CONTINUOUS QUALITY ASSESSMENT OF MPEG2 VIDEO WITH REDUCED REFERENCE

Patrick LE CALLET, Christian VIARD-GAUDIN, Dominique BARBA

Institut de Recherche en Communications et Cybernétique de Nantes – UMR CNRS 6597
Polytech’Nantes, rue Christian Pauc La Chantrerie
BP 60601 44306 NANTES CEDEX France

ABSTRACT

We propose an objective quality metric with reduced reference for color video that provides quality score highly correlated with human judgments. Since the size of the reduced reference data is very small (12 features per frame) and computationally simple, such a metric could be very useful to monitor quality for transmission services. The main novelties of the approach rely not only on the original way to proceed with color using a psychovisual colorspace but overall on a new promising idea to deal with the critical steps of comparison, combination and temporal pooling according to high level human visual system behavior. We show that time delay neural networks (TDNN) topology is well suited to achieve these latter steps.

1. INTRODUCTION

The main goal of image quality assessment is to find an automatic metric which provides computed quality scores well correlated with the ones given by human observers. Such metrics can provide quality control of the compressed images and more generally Quality of Service (QoS) for images transmission and especially broadcasting. Image quality metrics can be divided in three categories:
- full reference metrics (FR) for which the original image and the distorted image are required,
- reduced reference metrics (RR) for which a description of the original image into some parameters and the distorted image are both required,
- and no reference (NR) metrics which only require the distorted image.

As the whole original information is available, FR metrics often return the most accurate and robust quality measurements. They should be generic (suitable for any kind of distortions) and so allow coding schemes comparison. In a broadcasting purpose, with an emitter and a receiver, only RR and NR metrics are convenient for QoS monitoring since transmitting the whole reference image is not realistic at all. Ideally for such applications, NR metrics are the best choices since no extra data is added to the bitstream. But these metrics are quite complex to develop since the only way to get the original image is through distortions. RR metrics represent good alternatives since the reduced reference can be coded and embedded in the bitstream and therefore constitutes a practical approach to quality evaluation, as long as the reduced reference size is not too large.

This paper presents a new RR metric for color video quality assessment that provides quality values at a rate of two values per second according to the data obtained from subjective tests under a SSCQE protocol with hidden reference removal and used for the performance assessment of the proposed metric. The reduced representation of the original image (computed by the emitter and transmitted in conjunction with the coded image) is compared to the reduced representation of the decoded image (computed by the receiver) using an original method. The reduced representation computation can be considered as a generic method since the features used are mainly content based and so are independent from the type of distortions. However, the way these reduced representations are compared involves parameters which depend on the type of artifacts. These parameters are learnt by a training system.

Next figure describes the proposed approach. For each frame of both original and distorted video, we extract a set of features. In order to take into account color in our metric according to the human visual system, we add at the front end of the features extraction a decomposition into
the three components of the Krauskopf’s color space [1]. So, each feature is computed on each of the three components of the color space. Each feature represents a scalar value for the entire frame, so it corresponds to a full integration along the spatial dimension.

From the feature values, we proceed next to a step leading to combine and to compare features of the two sequences, and to pool them temporally on a determined time slot in order to produce a quality score regularly along the time axis (typically every 0.5 s). In order to construct the steps achieving combination/comparison and temporal pooling, we have made the following observations:

- the human visual system achieves a systematic local temporal analysis in order to produce high level perceptual descriptors of distortions;
- the subjective quality is constructed from the local analysis on a greater time slot. This effect is known as the memory effect.

Since these two contributions cannot be directly derived from psychophysics (too high level processing), we propose to use a neural network to learn these behaviors. We have selected a specific topology for the neural networks that fits very well with the process of local combination/comparison and temporal pooling: a time delay neural network (TDNN). TDNN is a specific case of convolutional neural network that enables invariance with respect to temporal shifting. It is based on three main ideas: shared weights, temporal window and delay. The goal of the TDNN is not to learn basically a signal but to extract the features from it. Layer by layer, we construct some new high level features coming from the combination and comparison of features of the under layer on a given temporal window. So we obtain a systematic local temporal analysis (combination and comparison on time dimension).

The last layer of the network is fully connected to the previous one, as with a standard multi-layer perceptron. It allows to emulate the memory effect, given the high level features for a larger temporal window.

2. FEATURES EXTRACTION

In order that the proposed metric incorporates aspects of the human vision, feature extraction is carried out on a perceptual color representation of the video sequence. It led to transform the YUV original images to a three perceptual component transformation: A, Cr1 and Cr2. Based on this representation, and according to the type of degradations that can affect the videos, one can imagine many different measurements. We have selected from the literature a set of N=4 features that are well suited in order to sum up the content of a frame. Three of them are totally content dependent (regarding frequency and temporal content). The last feature is more focused on distortion a priori related to blocking effect. Each of these 4 features is computed independently on the three perceptual components. Consequently, the global size of the feature vector describing every frame is $3 \times 4 = 12$ features.

Several authors have proposed some features allowing to represent a distorted video. In our case, it is important to consider only blindly extracted features (without reference image). In this context, we can find two classes of features, the first one related to distortions, the second related to the content. We have selected the following features:

- two features to represent the spectrum content (so related to content);
- one feature to represent the temporal evolution (so related to content);
- one feature of blocking effect (related to distortions).

2.1. Frequency content features: GHV and GHVP

The two first features, termed as $GHV$ and $GHVP$, are derived from the work of [2]. They have been previously elaborated to detect the blurring artifacts but are also sensitive to tiling distortions. These two features are computed from the two-dimensional histogram $SIH(r, \theta)$ where $r$ is the magnitude of the gradient vector, and $\theta$ is the orientation of the gradient vector with respect to the horizontal axis and $SIH(r, \theta)$ is the number of pixels in the gradient image whose gradient radius and angle is $r$ and $\theta$, respectively.
The feature \( GHV \) whose value increases as the number or sharpness of horizontal and vertical edges increase is given as:

\[
GHV = \frac{1}{p} \sum_r \sum_\theta SIH(r,\theta) r,
\]

with

\[0 < C_a \leq r \leq C_b,
\]

and

\[
\theta = \frac{k\pi}{2}, \quad (k=0,1,2,3),
\]

where \( r \) and \( \theta \) are as defined above and \( C_a \) and \( C_b \) are clipping limits, and \( p \) is the number of pixels in the image.

In order to separate blurring from tiling, the \( GHVP \) feature that characterizes the edge content of the image without the inclusion of horizontal and vertical edges is also computed:

\[
GHVP = \frac{1}{p} \sum_r \sum_\theta SIH(r,\theta) r,
\]

with

\[0 < C_a \leq r \leq C_b,
\]

and

\[
\theta \neq \frac{k\pi}{2}, \quad (k=0,1,2,3)
\]

2.2. Temporal content feature: Power of frame difference

The next extracted feature, \( P \), is derived from the work of [3]. They consider the following distortions: flicker, jadder, moving blurred images, random noise and edge jitter, and define linear combinations of some distortion factors using properties of visual perception. These combinations, which are explicitly defined in their work, are based on the power of the frame difference images computed respectively on the original and on the distorted video sequences. In our work, we will just keep the computation of the power of the frame difference and use it as an input feature for the NN. It will be the responsibility of the NN to model the distortions. The following computations are proceeded:

Frame difference:

\[
d(t, m, n) = I(t, m, n) - I(t-l, m, n)
\]

Power of frame difference:

\[
P(t) = \sum_{m,n} \left\{ d(t,m,n) \right\}^2
\]

2.3. Blocking measure: \( B \)

This last measurement is mainly dedicated to exhibit blocking effects. It is based on the method described in figure 2 and proposed by [4].

They apply 1-D FFTs to horizontal and vertical difference signals or rows and columns in the image to estimate the average horizontal and vertical power spectra. Peaks in these spectra due to 8x8 block structures are identified by their locations in the spectra. The power spectra of the underlying non-blocky images are approximated by median-filtering these curves. The overall blockiness measure, feature \( B \), is then computed as the difference between these power spectra at the locations of the peaks. Integration of masking effects is possible with this scheme while it has not been used in our implementation.

3. FEATURES COMBINATION AND TEMPORAL POOLING

3.1. TDNN topology

Multi Layer Perceptron (MLP) are the most common neural network architecture encountered. They consist of several layers of fully-connected hidden units. However, when the number of input variables is quite large, which is the case with image application, this architecture leads to several tens of thousands of weights. Such a large number of parameters increases the capacity of the system but at the same time requires a larger training set. To overcome the dilemma between small NN with low capacity and large NN that appear overparameterized with respect to the size of the training database, one can design specific architectures that aim to detect and combine local features.
The idea is to perform the same kind of computation at every place in the video stream based on a local receptive field. This is typically the principles involved with convolutional NN (CNN). In our case, the convolution kernels will be defined along the temporal axis, leading to the so-called Time Delay Neural Network (TDNN). TDNNs are well suited to sequential signal processing. They allow to preserve the sequential nature of data, in contrast with standard MLP where the topology of the input is entirely ignored. On the contrary, video sequences have a strong local structure: frames that are temporally nearby are highly correlated. Local correlations are the reasons for the well-known advantages of extracting and combining local features before processing temporal objects. With CNN, a given neuron detects a particular local feature of the video stream. It performs a weighted sum of its inputs followed by a non-linear squashing function (sigmoid). Its receptive field is restricted to a limited time window. The same neuron is reused along the time axis to detect the presence or absence of the same feature at different position of the video stream. A complete convolutional layer is composed of several feature maps, so that multiple features can be extracted at each temporal position. This weight sharing technique greatly reduces the number of free parameters and hence trained networks run much faster and require much less memory than fully connected NN. With local receptive fields, neurons can extract elementary visual distortions in videos. These distortions are then combined by the subsequent layers in order to detect high-order features.

In addition to the TDNN layers, the upper layers are standard fully connected layers. With this application, the last layer consists of a single neuron fully connected to the previous layer; the output of this neuron will be trained to estimate the DMOS value as it has been provided by human observers.

A detailed view of the TDNN architecture is presented in Figure 3.

### 3.2. Application and training

Different values for the TDNN parameters have been experimented. However, some of these parameters have been set once for all. For example, the number of layers has been set globally to 4, including 2 layers for the local feature extraction sub-system, and 3 for the fully connected NN at the upper level, which correspond to one input layer – actually, the output layer of the TDNN sub-system, one hidden layer and an output layer with a single neuron. The value of T, which refers to the number of frames involved in the computation of a score, has also been kept always the same; we supposed that at last the 5 last seconds influence the perceived quality at a given time, consequently, we set T to 5s × 25 f/s = 125 frames.

Accordingly, for every new frame we compute the four features (GHV, GHVP, P, B) previously described on each of the three perceptual components A, Cr1 and Cr2. Hence the input vector for the NN has a size of 4 × 3 × T for a NR video assessment system and twice this size for a RR system, where T is the number of frames taking into account in the computation of a scoring.

In order to train the TDNN we have used subjective quality assessment material provided by TDF (Télé Diffusion de France). This material is composed of different sequences of about three minutes long encoded and/or transcoded in MPEG2 at different rates. From four different initial contents (named Road, Cooking, Horses and Football), we get different distortions for each (see table 1 for details). For all of these video sequences associated with different distortions, we have access to subjective assessments provided by human observers, which consist of a quality rating sampled twice a second. It is available as a DMOS with an associated Confidence Interval (CI) obtained according to subjective measurement procedures. The reference sequence has been produced by a MPEG codec at a bit rate of 8 Mbits/s which ensures a very high quality, very close to the original videos.
To share this material between a training and a testing database, a separate video content is associated with the training and the testing video sequences. Considering the four sets of video contents (Road, Cooking, Horses and Football), sequentially, we trained the system on three out of four of these sequences and tested it on the remaining sequences. For example: Road, Horses, Football (13 videos) compose the training set, and the test set is composed of the remaining group of videos, in this case: Cooking (4 videos). With this procedure, we are sure that the sets of images of the training set and test set come from disjoint video contents.

### 4. RESULTS

As a measure of performance of the proposed objective scoring method, three main indicators are presented. One is the root mean squared error on the test set between DMOS and the output of the TDNN. The second one is the Linear Correlation Criteria (LCC), which expresses the monotony between DMOS and objective scoring. The third one represents the percentage of the marks given by the TDNN that lays inside half the confidence interval margins (Count). Results are given in table 2 for the four possible training sets.

The average mean quadratic error of 0.086 on a range of 1 indicates a rather good accuracy according to the accuracy on DMOS. Considering the correlation, the average value reaches 0.942, which is very good for a reduced reference metric. For the four training conditions, we obtain good performances on the test set indicating that the method is robust.

The results we obtained when the cooking videos are used for testing, whereas Football, Horses and Tancarville were used for training, are presented in more details in figure 4. It corresponds successively to four different bit rates: a) 3 Mbits/s, b) 5 Mbits/s, c) 2 Mbits/s and d) 3.5 Mbits/s. The DMOS are computed from the scores given by the human observers, and the DMOS+CI and DMOS-CI curves are also displayed, where CI represents the 95% Confidence Interval. The dashed-dotted line is the output of the TDNN, which represents the predicted score from the reduced reference features. In most of the situations, the predicted score lies inside the confidence interval range and follow quite well the human assessment. For every sequences, the root mean squared error between DMOS and TDDN output is displayed (rmse) and the percentage of objective marks that lies inside the [DMOS+CI, DMOS-CI] interval is also given. Except in the case of the most disturbed video at 2 Mbits/s, the predicted quality rate remains very close to the DMOS subjective marks.

### 5. CONCLUSION

In this paper, we have demonstrated that TDNN can be used to assess the perceived quality of video sequences by realizing a non-linear mapping between non subjective features extracted on the video frames and subjective grades. TDNN are so well suited to answer the problem of error pooling, we introduce a new way to use neural networks in the context of quality assessment (see [5] and [6] for others purposes). We have validated our approach using quite a large database that is composed of different video contents and different bit rates. On the test set, which was independent of the learning set, an average linear correlation criteria of 0.942 has been obtained between the output of the RR system and the subjective score provided by human observers. It shows a remarkable generalization capability of the neural network. The key factor of the proposed architecture relies on the set of convolutional neurons, which slides along the time axis sharing the same set of weights. They allow to perform the time integration function, which is not obvious to model and for that reason not always taken into account in other systems.

### 6. ACKNOWLEDGMENT

The authors wish to thank TDF for providing the databases used in the experiments related in this paper and Fabrice Alleau and Marti Vilarnau for their assistance in performing the experiments described in the paper.
7. REFERENCES


Figure 4 : Results for testing set cooking