta of degradation inflicted by the compression and transmission. Thus our measurement follows the same trend as the RMSE, where 'zero' denotes no degradation and 'one' full degradation.

Depending on the type of videos, network conditions envisioned and characteristics of the service, the RBM modeling and the NR metrics can be slightly changed and modified. In addtion, given the low overhead that the transmission of a model to the client inflicts in the network, our method allows to create condition or video dependent models. This characteristic further improves the adaptability of our unsupervised learning-based methodology. First, in order to formalize the RBM, it is crucial to choose appropriate learning rules to be used for fitting the input values into the model [10, 39, 40]. The most used approach is the Contrastive Divergence (CD) method proposed by Hinton [13], which performs an approximation of the maximum likelihood learning. The update number, learning rate, momentum, and weights decay together determine the learning rules [14]. Our solution makes use of this well-known learning method.

Second, the set of NR metrics used both for training the UDL model in the server and assessing the quality of the received will depend on the type of video content in the server, the network conditions to tackle and the compression and bandwidth constraints. In its more general version, our method assesses a set of 10 NR features in the frame (blur, noise and blockiness), inter frames (jerkiness and motion intensity) and video content descriptors (bitrate, spatial and temporal information). Table 1 presents the name, acronym and description of each of these 10 NR features.

In order to select the features to characterize the videos, we built on state-of-the-art research on NR features and artifacts which can be measured in real-time and have been demonstrated to show correlation to human perception. In order to characterize the videos as accurately as possible, our method performs measurements on the bitstream, inter and intra frame and on the content type. After a thorough research on the state-of-the-art and also based on some of our prior research [41], we ended up with 10 NR features: one on the bitstream, five on the frame, two on the inter-frame and two content characteristics.

Blur, noise, blockiness, ringing or temporal impairments have been quantified for measuring the end-user's quality [34]. For the NR feature selection we decided to measure frame's noise and blur (mean and ratio of both metrics) [9] and the level of blockiness [31, 45]. The temporal artifacts are measured by the video's motion intensity (the movement of video objects between frames by means of the compared level of intensity) and the inter-frame jerkiness [7].

Intuitively, quality is related to the bitstream bitrate (i.e. the number of bits received per time interval), whereby higher bitrates lead to better quality. However, this relation is highly non-linear, following a psychometric curve [12]. Earlier studies (some from us [23]) have shown how the parameters of the perception curve vary considerably across video types, compression values, bitrate etc. Bitrate is therefore a critical input to derive the prediction model.

Finally the content type can be expressed in terms of the spatial and temporal perceptual information [29]. In [41], for MPEG4 videos we defined those two parameters in terms of the scene complexity (spatial) of the frame and the motion (temporal) of the video. While the spatial and temporal perception are calculated by means of frame and inter-frame measurements [29], the complexity and motion are calculated on the bitstream based on the numbers of I and B frames [15, 19]. Our method measures either complexity and motion (for MPEG4 and other I,B, P frames based encoding) and in case that this is not possible (such as for MPEG2 and other older encoding), measurements of spatial and temporal perceptual information are performed.

Depending on the complexity and variety of the videos, either the full set of features or a clear subset of them can be used. Details on the selections of the subsets can be found in Section 4. We believe our methodology represents a new way to perform learning-based

real-time quality evaluation, in which prior data labelling is not required. This, in addition to the low overhead of the transmission of the model through the network (enabling condition and /or video based models), provides a compact, flexible and adaptable methodology for real-time quality assessment. This type of assessing method is envisioned to be used as a feedback loop for multimedia content providers or autonomic network management.

4 Video datasets characterization and benchmarking

The purpose of the evaluation of our UDL-based method is to assess its capabilities to measure the delta of degradation inflicted by the compression and network transmissions. Most of the video datasets available in the literature have been distorted by means of synthetic impairments (such as with network simulation tools or manually impairing the frames) [33]. However, for our assessment we had the aim to look for realistically network impaired video sets, i.e. by means of network emulators or real test-beds. With this purpose, we found the ReTRiEVED video quality database [4, 29]. This dataset is composed by 184 videos encoded to MPEG2 and impaired by delays, delay jitters, bandwidth and packet loss. From it, we assessed the most affecting impairments, i.e. bandwidth compressions and packet loss, but also got the trigger to start and develop our own dataset, the LIMP video quality database [41, 42], which provided a more extensive analysis (693 MPEG4 videos) of the effects of packet loss and bandwidth compression. The evaluation on both video set provides a complete evaluation assessment, not only of the performance on the general conditions (ReTRiEVED) but also with the extremely lossy networks cases (LIMP).

The video sets are characterized following the same procedure we presented in [41]. First, the quality is benchmarked. In order to benchmark the quality, our initial idea was to use subjective studies. However, although these are the best methods to understand the subject's absolute perception, their time and complexity requirements as well as their biased nature, make them unfit as a benchmark to objectively assess the degradation of quality inflicted by compression and networks on real-time video services. For this reason we turned our view to objective, FR metrics. Among them we selected VQM [32], given its well-known correlation to subjective studies [8, 32] while keeping computational complexity and time within certain boundaries. Second, the NR metrics are obtained in all the impaired videos. Finally, the Pearson correlations (PCC) between NR assessments and the benchmark quality are obtained. The correlations are evaluated both in terms of the overall value as well as per video and impairment levels.

The purpose of this process was two-folded. First, we aimed to understand the accuracy of NR metrics in detecting the delta of degradation inflicted by compressions and network impairments. Second, we wanted to pre-select a set of NR metrics which could, in turn, be used as input to the UDL-based method. Section 4.1 shows the characterization of the ReTRiEVED Video Quality Database [29]. Section 4.2 presents the same characterization of the LIMP Video Quality Database [41].

4.1 ReTRiEVED video quality database

The ReTRiEVED Video Quality database [4, 29] is composed by 184 test video-sequences obtained from 8 different original sources. These videos (encoded to MPEG2, with a duration of 10 seconds) are characterized by a broad range of spatial and temporal information, which allows drawing general conclusions out of the assessment. The 8 original videos are subjected to practical transmission impairment scenarios generated by a Network Emulator

(NETEM) and Video LAN [28]. Packet loss rate, jitter, delay, and throughput were the considered distortions resulting from video transmission, whereas their values were chosen based on ITU and ETSI recommendations [28, 29].

As we previously explained, the purpose of benchmarking the datasets is to understand how well the low complexity real-time metrics (either deterministic or machine learning based) correlated to the ground truth quality. For this reason, we picked VQM, an objective FR metric, as our benchmark quality instead of dealing with individual scores. This is the case, also for this video set, even if a small (40 subjects) subjective study was performed to it. As an additional confirmation of the suitability of VQM as an alternative for the subjective studies, we performed a preliminary fit of the MOS of this set to the VQM indexes on it (Table 2).

While each of the columns of the Table shows the PCC results for each of the videos subjected to a particular impairment, the rows present the averaged PCC per video type. Overall correlations across all network conditions and video types are shown in the last column and the last row, respectively. Looking at the Table, we can see that VQM shows an overall correlation of more than 70% in all the dataset. However, the standard deviation is close to 40%. Looking for a reason for this misbehavior, we looked at the impairments individually. While the results on jitter (column 2), throughput (column 3) and PLR (column 4) are very high, reaching values between 85 and 95%, the overall correlation in the case of the delay barely reaches 20%. Now if we try to pinpoint the cause of it, we can find it in each of the videos. The low correlation comes from videos 6 and 7, which show full anticorrelation to the MOS in the case of delays (roughly −70% for both videos). In addition, video 2 also shows heavy anticorrelation (−32%). The reason for this very low correlations with delay comes from the fact that the perception of delays is a very subjective task, where some people detect it faster than others. Thus, it becomes highly biased and with a very unpredictable behaviour. Our conclusion form this analysis was that apart from the imperfect fit of the delays. This imperfect correlation between objective (VQM) and subjective (MOS) metrics is, however, well-known and certainly within the acceptability boundary. Our conclusion from this study was that VQM could be confidently used as ground-truth benchmark for the remainder of our study, i.e. to evaluate our method in the ReTRiEVED dataset.

Table 2 PCC correlations of the FR metric VQM to the subjective MOS for all videos of the ReTRiEVED data set averaged per video type and network condition

	Network Impairment				
Video Type	Delay	Jitter	Throughput	PLR	ALL
1	0.308	0.994	0.954	0.842	0.775±0.318
2	−0.32	0.747011	0.754	0.78	0.49±0.54
3	0.112	0.996	0.99	0.923	0.755±0.43
4	0.7553	0.971	0.974	−0.456	0.871±0.69
5	0.7323	0.895	0.928	0.879	0.859±0.087
6	−0.773	0.985	0.934	0.887	0.508±0.855
7	−0.42	0.98	0.98	0.828	0.59±0.68
8	0.507	0.95	0.99	0.86	0.83±0.23
All	0.113±0.566	0.9398±0.084	0.939±0.078	0.84±0.05	0.709±0.4

Table 3 PCC correlations of all the NR features to VQM for the ReTRiEVED video set, averaged for all impairments per video type

Video Type	BIT	BUM	BUR	NOM	NOR	BLO	MOI	JER	SPI	TEI
1	0.85	0.12	0.12	0.08	-0.1	0.03	0.4	-0.18	0.31	0.47
2	0.95	0.02	0.74	0.11	0.58	-0.65	0.69	-0.49	0.81	0.54
3	0.76	0.63	0.59	-0.09	0.76	-0.8	-0.38	-0.44	0.6	0.47
4	-0.41	0.13	-0.06	0.84	-0.64	0.94	-0.55	-0.6	0.905	0.4
5	0.85	0.81	0.69	-0.26	0.07	0.83	0.69	-0.21	-0.37	-0.16
6	0.81	0.47	0.45	-0.06	0.76	0.31	-0.04	-0.82	-0.16	-0.1
7	0.26	-0.63	0.46	-0.65	-0.48	0.78	-0.37	-0.26	-0.79	0.63
8	0.71	0.47	0.57	-0.28	0.23	0.76	-0.47	-0.43	-0.8	0.42
All	0.6 ±0.46	0.25 ±0.45	0.45 ±0.28	-0.038 ±0.43	0.15 ±0.54	0.28 ±0.69	-0.005 ±0.52	-0.43 ±0.22	0.0643 ±0.7	0.33 ±0.29

Cell colors give qualitative correlation levels: green (best), and red (worst)

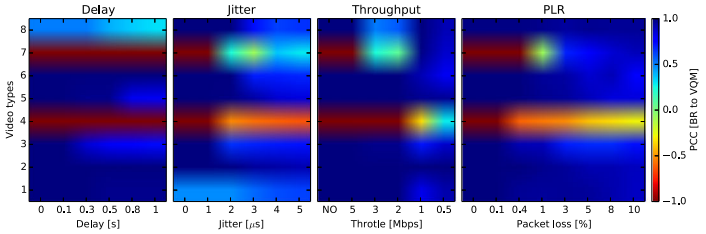
Having verified the validity of VQM as benchmark, we proceeded our experiments as described in [41]. We performed an accuracy analysis of the 10 NR metrics, by means of a PCC to the benchmark quality. Tables 3 and 4 provide with the overall correlation values per video type and per impairment, respectively. Per video type (Table 3), the best correlated is the bitrate (BIT). From the frame based features (BUM, BUR, NOM, NOR and BLO), the two blurs (BUM, BUR) and the blockiness (BLO) provide better overall performance (between 25 and 45%). Jerkiness (JER) and motion intensity (MOI) fail overall and in most video types. Finally, the content types (SPI and TEI) while failing in an overall measurement, provide good results in some of the cases, SPI is the best performer in the video type 4. Per network impairment (Table 4), while the bitrate (BIT) still provides the best correlations, it is interesting to point out that the accuracy of the metric for the delay condition barely reaches a 35% with a variability close to the 80%, while still being best. This situation made us reflect of the difficulty of the NR metrics to detect delay related impairments.

To further explore the accuracies of each of the NR metrics, Figs. 2 and 3 show the colormaps for each of the metrics. As expected the bitrate (BIT), Fig. 2a presents the best performance (more blue over all the set). However, video types 4 and 7 provide an almost anti-correlated pattern. These two videos follow quite an anti-correlated pattern for most of the metrics meaning that their behavioral pattern is difficult to assess by traditional NR metrics. However, the blur mean (BUM, Fig. 2b) seems to provide a better assessment (while far from perfect) of video type 7 (especially for the throughput and the PLR impairments). Further on, the spatial perceptual information (SPI, Fig. 3d) correlates better for the video type 4.

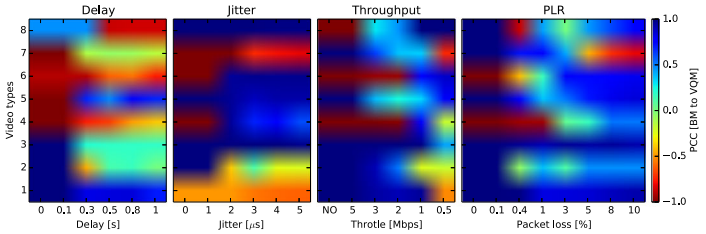
Table 4 PCC correlations of the NR features to VQM for the ReTRiEVED video set, averaged for all video types per impairment

IMP.	BIT	BUM	BUR	NOM	NOR	BLO	MOI	JER	SPI	TEI
Del.	0.35 ±0.86	-0.75 ±0.6	0.04 ±0.7	0.21 ±0.58	0.15 ±0.39	-0.046 ±0.81	0.12 ±0.55	0.03 ±0.74	0.008 ±0.77	-0.23 ±0.67
Jit.	0.5 ±0.5	0.35 ±0.73	0.46 ±0.75	-0.51 ±0.57	0.26 ±0.74	0.26 ±0.84	0.37 ±0.69	-0.3 ±0.56	0.125 ±0.89	0.64 ±0.39
Thr.	0.82 ±0.23	0.18 ±0.67	0.46 ±0.68	-0.32 ±0.58	0.11 ±0.77	0.47 ±0.64	0.23 ±0.69	-0.64 ±0.44	-0.1 ±0.74	0.53 ±0.48
PLR	0.706 ±0.41	0.55 ±0.58	0.82 ±0.16	0.47 ±0.53	0.09 ±0.74	0.377 ±0.82	0.005 ±0.67	-0.81 ±0.13	0.22 ±0.79	0.4 ±0.45
All	0.6 ±0.46	0.25 ±0.45	0.44 ±0.28	-0.04 ±0.43	0.15 ±0.54	0.27 ±0.69	-0.005 ±0.52	-0.43 ±0.21	0.06 ±0.7	0.33 ±0.29

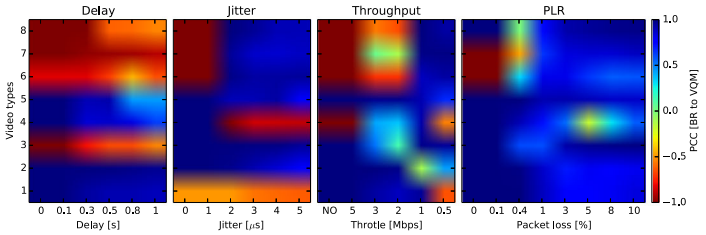
Cell colors give qualitative correlation levels: green (best), and red (worst)



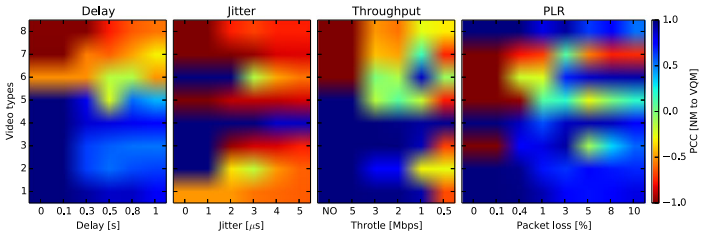
(a) Correlation maps for the NR feature bitrate (BIT).



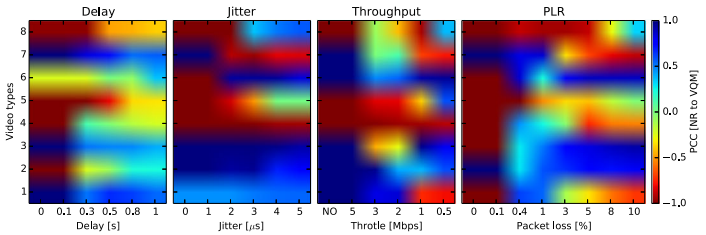
(b) Correlation maps for the NR feature blur mean (BUM).



(c) Correlation maps for the NR feature blur ratio (BUR).

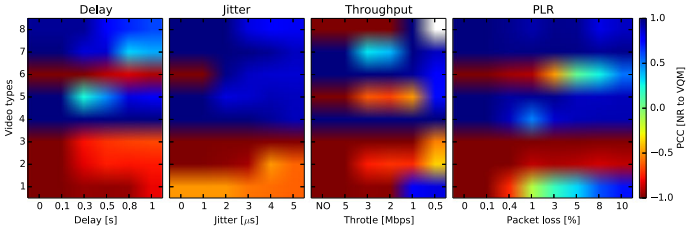


(d) Correlation maps for the NR feature noise mean (NOM).

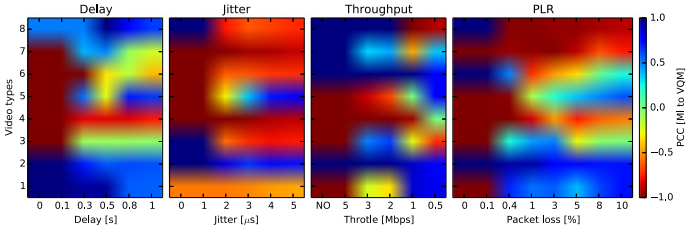


(e) Correlation maps for the NR feature noise ratio (NOR).

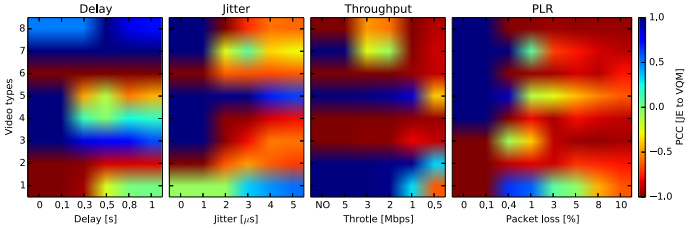
Fig. 2 Pearson correlation to VQM of 5 of the 10 NR metrics (BIT, BUM, BUR, NOM, NOR), considering the whole ReTRiEVED Video Quality Database [29]



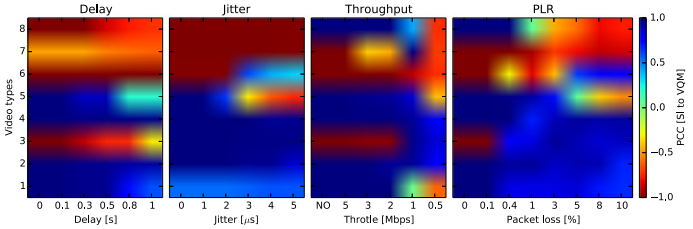
(a) Correlation maps for the NR feature blockiness (BLO).



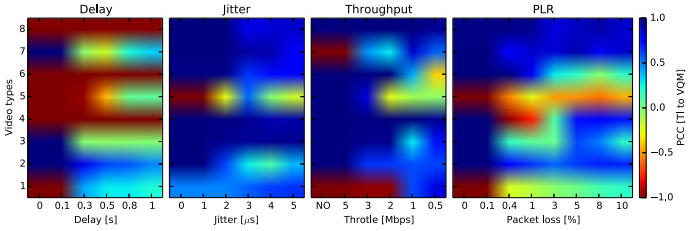
(b) Correlation maps for the NR feature motion intensity (MOI).



(c) Correlation maps for the NR feature jerkiness (JER).



(d) Correlation maps for the NR feature spatial perceptual information (SPI).



(e) Correlation maps for the NR feature temporal perceptual information (TEI).

Fig. 3 Pearson correlation to VQM of 5 of the 10 NR metrics (BLO, MOI, JER, SPI, TEI), considering the whole ReTRiEVED Video Quality Database [29]

With this knowledge we concluded to try to combine the measurement of BIT, BUM and SPI for our general RBM model. However, given the variety of behavioral patterns among the videos and network conditions, we also considered the possible need of generating video based RBM models or even network condition and video based models. Section 5 presents the results of applying this derived knowledge to our UDL-concept.

4.2 LIMP video quality database

The LIMP Video Quality database's purpose is to further explore the effect of lossy networks (packet losses) and bandwidth constrains on video quality [42]. In [41] we used it to present an in-depth analysis of the accuracy of NR metrics on videos impaired by compression and packet loss. In this paper, we use a subset of the LIMP video quality set to evaluate our UDL-method's performance. This Section provides a description, summary and analysis of this subset of the LIMP video quality dataset.

The set consists of 9 high quality videos (bs1, mc1, pa1, pr1, rb1, rh1, sf1, sh1, tr1) from the Live Quality Video Database [33] (10 seconds, 25fps, 768×432), encoded at MPEG-4 part 10/H.264 to 7 bitrates levels (640, 768, 1024, 2048, 3042, 4096, and 5120 kpbs). Each of these 63 videos (9 videos, 7 bitrates) was streamed in a controlled network environment (using the PacketStorm Hurricane II network emulator [27]) and subjected to 11 levels of packet loss (0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5, and 10%). We focused on the effect of packet loss due to its being the most impairing network condition [29, 38]. This makes a total of 693 different videos impaired by packet loss on which to assess the accuracy of our UDL-based quality method.

Table 5 presents the results of the correlations of the 10 NR measured features to VQM. This video set encoding (MPEG-4 part 10) allows measuring the spatial and temporal perception by means of the scene complexity and video motion as defined it in Section 3, i.e. using the numbers of I and B frames of the decoder.

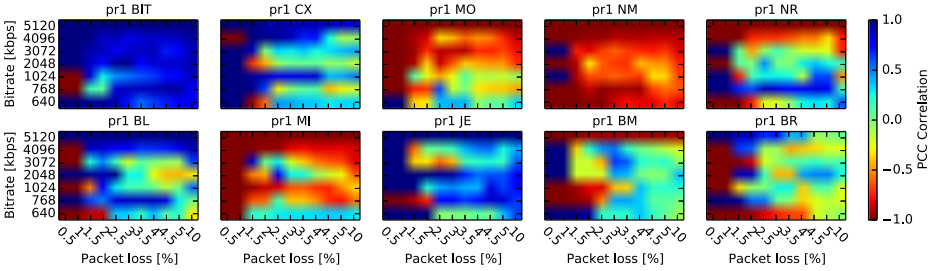
As in the case of the ReTRiEVED dataset, the bitrate (BIT) is the feature that better correlates to the benchmark quality, although BIT cannot be used independently. Yet, the correlation is far from perfect and some of the videos barely reach 60% correlation. Complexity (the content type regarding the spatial perception) obtains the second-best performance, with close to a 50% overall correlation and a 10% standard deviation. All the other metrics have worst overall performance, and show very large standard deviations. This can be explained by the fact that some of them while performing poorly for some of the videos, show high levels of correlation for others. Such as the case of the blur ratio (BUR), that while completely failing for videos such as bs1 or mc1, outperforms the bitrate for rb1 and pa1.

Table 5 PCC correlations to VQM of the ten NR metrics (BIT, BUM, BUR, NOM, NOR, BLO, MOI, JER, COX, MOT)

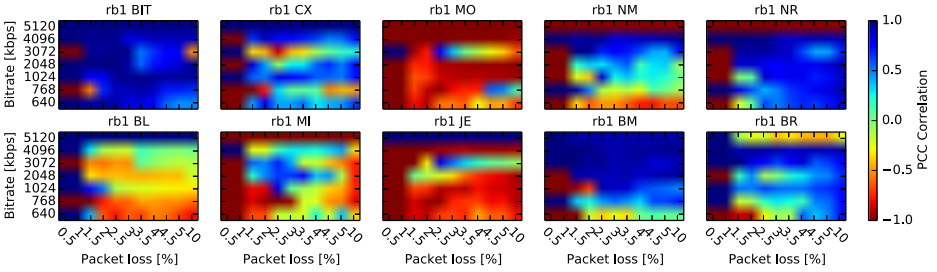
Video Type	BIT	BUM	BUR	NOM	NOR	BLO	MOI	JER	COX	MOT
bs1	0.88	-0.009	-0.267	-0.577	0.062	0.131	-0.9	-0.65	0.226	-0.23
mc1	0.89	-0.11	-0.74	-0.76	0.412	0.325	-0.114	0.37	0.72	-0.28
pa1	0.85	-0.021	0.88	0.48	0.312	0.542	0.602	0.19	0.552	-0.35
pr1	0.885	-0.045	-0.06	-0.79	-0.3	0.348	-0.563	0.68	0.309	-0.5
rb1	0.63	0.575	0.71	-0.006	0.59	-0.148	-0.68	-0.51	0.655	-0.71
rh1	0.93	0.51	0.65	0.273	0.0544	-0.614	-0.645	-0.65	0.6	0.426
sf1	0.95	0.72	0.7134	-0.86	0.019	0.542	-0.795	-0.46	0.5	0.29
sh1	0.9	0.13	-0.554	-0.502	0.221	0.498	-0.359	0.444	0.231	-0.375
tr1	0.94	0.489	0.721	0.2725	0.4554	-0.167	0.092	0.72	0.515	-0.752
All	0.88	0.25	0.23	-0.27	0.2	0.16	-0.37	0.015	0.479	-0.28
	±0.094	±0.3	±0.6	±0.5	±0.26	±0.375	±0.46	±0.54	±0.17	±0.38

Cell colors give qualitative correlation levels: green (best), and red (worst)

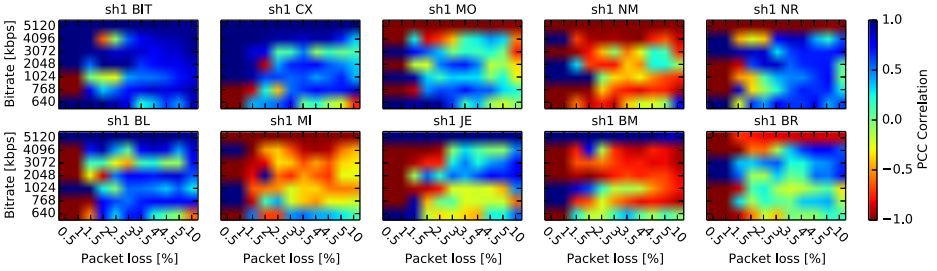
In order to discover the working limits of the various NR metrics, we analyzed the different video types individually (Fig. 4, shows videos pr1, rb1 and sh1 as example), with particular attention to compression level (Y axes) and packet loss (X axes). In Fig. 4, maximum correlation to VQM is shown in dark blue, while maximum anti-correlation is in dark red. Again we see that, even if bitrate leads to the highest correlation of all the metrics, it also fails for some videos and conditions. Regarding the other metrics, drill-down (instead of being averaged across the whole dataset) has allowed us to understand the conditions (and videos) under which the different metrics provide good performance. In this way, an RBM based on the bitrate in combination with some of the other metrics (based on the video under scrutiny) would make the perfect combination for our UDL-based method.



(a) Correlation maps for the video pr1.



(b) Correlation maps for the video rb1.



(c) Correlation maps for the video sh1.

Fig. 4 Pearson correlation to VQM of the 10 NR metrics (BIT,COX, MOT, NOM, NOR, BUM, BUR, BLO, MOI, JER), considering bitrates between 640 and 5,120 Kbps and packet losses between 0.5 and 10%. Video types: (**a**) Park run (pr1); (**b**) River bed (rb1); and (**c**) Shields (sh1). The original (unimpaired) videos were obtained from the Live Quality Video Database [33]. Network impairments were incurred by streaming videos through a network emulator (PacketStorm Hurricane II [27]) [41]

5 Evaluation on the ReTRiEVED video set: assessment of four network impairments

The purpose of this first evaluation was to understand the suitability of our method to detect the relative quality degradation inflicted by networks to videos, for a broad variety of network conditions. For this reason, we used the ReTRiEVED video quality database.

The RBM is structured with as many visible neurons as features (depending on the case studied) and 100 hidden neurons. Based on insights from previous works [13], the learning rate is set to 0.01, the number of CD steps to 1, the weight decay to 0.0002, the momentum to 0.9, and we trained to models for 100 epochs. For the evaluation, we considered three learning methodologies, as detailed in the three following subsections.

5.1 Case 1: a single model for all video types and impairment conditions

This first experiment evaluates the case-scenario in which the system does not have any prior information of the individual behavioural patterns of the videos. In such a situation, one single RBM would be trained in the server using as inputs all 8 original videos. This model could in turn be used to assess all the conditions (Table 6 row 1).

Table 7 shows the PCC correlations of our UDL-based method to VQM per video type and network condition. This configuration of the method obtains an overall correlation higher than the 90%. If we have a look at the accuracy of each of the network conditions independently, three out of four conditions are detected with very high correlation to the benchmark, accuracy values over 80%. Only for the delay, the assessment is worse. As we saw in the dataset characterization (Section 4.1), the delay impairment is the most difficult to assess. However, while the individual NR metrics were not able to achieve correlation values higher than 35% for the videos affected by delay (Table 4), with our UDL-based method, the performance reaches 67% overall correlation. However, the standard deviation, while lower than in the case of the classical NR metrics (85% reported in Table 4), is still far from good (70%).

In order to understand the reason behind the delay misbehavior and, to explore the performance details of our method, we again performed a per-impairment level correlation that can be seen in Fig. 5a. As it can be seen, the reason for the lower correlation on the delay

Table 6 Feature selection for the different cases tested on the ReTRiEVED Video Quality Database

Case	Features
Case I	BIT-BUM-SPI (1 model for all)
Case II	BIT-BUM-SPI (1 model per video type)
Case III	Delay:
	BIT-BUR-NOM (1 model per video type)
	Jitter:
	BIT-BUM-SPI (1 model per video type)
	Throughput:
	BIT-BUR-NOR (1 model per video type)
	Video 4 BIT-BUM-BUR-NOR-BLO-JER
	PLR:
	BIT-BUR-BLO-MOI

Table 7 PCC correlations of the UDL-based method in the case 1: One model for the whole dataset

Video Type	Network Impairment				
	Delay	Jitter	Throughput	PLR	All
1	0.9942	0.9854	0.99	0.992	0.9904±0.004
2	0.9999	0.9993	0.999	0.999	0.999±0.0003
3	0.3774	0.9997	0.998	0.995	0.8425±0.31
4	−0.999	0.9995	0.992	0.9984	0.497±0.9978
5	0.999	0.9998	0.998	0.998	0.9987±0.0008
6	0.999	0.9996	0.9994	0.999	0.9993±0.0003
7	0.999	0.9995	0.9995	0.999	0.9993±0.00028
8	0.992	0.9998	0.999	0.999	0.9975±0.0037
All	0.67 ±0.7	0.99788 ±0.003	0.9974 ±0.0032	0.998 ±0.0029	0.9158 ±0.1639

impairment comes from the low prediction on video type 4. Furthermore, this same video seems to give prediction issues in two of the other three conditions (jitter and PLR). This made us go one step further on the analysis and to put into practice the multiple RBM approach.

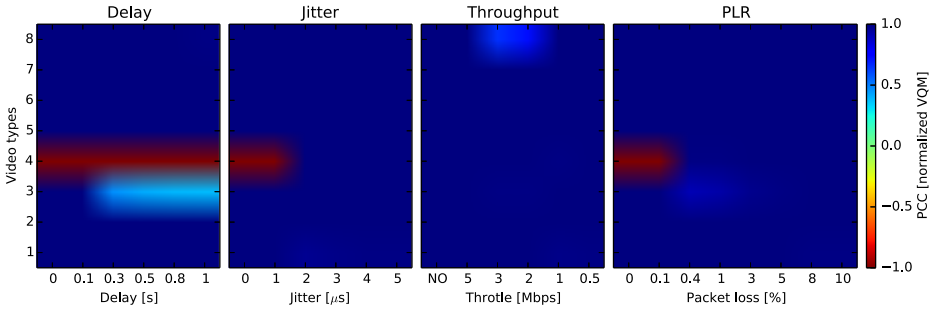
5.2 Case 2: individual models for each video type, for all impairments

In this analysis we trained one RBM per video type (Table 6, column 2). All of the RBMs are trained only on their own original video and tested on it as well. This means we have a total of 8 RBMs, which are then selected in the client according to the video streamed.

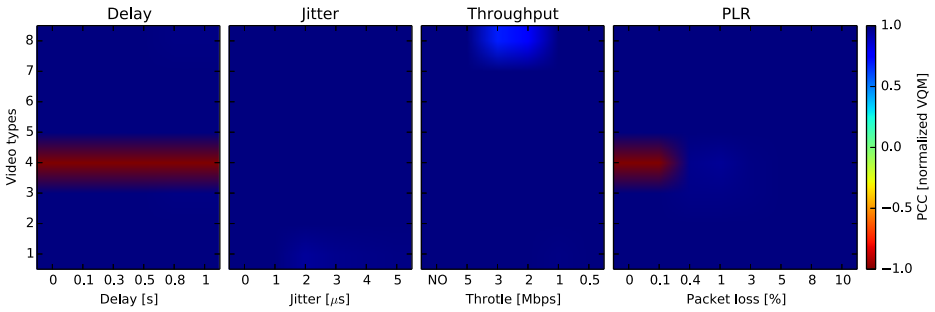
Table 8 shows the results of this evaluation. In this case, the overall correlation has increased by 4% compared to the first case, while the variability is kept nearly to the same values (under 20%). Most of the improvement corresponds to the better assessment of the video type 4. By contrast, if we check the corresponding Fig. 5b, while the performance of video type 4 has significantly improved on jitter and PLR, our method still shows full anti-correlation in the presence of delay for video type 4. These results made us understand the need for impairment-based models, which is the topic of the last evaluation.

5.3 Case 3: individual impairment based-models for each video type

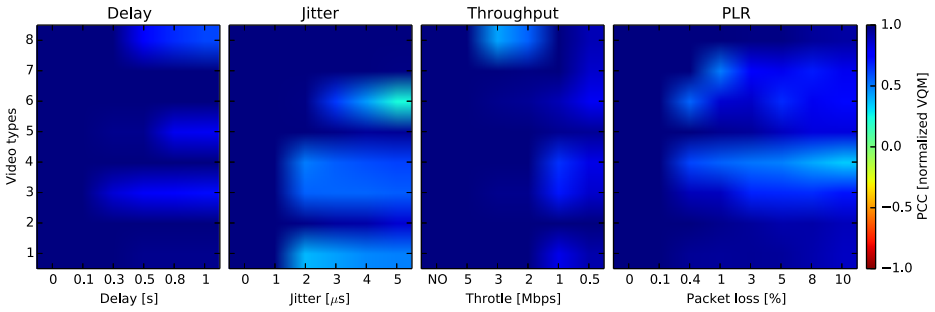
Finally, this analysis puts into practice the concept of one RBM model per video type and network impairment (Table 6, row 3). Table 9 shows the correlation results for this approach. The overall correlation has again increased by 4%, compared to the previous case (Table 8, last column and row), making this approach the best performer of the three analyses. Furthermore, the overall standard deviation has decreased from the 20% shown in the previous two cases to a final 9%). If we have a look at the colormaps of Fig. 5c, the



(a) Case 1: One model for all video types and conditions.



(b) Case 2: One model per video type, for all conditions (8 models).



(c) Case 3: One model per video type and network condition (32 models).

Fig. 5 PCC correlation to VQM of our UDL-based method for all the videos in the ReTRiEVED video set. Three cases: (**a**) 1 RBM model for all videos; (**b**) 1 RBM model per video type (8 models for 8 video types); and (**c**) 1 RBM model per video type and network condition (32 models)

prediction of video type 4 has significantly improved. In addition, the red spots (anticorrelation) on other parts of the colormaps have completely disappeared and only lighter blue is to be found, confined to the extreme conditions.

Encouraged by the performance of our UDL-method on the ReTRiEVED videoset, we set to evaluate it in the presence of extremely lossy networks. This was done with the LIMP database and the analysis is provided in the next section.

Table 8 PCC correlations of the UDL-based method in the case 2: One model for each video with the same configuration (8 models)

Video Type	Network Impairment				
	Delay	Jitter	Throughput	PLR	All
1	0.974	0.4976	0.9232	0.8811	0.819±0.2176
2	0.997	0.862	0.9955	0.8988	0.938±0.068
3	0.731	0.564	0.8843	0.748	0.732±0.131
4	−0.999	0.624	−0.247	0.298	0.4188±0.5285
5	0.796	0.956	0.9824	0.8433	0.8944±0.089
6	0.997	0.237	0.7655	0.8296	0.7074±0.3286
7	0.882	0.9895	0.82	0.87	0.8906±0.0711
8	0.716	0.9987	0.8427	0.923	0.87±0.12
All	0.887 ±0.123	0.716 ±0.278	0.745 ±0.4	0.998 ±0.2	0.784 ±0.194

6 Evaluation on the LIMP video quality database

The purpose of this evaluation was to analyse in depth, the working boundaries of our UDL-method in the presence of lossy conditions in combination to bandwidth constrains. For this reason we made use of the LIMP video set.

The RBMs are structured with as many visible neurons as features filtered (one per feature) and 50 hidden neurons (obtained through experimental analysis). Based on insights from previous works [13], the learning rate is set to 0.01; the number of CD steps to 1; the weight decay to 0.0002; the momentum to 0.9; and we trained to models for 100 epochs. As for the ReTRiEVED assessment (Section 5), we again considered three different learning methodologies.

Table 9 PCC correlations of the UDL-based method in the case 3: One model for each video and network configuration (32 models)

Video Type	Network Impairment				
	Delay	Jitter	Throughput	PLR	All
1	0.974	0.4976	0.9189	0.88	0.8189±0.216
2	0.996	0.862	0.998	0.898	0.938±0.07
3	0.7312	0.564	0.8633	0.714	0.7188±0.121
4	0.999	0.624	0.8119	0.344	0.695±0.279
5	0.7942	0.956	0.96	0.8346	0.8856±0.084
6	0.9988	0.237	0.79	0.742	0.6923±0.323
7	0.9978	0.9895	0.86	0.8	0.912±0.097
8	0.616	0.9987	0.9	0.932	0.8626±0.17
All	0.888 ±0.153	0.716 ±0.07	0.888 ±0.4	0.998 ±0.2	0.815 ±0.095

Table 10 Feature selection for the different cases tested on the LIMP Video Quality Database

Case	Features
Case I	BIT-BUM-NOR-BLO-MOI-COX-MOT (1 for all)
Case II	BIT-BUM-NOR-BLO-MOI-COX-MOT (1 per video type)
Case III	BIT-BUM-NOR-BLO-MOI-COX-MOT (3-fold cross-validation, 3 for 9)

6.1 Case 1: a single model for all video types

In this first set of experiments we focus on the simplest case in which the service provider would choose to train one single model to be used across all video cases and conditions. This corresponds to a theoretical worst-case scenario. Based on the dataset accuracy study presented in Section 4.2, we selected the feature set shown in the first column of Table 10.

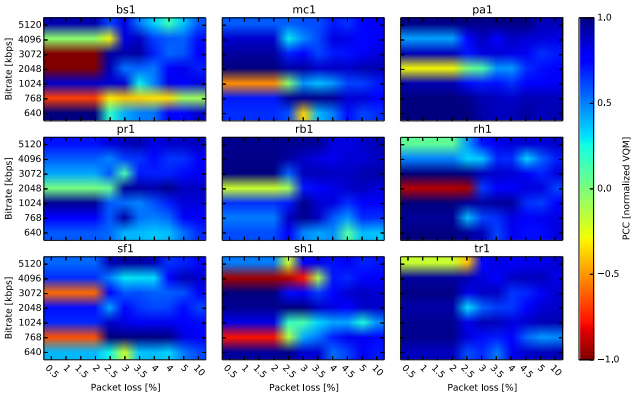
Table 11, first column presents the overall correlation results of our method to VQM for this first approach. This case shows an 86% correlation to the benchmark FR quality with a standard deviation of close to 2%. This is already an improvement, if we compared to the values showed by the NR features presented in Section 4.2 (apart from the bitrate, the NR metrics in Table 5 did not reach values higher than 50%). Furthermore, this method has brought stability to the solution, as can be seen first by the standard deviation, which stays in the area of 1% (the NR features of Table 5 showed variabilities of at least 10%). The accuracies have also dramatically improved for all the particular videos (all of them show values in the range of 82 to 87% while the NR metrics presented values from −75 to 90%).

In order to understand possible improvements on the feature selection and number of RBMs, we performed the colormap analysis shown in Fig. 6a. The colormaps show a situation in which most of the results are full blue (full correlation). Only some of the assessments

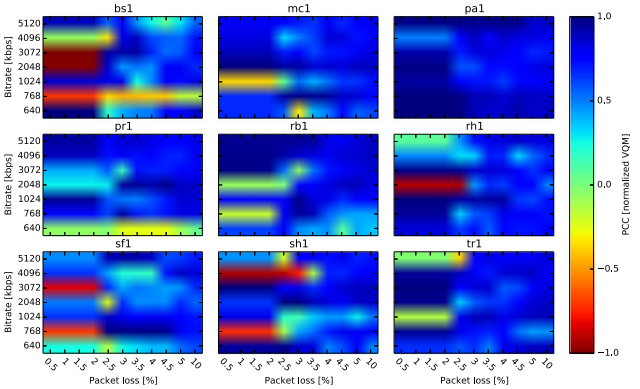
Table 11 UDL-based method PCC correlations to VQM for the LIMP Video Quality dataset in each of the three cases evaluated

Video Type	Case1	Case2	Case3
bs1	0.8192	0.88	0.8
mc1	0.8716	0.86	0.86
pa1	0.8653	0.864	0.85
pr1	0.8454	0.824	0.811
rb1	0.8508	0.82	0.83
rh1	0.8777	0.864	0.87
sf1	0.8827	0.8	0.869
sh1	0.8633	0.87	0.873
tr1	0.8657	0.85	0.855
All	0.86 ±0.0194	0.85 ±0.0289	0.844 ±0.0267

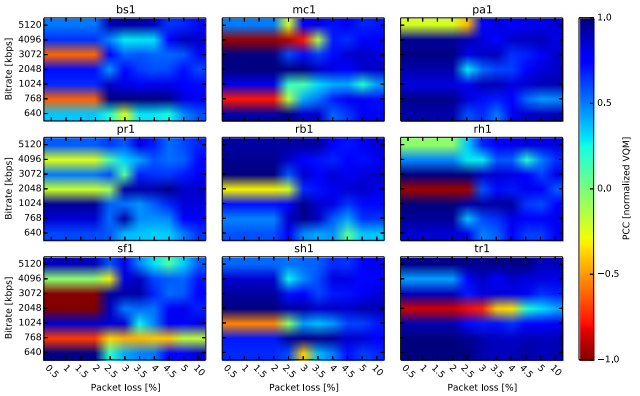
Results are averaged per video type. Column 1 presents the method's performance for Case 1, 1 model for the whole dataset (Section 6.1). Column 2 shows the results of Case 2, 1 video per video type (Section 6.2). Column 3 provides the results of the 3-fold cross validation test (Section 6.3)



(a) Case 1: A single model for all video types.



(b) Case 2: Individual models for each video type (9 models).



(c) Case 3: 3-fold crossvalidation (3 models for 9 video types).

Fig. 6 PCC correlation to VQM of our UDL-based method on the LIMP Video Quality Dataset, considering bitrates between 640 and 5,120 Kbps and packet losses between 0.5 and 10%. Three cases: (**a**) 1 model for all videos; (**b**) 1 model per video type and (**c**) 3-fold crossvalidation (3 models for 9 video types)

of certain videos tend to fail slightly. Such is the case of bs1 for high quality variants or certain bitrates of pa1. This encouraged us to carry out a second set of experiments in which each video type would be trained on its own RBM model.

6.2 Case 2: individual models for each video type

The purpose of this set of experiments is to explore the working possibilities of our method by means of using video defined models. In this case the video server trains 9 different RBMs, each of which is used on its own video type. As it can be seen in Table 10 column 2, the configuration of these RBMs is the same (same features) as the previous case and the only difference being the independent training of each video type.

The results of this experiment are found in Table 11 column 2 and in Fig. 6b. The overall correlation stays nearly on the same value as in the first evaluation (0.5% decrease). However, some of the videos show improved performance. This can be clearly seen in Fig. 6b, where videos such as the pa1 show no sign of red, uncorrelated regions. Counter-intuitively, this case shows a slightly worst performance than the first case and slightly higher variability (deviation went from 2 to 3%). The reason for this behaviour can be derived from the high lossy and very high compression conditions. When in the presence of high losses, the video under-scrutiny might suffer such a degradation (compared to its original version) that an RBM trained only on it provides less information than one trained on multiple types of those original videos.

6.3 Case 3: three-fold cross-validation

In the situation that a new video type is added to the content provider after the RBM assessment model has been computed, the service administrator can decide not to retrain a new model, but to use the already trained to assess the unseen content type. Herein, we assess the accuracy of the model on videos that have not been used in the training process, adopting a 3-fold cross-validation approach. In it, 3 RBMs are generated based on the original version of 3 videos. Then, each of this models is tested on 3 videos on which it was not trained for.

The last column of Table 11 and Fig. 6c show the results of this analysis. The overall correlation barely decreases (1% from the first case and 0.4% from the second), as it is for the standard deviation (only 0.6% worse). However, considering this is the worst-case scenario, the overall and per video accuracies are still within the acceptable limits (over 80%). These results demonstrate the strength of our method.

7 Conclusion

Clients of video streaming services require having their instantaneous needs fulfilled. In turn, the network and service providers require effective processes for real-time quality assessment that well align to human perception factors. Furthermore, given the increasing number of the available video content, in addition to the timeliness, methods are required to be adaptable to video content and scalable.

In this paper we have presented a novel approach for learning-based, real-time, adaptable, and scalable video quality assessment. By means of the outstanding density estimation features of the Unsupervised Deep Learning Restricted Boltzmann Machines and NR features, our method assesses in real-time the delta of degradation of videos, achieving a performance as accurate as the full-reference counter parts (VQM). In this paper we have presented

the evaluation on two different, network-impaired video datasets, namely ReTRiEVED and LIMP. Accuracies higher than 85% to the benchmark have been demonstrated in ReTRiEVED (184 videos, 4 different network conditions) [29] and in LIMP (960 videos under lossy conditions).

Adaptability is fulfilled by the unsupervised essence of our approach which, unlike other learning-based approaches requiring labeled data for training (i.e. FR or subjective index), does not need any labeling on the data and can adapt to new data, just as new samples arrive. In addition, our methodology allows the generation of multiple RBM models adapted to different network conditions and videos, providing a scalable solution. Finally, the scalability of our approach has also been demonstrated in the experimental evaluation, whereby only the original videos were required to obtain overall correlations higher than 85%.

Our methods and the findings presented in this paper open a new research venue for unsupervised deep learning-based, real-time quality assessment. Future work may follow three directions. First, we will extend our analysis to other available network impaired datasets. Second, we are currently researching for ways to improve the accuracy of our method (now below 90%). Finally, we intend to demonstrate the applicability of our real-time quality assessment to provide the feedback response in network management or real networks.

Acknowledgements This work has been carried out in the context of the European Research Council project BROWSE (Beam-steered Reconfigurable Optical-Wireless System for Energy-efficient communication - Grant 291632) and the ICT COST Action 3D-ConTourNet (IC1105).

References

1. Agboma F, Liotta A (2012) Quality of experience management in mobile content delivery system. Telecommun Syst 49(1)
2. Atzori L, Chen CW, Dagiuklas T, Wu HR (2012) Qoe management in emerging multimedia services. IEEE Commun Mag 50(4):18–19. doi:10.1109/MCOM.2012.6178829
3. Atzori L, Floris A, Ginesu G, Giusto DD (2014) Quality perception when streaming video on tablet devices. J Vis Commun Image Represent 25(3):586–595. doi:10.1016/j.jvcir.2013.08.013
4. Battisti F, Carli M, Paudyal P ReTRiEVED video quality database. http://www.comlab.uniroma3.it/retrieved.htm
5. Bengio Y (2009) Learning deep architectures for ai. Found Trends Mach Learn 2(1):1–127. doi:10.1561/2200000006
6. Bengio IGY, Courville A (2016) Deep learning. http://www.deeplearningbook.org. Book in preparation for MIT Press
7. Borer S (2010) A model of jerkiness for temporal impairments in video transmission. In: Proceedings of the second international workshop on quality of media experience (QoMEX). doi:10.1109/QOMEX.2010.5516155
8. Chikkerur S, Sundaram V, Reisslein M, Karam LJ (2011) Objective video quality assessment methods: a classification, review, and performance comparison. IEEE Trans Broadcast 57(2):165–182. doi:10.1109/TBC.2011.2104671
9. Choi MG, Jung JH, Jeon JW (2009) No-reference image quality assessment using blur and noise. Int J Electric Comput Energ Electron Commun Eng 3(2):184–188
10. Desjardins G, Courville A, Bengio Y, Vincent P, Delalleau O (2010) Tempered Markov Chain Monte Carlo for training of restricted Boltzmann machines. In: Teh YW, Titterington M (eds) Proceedings of the thirteenth international conference on artificial intelligence and statistics, Sardinia, pp 145–152

11. Ferzli R, Karam LJ (2006) Human visual system based no-reference objective image sharpness metric. In: ICIP. IEEE, pp 2949–2952
12. Gescheider G (2013) Psychophysics: the fundamentals. Taylor & Francis. https://books.google.es/books?id=gATPDTj8QoYC
13. Hinton G (2002) Training products of experts by minimizing contrastive divergence. Neural Comput 14(8):1771–1800. doi:10.1162/089976602760128018
14. Hinton G (2012) A practical guide to training restricted boltzmann machines. In: Neural networks: tricks of the trade, lecture notes in computer science, vol 7700. Springer, Berlin Heidelberg, pp 599–619
15. Hu J, Wildfeuer H (2009) Use of content complexity factors in video over ip quality monitoring. In: International workshop on quality of multimedia experience, 2009. QoMEx 2009, pp 216–221. doi:10.1109/QOMEX.2009.5246950
16. Huynh-Thu Q, Ghanbari M (2009) No-reference temporal quality metric for video impaired by frame freezing artefacts. In: Proceedings of the international conference on image processing, ICIP 2009, 7–10 November 2009. Cairo, pp 2221–2224.
17. Kendall MG, Stuart A, Ord JK (eds) (1987) Kendall's advanced theory of statistics. Oxford University Press, Inc., New York
18. Kordelas A, Politis I, Dagiuklas T (2016) Transport analysis and quality evaluation of MVC video streaming. Multimed Tools Appl 75(10):5619–5644. doi:10.1007/s11042-015-2530-8
19. Liotta A, Mocanu D, Menkovski V, Cagnetta L, Exarchakos G (2013) Instantaneous video quality assessment for lightweight devices. In: Proceedings of the international conference on advances in mobile computing, MoMM '13. New York, pp 525:525–525:531. doi:10.1145/2536853.2536903
20. Ma L, Li S, Ngan KN (2012) Reduced-reference video quality assessment of compressed video sequences. IEEE Trans Circ Syst Video Techn 22(10):1441–1456. doi:10.1109/TCSVT.2012.2202049
21. Menkovski V, Exarchakos G, Liotta A (2011) The value of relative quality in video delivery. J Mob Multimed 7(3):151–162
22. Mocanu D, Exarchakos G, Liotta A (2014) Deep learning for objective quality assessment of 3d images. In: 2014 IEEE International conference on image processing (ICIP), pp 758–762. doi:10.1109/ICIP.2014.7025152
23. Mocanu D, Liotta A, Ricci A, Vega M, Exarchakos G (2014) When does lower bitrate give higher quality in modern video services? In: Network operations and management symposium (NOMS), 2014. IEEE, pp 1–5. doi:10.1109/NOMS.2014.6838400
24. Mocanu D, Exarchakos G, Ammar H, Liotta A (2015) Reduced reference image quality assessment via boltzmann machines. In: 2015 IFIP/IEEE International symposium on integrated network management (IM), pp 1278–1281. doi:10.1109/INM.2015.7140481
25. Mocanu D, Mocanu E, Nguyen PH, Gibescu M, Liotta A (2016) A topological insight into restricted boltzmann machines. Mach Learn 104(2):243–270. doi:10.1007/s10994-016-5570-z
26. Oelbaum T, Diepold K (2008) Building a reduced reference video quality metric with very low overhead using multivariate data analysis. Syst Cybern Inf 6(5)
27. PacketStorm: Packetstorm hurricane ii network emulator. Available at http://packetstorm.com/packetstorm-products/hurricane-ii-software/
28. Paudyal P, Battisti F, Carli M (2014) A study on the effects of quality of service parameters on perceived video quality. In: Poceedings of the 5th European workshop on visual information processing, EUVIP 2014
29. Paudyal P, Battisti F, Carli M (2016) Impact of video content and transmission impairments on quality of experience. Multimed Tools Appl. doi:10.1007/s11042-015-3214-0
30. Paudyal P, Liu Y, Battisti F, Carli M (2016) Video quality of experience metric for streaming services. In: Image processing: algorithms and systems XIV. San Francisco, pp 1–5
31. Perra C (2014) A low computational complexity blockiness estimation based on spatial analysis. In: IEEE 22nd telecommunications forum. doi:10.1109/TELFOR.2014.7034606
32. Pinson MH, Wolf S (2004) A new standardized method for objectively measuring video quality. IEEE Trans Broadcast 50(3):312–322. doi:10.1109/TBC.2004.834028
33. Seshadrinathan K, Soundararajan R, Bovik AC, Cormack LK (2010) Study of subjective and objective quality assessment of video. Trans Img Proc 19(6):1427–1441. doi:10.1109/TIP.2010.2042111
34. Shahid M, Rossholm A, Lövström B, Zepernick H (2014) No-reference image and video quality assessment: a classification and review of recent approaches. EURASIP J Image Vid Process 2014:40. doi:10.1186/1687-5281-2014-40
35. Shahid M, Panasiuk J, Wallendael GV, Barkowsky M, Lövström B (2015) Predicting full-reference video quality measures using HEVC bitstream-based no-reference features. In: Seventh international workshop on quality of multimedia experience, QoMEX 2015. Pilos pp 1–2. doi:10.1109/QoMEX.2015.7148118

36. Shanableh T (2015) A regression-based framework for estimating the objective quality of HEVC coding units and video frames. Signal Process Image Commun 34:22–31. doi:10.1016/j.image.2015.02.008
37. Smolensky P et al (1987) Information processing in dynamical systems: foundations of harmony theory. In: Rumelhart DE, McClelland JL (eds) Parallel distributed processing: volume 1: foundations. MIT Press, Cambridge, pp 194–281
38. Suárez FJ, García A, Granda JC, García DF, Nuño P (2016) Assessing the qoe in video services over lossy networks. J Netw Syst Manage 24(1):116–139. doi:10.1007/s10922-015-9343-y
39. Tieleman T (2008) Training restricted boltzmann machines using approximations to the likelihood gradient. In: Proceedings of the 25th international conference on machine learning, ICML '08. ACM, New York, pp 1064–1071, doi:10.1145/1390156.1390290
40. Tieleman T, Hinton G (2009) Using fast weights to improve persistent contrastive divergence. In: Proceedings of the 26th annual international conference on machine learning, ICML '09. ACM, New York, pp 1033–1040, doi:10.1145/1553374.1553506
41. Torres M, Sguazzo V, Mocanu D, Liotta A (2016) An experimental survey of no-reference video quality assessment methods. Int J Pervasive Comput Commun 12(1):66–86. doi:10.1108/IJPCC-01-2016-0008
42. Torres Vega M, Liotta A LIMP-video quality database. https://www.tue.nl/index.php?id=53688
43. Vink JP, de Haan G (2011) No-reference metric design with machine learning for local video compression artifact level. J Sel Topics Signal Process 5(2):297–308. doi:10.1109/JSTSP.2010.2055832
44. Wolf S, Pinson MH (2005) Low bandwidth reduced reference video quality monitoring system. In: Proceedings of the international workshop video process. quality metrics consumer electr
45. Wu HR, Yuen M (1997) A generalized block-edge impairment metric for video coding. IEEE Signal Process Lett 4(11):317–320. doi:10.1109/97.641398
46. Xue Y, Erkin B, Wang Y (2014) A novel no-reference video quality metric for evaluating temporal jerkiness due to frame freezing. CoRR arXiv:abs/1411.1705
47. Yang S (2011) Reduced reference mpeg-2 picture quality measure based on ratio of dct coefficients. Electron Lett 47(6)
48. Zeng K, Wang Z (2010) Quality-aware video based on robust embedding of intra- and inter-frame reduced-reference features. In: Proceedings of the international conference on image processing, ICIP 2010. Hong Kong, pp 3229–3232. doi:10.1109/ICIP.2010.5651106
49. Zeng K, Z W (2010) Temporal motion smoothness measurement for reduced-reference video quality assessment. In: Proceedings of the IEEE international conference on acoustics, speech, and signal processing, ICASSP 2010. Sheraton Dallas Hotel, pp 1010–1013. doi:10.1109/ICASSP.2010.5495316
50. Zhang M, Muramatsu C, Zhou X, Hara T, Fujita H (2015) Blind image quality assessment using the joint statistics of generalized local binary pattern. IEEE Signal Process Lett 22(2):207–210. doi:10.1109/LSP.2014.2326399

Maria Torres Vega received her M.Sc. degree in Telecommunication Engineering from the Polytechnic University of Madrid, Spain, in 2009. Between 2009 and 2013 she worked as a software and test engineer in Germany with focus on Embedded Systems and Signal Processing. In October 2013, she decided to go back to academia and since then she is a Ph.D. student at the Eindhoven University of Technology. Her research interests include, but are not limited to, computer vision, quality of service and quality of experience in multimedia systems, autonomic management of wireless networks, artificial intelligence, and machine learning.

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.

Antonio Liotta (www.tue.nl/staff/a.liotta) holds the Chair of Communication Network Protocols at the Eindhoven University of Technology (NL), where he leads the Smart Networks team. Antonio is Editor-in- Chief of the book series Internet of Things: Technology, Communications and Computing (Springer) and associate editor of the Journal of Network and System Management (Springer) and of the International Journal of Network Management (Wiley). He explores network and data science in the context of complex communications and smart sensing, particularly Data Mining and the Internet of Things. He studies new network concepts that will allow us to tackle the IoT data deluge as a distributed datamining problem, bringing future internet research efforts alongside data science. He is the author of Networks for Pervasive Services: six ways to upgrade the Internet (Springer). His work has recently been featured in IEEE Spectrum in The cognitive Net is coming.