Neural network-based assessment of the impact induced in video quality assessment by the semantic labels

C. Hernandez, Bensaied Rania, Z. de la Lande Dolce, Mihai P Mitrea

Neural network-based assessment of the impact induced in video quality assessment by the semantic labels

C. Hernandez, Z. De La Lande Dolce, R. Bensaied, M. Mitrea
Telecom SudParis – Institut Polytechnique de Paris, ARTEMIS Department, SAMOVAR Laboratory
9, rue Charles Fourier, 91011 Evry, mihai.mitrea@telecom-sudparis.eu

Abstract

Subjective video quality assessment generally comes across with semantically labeled evaluation scales (e.g. Excellent, Good, Fair, Poor and Bad on a single stimulus, 5 level grading scale). While suspicions about an eventual bias these labels induce in the quality evaluation always occur, to the best of our knowledge, very few state-of-the-art studies target an objective assessment of such an impact. Our study presents a neural network solution in this respect. We designed a 5-class classifier, with 2 hidden layers, and a softmax output layer. An ADAM optimizer coupled to a Sparse Categorical Cross Entropy function is subsequently considered. The experimental results are obtained out of processing a database composed of 440 observers scoring about 7 hours of video content of 4 types (high-quality stereoscopic video content, low-quality stereoscopic video content, high-quality 2D video, and low-quality 2D video). The experimental results are discussed and confrontment to the reference given by a probability-based estimation method. They show an overall good convergence between the two types of methods while pointing out to some inner applicative differences that are discussed and explained.

1. Problem statement

The subjective visual quality assessment methods are essentially used to gauge the performance of multimedia systems with the help of responses obtained from observers who investigate the content displayed by the system under test. Thus, well-configured, consensual evaluation conditions are particularly required and the ITU Recommendations serve as a ground in this respect. Such recommendations are intensively used in a large variety of research studies, no matter their applicative field (device evaluation/calibration, compression, 3D image reconstruction, watermarking, *etc*.) or the type of content under evaluation (still 2D/3D images, video, 3D graphics, …).

Despite their consensual usage, some questions remain still open, mainly about the type of grading scale (continuous *vs*. discrete), the number of the levels on the grading scale (3, 5, 7, 11, …), and the semantic labels associated to those levels (e.g. *Excellent*, *Good*, *Fair*, *Poor* and *Bad*). Such questions are broad and may range from the very conceptual differences among and between the underlying psyche-cognitive mechanisms to the practical impact in the evaluation result. For instance, the impact of semantic labels is discussed and detailed in various research studies. On the one hand, some studies state that adjacent ITU labels are characterized by non-uniform semantic distances [1], [2]; yet, such a behavior is not quantified. On the other hand, some studies claim the contrary [3], *i.e.* that the semantic of adjacent ITU labels does not impact the results.

In order to elucidate this aspect, some previous studies published by the author team investigated the case of 5-levels grading scales (*Excellent*, *Good*, *Fair*, *Poor* and *Bad* labels) and brought to light a non-linear filtering formula that was coupled to a new statistical investigation method, thus obtaining the reference values for the semantic impact of these labels [4-6]. Just for illustration, it was thus spotted out that the *Excellent* label induce a reluctance effect of about 35% (that is, observers rather assign the *Good* label instead of the *Excellent* label, although they consider the content to be in top 20%). Conversely, the observers avoid the label *Bad* and tend to replace it by *Poor*. The inverse problem was also analyzed through statistical tools (that is, assuming an observer scores *Good* while he thinks the content is in top 20%, how can we statistically correct that score towards *Excellent*?) but the results are inherently bounded in precision [7].

The present paper has as objective to reconsider our previous study and to solve this issue through a neuronal network approach. This way, three types of results are targeted: (1) *an a posteriori* investigation of our previous results, (2) building an easier, more versatile tool for integrating such results into quality evaluation procedures and (3) eventually paving the way towards more accurate semantic impact cancelation.

The paper is structured as follows. Section 2 presents a panorama of the state-of-the-art results. Section 3 recalls the principles of the statistical-based investigation method and Section 4 presents the neural network architecture considered in the present study. Section 5 is devoted to the experimental results while Section 6 concludes the paper.

2. State of the art

The study in [8] investigates the reliability of 18 different discrete scales, with a variable number of classes ranging from 2 to 19. The reliability of the scales is discussed with respect to three factors that are a *priori* important in determining the number of alternatives to employ: (a) the proportion of the scale which is effectively considered by the subjects when scoring, (b) the duration required for testing, and (c) whether or not an "uncertain" category is provided. 360 students (20 students for each scale) are considered in the scoring sessions. The scale reliability was assessed by an analysis of the variance of the scores; in this respect, the Fisher's test is used. The conclusion is that, at least from this point of view, the scores are largely independent from the number of rating points on the scale: 16 out of the 18 scales examined did not differ significantly. The two exceptions correspond to the 2 level and 3 level grading scales. The results also show that the testing time increases with the number of levels on the scale while the usage of the "uncertain" category decreases as the number of rating steps increases.

The study reported in [9] investigates the reliability and validity of the scores assigned on a continuous scale and on discrete scales with 5, 7 and 11 categories. For both continuous and discrete scales, both labeled and unlabeled versions are presented to 30 subjects involved in the experiment. The conclusions are of different

types. First, the continuous scale is "*most pleasing"* to be used. Secondly, the results brought no evidence that the continuous scale would provide either more discrimination or better accuracy than the discrete scales. Concerning the discrete scales, the results brought to light that 5 or 6 categories should be considered for evaluation. It is also stated that even on a continuous scale, the scores assigned by the observers are somewhat clustered into 5 or 6 classes.

The problem of the optimal number response alternatives on a discrete scale is also raised in [10]. By using information theory tools, it is considered that the optimal number is the one that maximizes the information provided by respondents in a test, while minimizing the response errors or the likelihood of the random responses. In is thud concluded that using a large number of scale categories (higher than 9) results in no benefit, while a very small number of categories (less than 5) could produce a loss in accuracy.

The study in [11] compares a five-category discrete semantic differential scale with the corresponding unlabeled continuous scale. By differential scale it is understood a scale whose extremities are labeled. The discrete scale is presented to the user with intermediate marks, yet without any labels associated to these marks. The scores from 176 participants are investigated according to a paired Student's test. It is thus demonstrated that only 4 out of the 30 pairs of scale were significantly different. It is also concluded that similar evaluation result can be obtained from the five-category discrete rating scale and from a continuous scale.

The study in [12] compares a visual analog scale (similar to a continuous unlabeled scale), a graphic rating scale (similar to the standard ITU labeled continuous scale) and a five-point verbal descriptor scale (similar to the 5-point discrete category scale). 174 students participated in the subjective assessment. As a general conclusion, it is stated that assessments on the discrete scale have the highest level of stability; particularly, it is shown that both the verbal descriptor scale and the graphic scale assessments provided a better order consistency compared to the visual analog scale assessment. The results also show that an increased number of possible responses did not guarantee a higher sensitivity of the assessments.

The study in [13] provides to 149 participants a questionnaire concerning the service elements quality. The questionnaire used scales with a number of judgment category from 2 to 11, and a 101-point scale (from 0 to 100). It is thus shown that the scales that produce the least reliable scores are those with the fewest response categories. However, it is also found that a decrease in reliability is encountered for scales with more than ten response categories. The most reliable scores are found to be those from scales featuring between 7 and 10 response categories.

The issues related to the usage of semantic labels for visual multimedia content evaluation is also a topic of particular research interest, as testified by a large variety of studies [14]-[20]. In order to objectively assess the issue, previous studies carried out by the authors team established a common theoretical ground for subjective quality evaluation, encompassing both continuous and discrete scale evaluations and based on statistical ground, as described in the next section. These results will serve as a ground for comparison with the results based on neural networks, that will be presented in Sections 4-5.

3. Statistical-based investigation

The statistical-based methodological framework for assessing the semantic impact of the labels was presented in [4-6].

First, in order to bring to light whether such a semantic impact exist, a comparison (based in the Student's paired test) between the average values (representing the MOS) corresponding to the continuous (unlabeled) and discrete, semantically labeled scales is carried out. As the experiments demonstrate that the semantic impact exists, a procedure for its quantification it is also defined. Hence, the second step is to define an auxiliary discrete random variable, which is characterized by uneven partition but by equal a posteriori probabilities. By comparing the differences in the partition classes length between this auxiliary random variable and the random variable corresponding to the semantically labeled scale, the semantic impact is quantified (by defining an underlying coefficient). This auxiliary random variable is estimated trough repeated binomial tests.

The method is briefly presented in the sequel.

Step 1: Continuous scale evaluation

Step 1.1 Perform the continuous evaluation experiment

The human observers are asked to score the content on a continuous scale, *e.g.* between 0 and M, thus obtaining the data set $[x_1, x_2, \dots, x_N]$.

Step 1.2 Estimate the $p_X(x)$ probability density function (pdf)

This step can be performed by any continuous *pdf* estimation method, applied on the data set obtained in the previous step. For instance, we can consider a Gaussian mixture model, whose parameters are estimated under an EM (expectation-maximisation) criterion. The result of this step is the $p_X(x)$ *pdf*.

Step 1.3 Compute the Y random variable,

The Y r.v. is the discretization of X according to a partition $[0 = y_1, y_2, \dots, y_q = M]$ of the $[0, M\,]$ interval

This step is performed by applying a non-linear random variable filtering operation $Y = f(X)$, according to the $f(.)$ function, where:

$$y = f(x) = \begin{cases} 0, & x \le 0 \\ i, & (i-1)M/q < x < iM/q, i \in \{1,2,\dots,q\} \\ 0, & x > M \end{cases}$$

No particular constraint is imposed on the partition $[0 = y_1, y_2, \dots, y_q = M]$: it can be uniform (*i.e.* corresponding to even sub-intervals) or not. The result of this step is the $p_Y(y)$ *pdf* that will serve as basis for the comparison when identifying the semantic bias.

Step 2: Discrete, semantic-labelled scale evaluation

Step 2.1 Perform the discrete evaluation experiment

The human observers are asked to score the content on a q levels semantic-labelled discrete scale, *e.g.* on a scale with the following labels: "*Bad*", "*Poor*", "*Fair*", "*Good*", and "*Excellent*"; the result of this step is the data set $[z_1, z_2, \dots, z_N]$. Such an evaluation implicitly supposes that the evaluation grades are evenly distributed, *i.e.* that the user evenly divide the evaluation scale covering from the "worst" to the "best" content into q intervals, each of them corresponding to a semantic label.

Step 2.2 Estimate the $p_Z(z)$ probability density function

This step can be performed by any discrete *pdf* estimation method, applied to the data set obtained in the previous step. For instance, we can consider a frequency based estimation. The result of this step is the $p_Z(z)$ *pdf*, where:

$$p_Z(z) = \sum_{i=1}^{q} p_Z(i)\,\delta(z-i)$$

Step 3: Discrete vs. semantic-labelled scale evaluation

Step 3.1 Find the identity condition

This step searches for the $[0 = y_1, y_2, \dots, y_q = M]$ partition ensuring identity between the Y and Z random variables. In this respect, the $p_Y(y)$ can be considered as reference (theoretical value) and $p_Z(z)$ as an experiment vale to be validated through a goodness-on-fit test (*e.g.* the binomial test).

Step 3.2 Compute the relative variation of the partition intervals with respect to the uniform partition

This step computes the set of coefficients $\rho_{q-i}, i = 0,1, \dots q-1$, where:

$$\rho_{q-i} = \frac{y_{q-i} - y_{q-i-1}}{M/q}.$$

A unitary value for such a coefficient demonstrates that the related semantic label does not modify the evaluation - that is, an even partition $[0 = y_1, y_2, \dots, y_q = M]$ ensures the identity between Y and Z. A value larger than 1 indicates that the related semantic label makes the observer more likely to score that way while, conversely, a value lower than 1 shows that the related label makes the observers more reluctant in assigning that label when scoring.

A retrospective view on this method brings to light that we do not a priori expect all the labels to jointly suffer from a semantic impact for a same investigated content. Actually, we rather expect that for some good quality content the higher labels (*Excellent*, *Good*) to be impacted while for some low quality content the semantic impact is more likely to affect the lower labels on the evaluation scale (*Poor* and *Bad*).

4. Neural network architecture

The neural network used for this task is a 4-perceptron network, with a size 1 input layer for the continuous grade, 2 hidden layers and a size 5 output layer (for each possible label), Figure 1.

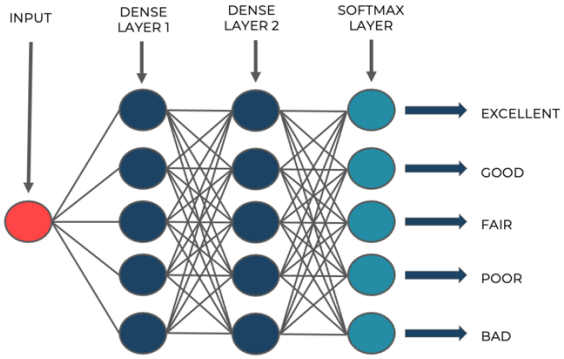

Figure 1: Network structure considered in the experiments

This architecture was chosen experimentally and incrementally to limit overshooting the complexity of the problem. Actually, note that while the dataset we shall process (details in Section 5.1) is very large (with respect to the state-of-the-art studies) for the quality evaluation investigation through conventional methods, it is quite small when compared to datasets usually processed in neural network applications.

The default and classical RELU activation function for the hidden nodes performed well enough to be kept and as for the problem itself, a SoftMax final activation seemed the most logical. The learning rate was experimentally set at 0.04 (value for which stability on validation was granted for all data) and it was used alongside an ADAM optimizer with default β values.

The model was trained for 100 to 400 epochs (according to the type of content – cf. details in section 5) with a batch size of 8 all of which granting satisfactory validation results for us to exploit (in the sense discussed in Section 5).

The network weights were not randomly initialized to keep track of the validity and scalability of the architecture.

Section 5 will present the numerical results and will also discuss the impact of some of the above-mentioned parameters.

5. Experimental results

5.1 Experimental testbed

For comparison's sake, we kept the same database we processed in our previous studies [4] – [7].

All the experiments reported in the present study are carried out on four types of content. The high-quality stereoscopic video content is sampled from the HD 3DTV corpus, and sums to a total of 2 hours 11 minutes and 24 seconds of stereoscopic video sequences (197'000 stereoscopic pairs), full HD encoded (1920×1080 pixels), shot in professional conditions. The low-quality stereoscopic video content is obtained by downgrading this content through general image processing operation that can be modelled by additive noise. The 2D content (both high quality and low quality) is obtained by considering only one view from the corresponding stereoscopic video content.

The general viewing conditions were set so as to meet the requirements expressed in ITU-R BT 500-11. A 47" LG LCD, full HD 3D monitor (1920×1080 pixels) with a 400cd/m² maximum brightness was used. The experiments involved 2 observers per session. The observers were seated in line with the center of the monitor, at a distance D equal to the height of the screen multiplied by factor $F = 3$ and defined as the Preferred Viewing Distance.

The test was conducted on a total of 110 naïve viewers (45 females and 65 males), with marginal knowledge of image quality assessment. The age distribution ranged from 20 to 37 years old with an average of 22. All observers are screened for visual acuity by using Snellen chart and the Ishihara test.

All the results reported in this study correspond to a single stimulus, 5 level grading scale a scale (*Excellent*, *Good*, *Fair*, *Poor* and *Bad*). At the beginning of the first session, from 2 to 5 training presentations are introduced to stabilize the observers' opinion. The data issued from these presentations are not taken into account in the results of the test. If several sessions are required, only two training presentations are done at the beginning of the next session.

5.2 Numerical results

For each of the four types of the investigated content, we shall first present the accuracy and the loss curves; then, Tables 1-4 will provide the values related to the ρ coefficient defined in Section 3.

The accuracy curves, for the four investigated types of content, are presented in Figures 2-5: the abscissa correspond to the epoch, the plots in red to the training accuracy while the plots in blue to the validation accuracy. The corresponding loss curves are presented in Figures 6-9.

It can be noticed that the accuracy curves related to 2D quality content have a particular behavior. Actually, an in-depth analysis of this case shown that the neural network behaves as a kind of binary classifier. The scores being quite evenly distributed among the central classes, the network does not manage to feature a global stability and some oscillations occur.

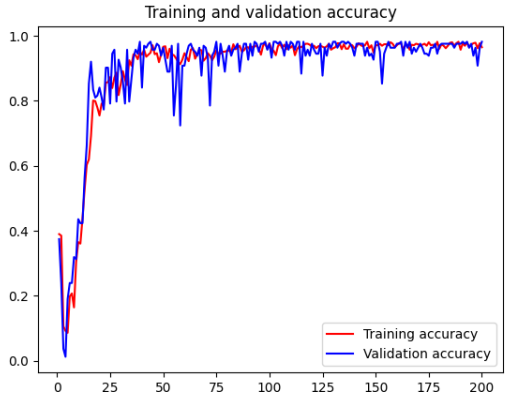

Figure 2: Accuracy plot for high quality stereoscopic video content

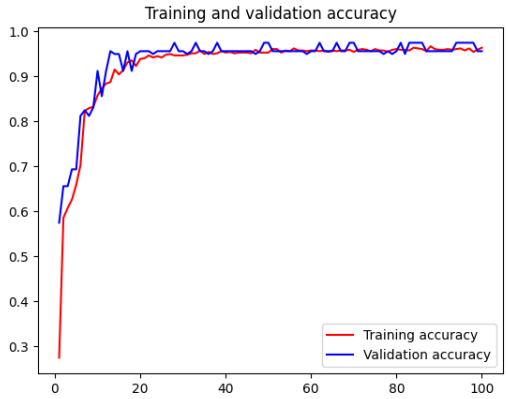

Figure 3: Accuracy plot for low quality stereoscopic video content

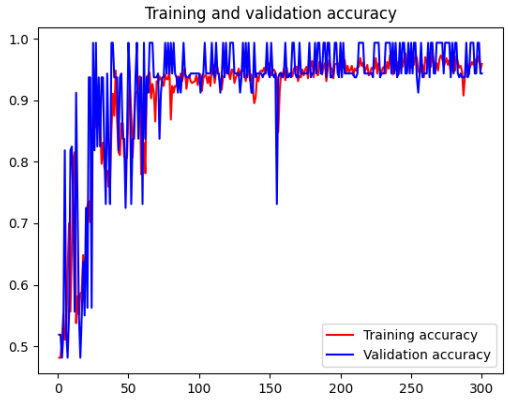

Figure 4: Accuracy plot for high quality 2D video content

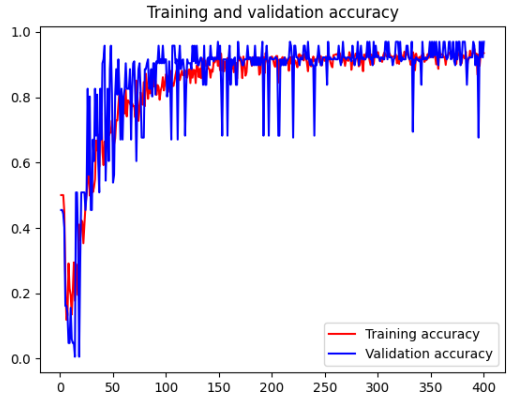

Figure 5: Accuracy plot for low quality 2D video content

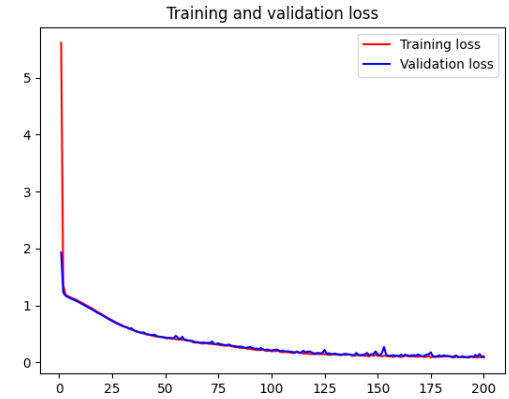

Figure 6: Loss plot for high quality stereoscopic video content

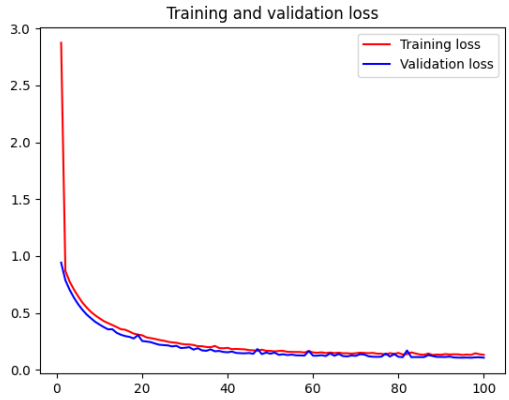

Figure 7: Loss plot for low quality stereoscopic video content

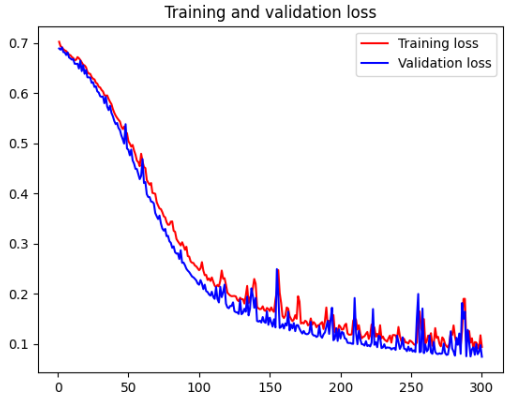

Figure 8: Loss plot for high quality 2D video content

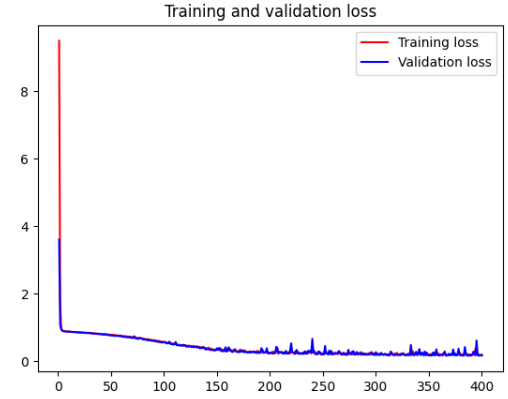

Figure 9: Loss plot for low quality 2D video content

The overall results related to the quantitative evaluation of the semantic impact are presented in the four tables below (Table 1 – 4), according to the type of content and, for each type of the content, to the label assigned by the evaluators. The grey-shadowed cells correspond to the results obtained by the state-of-the-art method [4-6] while the blue-shadowed cells to the neural-network based results.

The column labeled "*limits*" provides information about the numerical intervals imposed by the semantic labels on a would-be continuous scale graded between 0 and 100 - in absence of any semantic impact, these limits would be [0..20), [20..39), [40..59) and [60..79) and [80..100).

The column labelled ρ provides the values of the corresponding ρ coefficient, as defined in Section 3, Step 3.2.

Tables 1-4 show that some of the labels are not always outputted by the network even though they were present in the dataset. The potentiality of an overfitting was considered for a while but the network behaved the same and gave the exact same results when asked to provide 2 labels or 5 labels for the 3D Low Quality for example. The explanation might be that the network purposefully chooses to ignore those labels to give importance to those most present in the dataset; thus, the data limit is the greatest threat to our approach.

When the two approaches are successful, their results are positively correlated : assuming a data pre-filtering is applied, the relative differences are lower than 10%.

As a final experimental result, note that the network structure we chose in our study is quite sensitive with the learning rate, as illustrate din Figure 10.

Table 1: Impact of the semantic labels for high quality stereoscopic video content

		limits	ρ
High Quality Stereoscopic Video Content	*Bad*	[0..20)	1
		NA	NA
	Poor	[20..40)	1
		[0..39)	1.95
	Fair	[40..60)	1
		[39..59)	1
	Good	[60..87)	1.35
		[59..85)	1.3
	Excellent	[87..100]	0.65
		[85..100]	0.75

Table 2: Impact of the semantic labels for low quality stereoscopic video content

		limits	ρ
Low Quality Stereoscopic Video Content	*Bad*	[0..20)	1
		NA	NA
	Poor	[20..30)	0.5
		[0..29)	1.45
	Fair	[30..60)	1.5
		[29..100)	3.55
	Good	[60..80)	1
		NA	NA
	Excellent	[80..100]	1
		NA	NA

Table 3: Impact of the semantic labels for high quality 2D video content

		limits	ρ
High Quality 2D Video Content	*Bad*	[0..20)	1
		NA	NA
	Poor	[20..38)	0.9
		NA	NA
	Fair	[38..60)	1.1
		[0..88)	4.4
	Good	[60..87)	1.35
		[88..100]	0.6
	Excellent	[87..100]	0.65
		NA	NA

Table 4: Impact of the semantic labels for low quality 2D video content

		limits	ρ
Low Quality 2D Video Content	*Bad*	[0..20)	1
		NA	NA
	Poor	[20..31)	0.55
		[0..65)	3.25
	Fair	[31..60)	1.45
		[65..89)	1.2
	Good	[60..83)	1.15
		[89..100]	0.55
	Excellent	[83..100]	0.85
		NA	NA

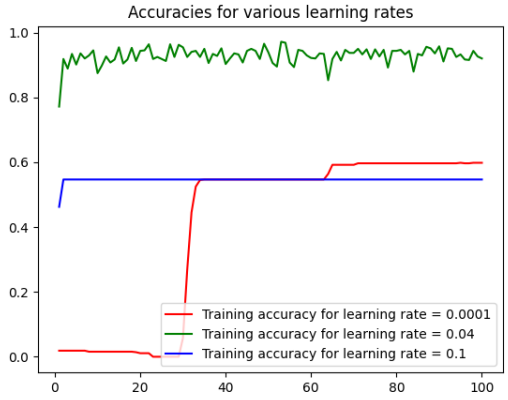

Figure 10: The impact of the training rate in the accuracy, for the case of high quality stereoscopic video content

6. Conclusion

The present paper is a study devoted to the use of basic neural network solutions for quantifying the bias induced in subjective video content evaluation by the semantic labels generally attached to discrete scales.

Under this framework, it is demonstrated that, despite the reduced size of the data set for a neural network application, when successful, the method produces results approximating by 10% the statistical -based state-of-the-art method. Yet, the main advantage of the neural network-based approach is its simplicity, thus becoming a good candidate for experimenters who would like to use such a tool in actual quality evaluation sessions.

Future work will be carried out in refining our experiments, e.g. by considering data augmentation techniques. Note that besides the direct expected results, data augmentation can also serve as guide when designing a new data-set allowing the increase of the results accuracy.

Future work will be also devoted to solving the converse problem, i.e. canceling the semantic label impact during an evaluation procedure.

References

[1] B.L. Jones, P.R. McManus, “Graphic scaling of qualitative terms”. SMPTE Journal, 1166–1171, 1986.

[2] N. Narita, “Graphic scaling and validity of Japanese descriptive terms used in subjective evaluation tests”, SMPTE J., vol. 102, no. 7, pp. 616–622, 1993.

[3] S. Zielinski, P. Brooks, and F. Rumsey F., “On the use of graphic scales in modern listening tests”, Proc. 123rd AES Convention, NY, 2007

[4] R. Bensaied, M. Mitrea, A. Chammem, T. Ebrahimi, “Continuous vs. discrete scale stereoscopic video subjective evaluation: case study on robust watermarking”, QoMEX, 2014 Sixth International Workshop on, pp. 238-244, Sept. 2014, Singapore

[5] R. Bensaied, M. Mitrea, “Assessing the impact of the semantic labels in subjective video quality evaluation”, in 11th IMA International Conference on Mathematics in Signal Processing, December 2016, Birmingham, UK

[6] R. Bensaied, “Subjective quality assessment: a study on the grading scales. Illustrations for stereoscopic and 2D video content", PhD Thesis, Pierre et Marie Curie University, Paris-France, July 2018

[7] M. Mitrea, R. Bensaied, P. Le Callet, “Semantic label bias in subjective video quality evaluation: A standardization perspective", IS&T Electronic Imaging 2019: Image Quality and System Performance XVI, Burlingame, California USA, Jan. 2019

[8] M.S. Matell, J. Jacoby, “Is there an optimal number of alternatives for Likert scale items? Study 1: reliability and validity”, Educational and Psychological Measurement, vol. 31, pp. 657–674, 1971.

[9] S.J. Mc Kelvie “Graphic rating scales—How many categories?”, British J. Psych., vol. 69, no. 2, pp. 185–202, 1978.

[10] E.P. Cox, “The optimal number of response alternatives for a scale: A review”, J. Marketing Res., vol. 17, no. 4, pp. 407–422, 1980.

[11] G. Albaum, R. Best, D. Hawkins, “Continuous vs. discrete semantic differential rating scales”, Psych. Reports, vol.49, pp.83–86, 1981.

[12] E. Svensson, “Comparison of the quality of assessments using continuous and discrete ordinal rating scales”, Biometrical J., vol. 42, no. 4, pp. 417–434, 2000.

[13] C.C. Preston, and A.M. Colman, “Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences”, Acta Psychologica, vol. 104, no. 1, pp. 1–15, 2000.

[14] K. Teunissen, “The validity of CCIR quality indicators along a graphical scale”, SMPTE J., vol. 105, no. 3, pp. 144–149, 1996

[15] A. Watson, and A. Sasse, “Measuring perceived quality of speech and video in multimedia conferencing applications”, Proc. ACM Multimedia Conf., pp. 55–60, 1998

[16] S. Winkler S., and R. Campos, “Video quality evaluation for Internet streaming applications”, Proc. SPIE Human Vision and Electronic Imaging, Santa Clara, CA, vol. 5007, pp. 104–115, 2003

[17] S. Winkler, “On the properties of subjective ratings in video quality experiments”, in QoMEX, San Diego, CA, 2009

[18] S. Pechard S., R. Pepion R., P. Le Callet , “Suitable methodology in subjective video quality assessment: a resolution dependent paradigm”, Proceedings of the Third International Workshop on Image Media Quality and its Applications, IMQA2008, 2008

[19] Q. Huynh-Thu, M. Brotherton, D. Hands, K. Brunnström, and M. Ghanbari, “Examination of the SAMVIQ methodology for the subjective assessment of multimedia quality”, in Proc. 3rd Int. Workshop Video Process. Consum. Electron., AZ, USA, 2007

[20] Q. Huynh-Thu, Q., M. Garcia, F. Speranza, P. Corriveau, A. Raake, “Study of rating scales for subjective quality assessment of High-Definition video” Broadcasting, IEEE Trans. on 57(1), 1–14, 2011

Author Biography

Celestin Hernandez and Zacharie De La Lande Dolce are engineering students (MS) at Telecom SudParis - Institut Polytechnique de Paris. Currently, Célestin starts his R&D career in the field of deep learning for video processing while Zacharie in the field of 3D reconstruction for VR.

Rania Besaied holds a PhD degree from Pierre and Marie Curie University in Paris (2018). She is currently R&D project manager at the ARTEMIS department of Telecom SudParis.

Mihai Mitrea holds and HDR degree Pierre and Marie Curie University in Paris (2010) and is currently Associate Professor at Telecom SudParis. He is vice-president of the Cap Digital's Technical Commission on Digital Content and serves as advisor for the French delegation at ISO/IEC JTC1 SC29 (a.k.a. MPEG).