Perceptual-based Quality Metrics for Image and Video Services: A Survey

Ulrich Engelke and Hans-Jürgen Zepernick
Blekinge Institute of Technology
PO Box 520, SE–372 25 Ronneby, Sweden
E-mail: {ulrich.engelke, hans-jurgen.zepernick}@bth.se

Abstract— **The accurate prediction of quality from an end-user perspective has received increased attention with the growing demand for compression and communication of digital image and video services over wired and wireless networks. The existing quality assessment methods and metrics have a vast reach from computational and memory efficient numerical methods to highly complex models incorporating aspects of the human visual system. It is hence crucial to classify these methods in order to find the favorable approach for an intended application. In this paper a survey and classification of contemporary image and video quality metrics is therefore presented along with the favorable quality assessment methodologies. Emphasis is given to those metrics that can be related to the quality as perceived by the end-user. As such, these perceptual-based image and video quality metrics may build a bridge between the assessment of quality as experienced by the end-user and the quality of service parameters that are usually deployed to quantify service integrity.**

I. Introduction

Multimedia applications are experiencing a tremendous growth in popularity in recent years due to the evolution of both wired and wireless communication systems, namely, the Internet and third generation mobile radio networks [1]. Despite the advances of communication and coding technologies one problem remains unchanged, the transmitted data suffers from impairments through both lossy source encoding and transmission over error prone channels. This results in a degradation of quality of the multimedia content. In order to combat these losses they need to be measured utilising appropriate quality indicators. Traditionally, this has been done with measures like signal-to-noise ratio (SNR), bit error rate (BER), or peak signal-to noise ratio (PSNR). It has been shown that those measures do not necessarily correlate well with quality as it would be perceived by an end-user [2].

Maximising service quality at a given cost is a main concern of network operators and content providers. Due to this, concepts such as Quality of Service (QoS) and Quality of Experience (QoE) [3], [4] have been introduced giving operators and service providers the capability of better exploitation of network resources that satisfy user expectations. In contrast to already standardised perceptual quality metrics for audio [5] and speech [6], the standardisation process for image and video seemed to have proceeded somewhat slower. This issue has also been recognised and addressed by the International Telecommunications Union (ITU). In 1997, two independent sectors of the ITU, the Telecommunication sector (ITU-T) and the Radiocommunication sector (ITU-R), chose to cooperate in the search for appropriate image and video quality measures suitable for standardisation. A group of experts from both sections was formed known as the Video Quality Experts Group (VQEG) [7]. The efforts which the VQEG has performed and the results are reported in [8], [9]. The application area for quality metrics is wide and can include in-service monitoring of transmission quality and optimisation of compression algorithms.

In this paper a survey and classification of contemporary image and video quality metrics is presented. A broad overview of available methodologies applicable to assess quality degradation occurring in communication networks is given. The survey is understood as a guide to find favorable metrics for an intended application but also as an overview of the different methodologies that have been used in quality assessment. Emphasis is given to those metrics that can be related to the quality as perceived by the end-user. As such, these perceptual-based metrics may build a bridge between QoE as seen by the end-user and QoS parameters quantifying service integrity.

The paper is organised as follows. In Section II classification aspects of quality measures are discussed. In Section III a class of metrics is reviewed that uses solely the received image respectively video for the quality evaluation. Similarly, in Section IV a class of metrics is considered that additionally utilises reference information from the original image respectively video. Finally, conclusions are drawn in Section V.

II. Classification of Quality Evaluation Methods

A. Subjective and objective methods

The evaluation of quality may be divided into two classes, subjective and objective methods. Intuitively one can say that the best judge of quality is the human himself. That is why subjective methods are said to be the most precise measures of perceptual quality and to date subjective experiments are the only widely recognized method of judging perceived quality [2]. In these experiments humans are involved who have to vote for the quality of a medium in a controlled test environment. This can be done by simply providing a distorted medium of which the quality has to be evaluated by the subject. Another way is to additionally provide a reference medium which the subject can use to determine the relative quality of the distorted medium. These different methods are specified for television sized pictures by ITU-R [10] and

1-4244-0857-1/07/$25.00 ©2007 IEEE

are, respectively, referred to as single stimulus continuous quality evaluation (SSCQE) and double stimulus continuous quality-scale (DSCQS). Similar, for multimedia applications an absolute category rating (ACR) and degradation category rating (DCR) are recommended by ITU-T [11]. Common to all procedures is the pooling of the votes into a mean opinion score (MOS) which provides a measure of subjective quality on the media in the given test set. Clearly, subjective quality assessment is expensive and tedious as it has to be performed with great care in order to obtain meaningful results. Also, subjective methods are in general not applicable in environments which require real-time processing. Hence, automated methods are needed which attempt to predict the quality as it would be perceived by a human observer. We refer to them as objective perceptual quality metrics. The existing methods have a vast reach from computationally and memory efficient numerical methods to highly complex models incorporating aspects of the human visual system (HVS) [12].

B. Psychophysical and engineering approach

Two general approaches have been followed in design of objective quality metrics which in [13] are referred to as the psychophysical approach and the engineering approach. Metric design following the former approach is mainly based on incorporation of various aspects of the HVS which are considered crucial for visual perception. This can include modeling of contrast and orientation sensitivity, spatial and temporal masking effects, frequency selectivity and colour perception. Due to the complexity of the HVS these models, and therewith the metrics, can become very complex and computationally expensive. On the other hand, they usually correlate very well with human perception and are usable in a wide range of applications. Fundamental work following the psychophysical approach has been performed in [14]–[20]. Methods following the engineering approach are primarily based on image analysis and feature extraction, which does not exclude that certain aspects of the HVS are considered in the design as well. The methods span from simple, numerical measures [21] to more complex extraction and analysis algorithms. The extracted features and artifacts can be of different kinds such as spatial and temporal information, codec parameters, or content classifiers. Simple methods are based on measuring single features whereas more complex algorithms combine various measures in a meaningful way. In any case, the metric outcomes can be connected to human visual perception by relating them to MOS obtained in subjective experiments.

C. Reference-based classification

Finally, we can classify quality metrics regarding their dependency on available reference information at the quality assessment equipment. The different methods that will be discussed are shown in Fig. 1.

In general, it is no problem for the HVS to judge the quality of a distorted visual medium without having any reference available. However, what seems to be so easy for the HVS is a highly complex task for a machine. Metrics following the approach of judging perceptual quality only based on the distorted medium are called no-reference (NR) or "blind" methods. These methods are readily applicable in a communication system as they would base the quality prediction solely on the received medium.

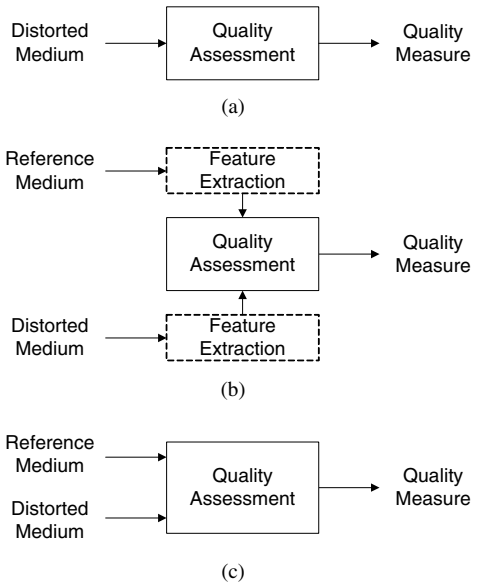

Fig. 1. Quality assessment methods: (a) No-reference method, (b) Reduced-reference method, (c) Full-reference method.

In order to quantify whether a change in quality between a reference and distorted medium has occurred, some degree of knowledge about the original medium would ease the related evaluation compared to using an NR method. This can be achieved by reduced-reference (RR) methods. Here, only a set of features from the reference medium is needed at the quality evaluation equipment instead of the whole medium itself. This set of features can then be transmitted piggy-backed with the medium or over an ancillary channel. At the receiver, the features can then be extracted from the medium and used along with the reference features for the quality prediction.

In cases where the reference is available at the evaluation equipment, one can use a full-reference (FR) method. These methods use the reference to predict the quality degradation of the distorted medium which eases the process substantially and provides in general superior quality prediction performance. It should be noted that most existing metrics following the psychophysical approach are FR methods [22]–[27]. The drawback of FR methods in a communication environment is that the reference is not available at the receiver where the quality assessment is performed. In the sequel, only existing NR and RR methods will be reviewed due to their applicability in communication systems.

III. No-reference Quality Metrics

The task of NR quality assessment is very complex as no information about the original, undistorted medium is available. Therewith, a NR method is an absolute measure of features and properties in the distorted medium which have to be related to perceived quality. An overview of NR metrics

TABLE I

OVERVIEW OF NO-REFERENCE QUALITY METRICS

Ref	Features/Artifacts	Domain	Medium	Image size
[28]	Blocking	DCT	JPEG	512×512
[29]	Blocking	Spatial	JPEG	240×480
[30]	Blocking	Frequency	MPEG-2	720×576
[31]	Blocking	Spatial	MPEG-1	352×288 (CIF)
[32]	Blur	Spatial	Image	768×512
[33]	Blur	DCT	MPEG, JPEG	-
[34], [35]	Sharpness	Spatial/DCT	Video	720×576
[36]	Frame-freeze	Temporal	Video	352×288 (CIF)
[37]	Motion vector information	Temporal/Spatial	MPEG-2	525 & 625 line [38]
[39]	Blocking, blur	Spatial	JPEG	Various
[40]	Blocking, blur, noise	Spatial/DFT	MPEG-2	-
[41]	Blocking, blur, jerkiness	-	MPEG-4	-
[42]	Natural scene statistics	DWT	JPEG2000	768×512
[43]	Frame rate, bit rate, f_{SI13}	Temporal/Spatial	H.263	176×144 (QCIF)
[44]	Bit rate, max/min quality levels	Temporal	MPEG-4	QCIF & CIF
[45]	Mean square error	Spatial/DCT	MPEG-2	720×486
[46]	DFT coefficient cross-correlations	Frequency	Image	-

that can be expected to perform favorable within the context of QoS engineering in wired and wireless networks is provided in Table I and will be discussed in the following.

A. Single feature metrics

Due to the difficulty in designing NR quality metrics, many metrics solely measure single spatial features such as blocking and blur. The former is among the most common artifacts in compression standards using discrete cosine transform (DCT), e.g. JPEG and H.263. On the other hand, blur and ringing are major artifacts in compression algorithms which are based on discrete wavelet transform (DWT) such as JPEG2000.

In [28] a method is proposed which for the reason of computational efficiency measures blocking artifacts entirely in the DCT domain. The blocking is modeled as two-dimensional (2-D) step functions and properties of the HVS are included by introducing visibility threshold relating to activity masking.

In [29] subjective experiments have revealed that blocking, blur, and ringing all correlate strongly with perceived quality. Based on this observation a quality measure for JPEG images was developed exclusively based on the blocking artifact. The decision was also motivated by the fact that blocking occurs as horizontal and vertical edges unlike blur and ringing which can have arbitrary shape and due to that would be harder to measure. The blocking model is divided into three steps. A front-end processing models luminance adaptation of the HVS. Then a block boundary estimation is performed based on the Gaussian blurred edge model. Finally, in an integration stage the estimated edge amplitudes are collapsed into a single scalar blocking value.

A blocking measure for video sequences has been proposed in [30] which is said to be insensitive to other artifacts. Here, each frame is partitioned into blocks and further sampled into subimages. These subimages are pairwise correlated within (intra-block) and across block boundaries (inter-block) to obtain similarity measures within and between the blocks, respectively. The correlation measures are performed on the frequency representation of each subimage. The final blocking measure is given by the ratio of intra-block to inter-block similarity. Values close to unity indicate low blocking while values significantly larger than unity yield strong blocking.

A generalized block-edge impairment metric (GBIM) for image and video coding is reported in [31]. It is the successor of the block-edge impairment metric (BIM). With BIM horizontally and vertically differences at 8×8 block boundaries are measured which by GBIM are perceptually weighted according to luminance masking properties of the HVS.

A blur metric is proposed in [32] which does not make any assumptions about the type or origin of the blur. The metric works in the spatial domain where basically an edge image is obtained by using a Sobel edge detector. Then either horizontal or vertical edge widths are measured and identified as local blur measures. An overall blur measure is attained by averaging the local blur values over all edge locations. The quality prediction performance of the metric has been testified with subjective experiments on a set of Gaussian blurred images and JPEG2000 compressed images. The Pearson linear correlation and Spearman rank order correlation show good agreement of the predicted and the subjective quality.

The blur metric in [33] is based on histogram computations of DCT coefficients and can therefore instantly be applied in the compressed domain of JPEG images or MPEG frames. The idea behind this is to take advantage of image analysis which has already been performed in the compression process. In a three step process, first the DCT information of the entire image is gathered, then it is evaluated with respect to contained DCT values that are equal to zero, and finally the measure is normalised to remove dependance on the image size. The prediction performance is validated with subjective experiments on a set of MPEG coded video sequences.

Intuitively one could consider image sharpness as an opposite measure to image blur. A content independent sharpness metric has been proposed in [34]. It is motivated by observations on statistical measures of image frequency distributions. Specifically, the kurtosis, as a measure of peakedness of a

signal distribution relative to the normal distribution, has been identified as a precise measure of image sharpness. The basic steps of the algorithm are composed of the creation of an edge image using a Canny edge detector, an assignment of 8×8 blocks to each edge pixel and transformation into the DCT domain, the calculation of the probability density function (PDF) of each block, and finally the computation of a 2-D kurtosis on the PDF. A good prediction performance has been verified with subjective experiments. The kurtosis method has been adopted in [35] but is said to provide more robustness to noisy images by computation solely in the wavelet domain using a 3-level discrete dyadic wavelet transform (DDWT).

In video sequences, distortions do not only occur in the spatial domain but also in the temporal domain. Common artifacts include jitter, which are abrupt variations resulting from asynchronous acquisition of video frames, and jerkiness, the perception of still images in a video sequence resulting from too low frame rates. The loss of entire frames is called frame-loss whereas a frame that is repeated in consecutive time instants is referred to as frame-freeze.

A quality measure for real-time video streams over Internet, exclusively measuring temporal artifacts, is reported in [36]. Here, temporal discontinuities, or frame-freeze, are object to quality prediction. They are detected when the temporal derivative of the frame luminance is null. A frame-freeze is considered perceptible when its duration exceeds a certain threshold. Furthermore, the model accounts for the regularity and density of the occurring discontinuities and also for their burst sizes. Abrupt scene changes and object displacements after frozen frames are also taken into account. The performance of the metric has been verified in subjective experiments achieving high correlations with perceived quality.

In [37] the assumption is made, that quality degradation in MPEG-2 is correlated to the accuracy of motion vector estimation. Specifically, the authors state that motion estimation is highly related to the mean absolute error (MAE), computed by subtracting each pel in a block with its corresponding motion compensated reference block, and to spatial activity (SA), as the amount of texture in a macro-block. A probability surface is established with the variables MAE and SA allowing for classification of macro-blocks into the categories well predicted, badly predicted, or uncertainly predicted. An additional measure looks into the spatial and temporal neighbourhood of macro-blocks and provides supportive information for a final probability measure of how well a macro-block is predicted. A final criticality index is then established as an average of the probabilities over all macro-blocks.

B. Metrics of combined features and structural information

Perceptual quality prediction based on structural properties of images, respectively video frames, is a common approach and is motivated by the fact that the HVS is highly adapted to the extraction of structural information [25]. Usually, this is achieved by quantifying different features in an image and combining them in a certain way. The weights for feature quantification are often derived from subjective experiments to find better accordance to perceived quality. In comparison to single-feature metrics, such multi-feature metrics offer more insight into the structural information of an image and also more robustness to different types of artifacts. A good example of a multi-feature metric for JPEG images utilising perceptual based weightings is proposed in [39].

In [40] experiments with videos are reported in which subjects had to vote for the annoyance of three different artifacts, blockiness, blurriness and noisiness, resulting in mean annoyance values (MAV) for each sequence. The artifacts were introduced into three different spatial regions (top/middle/bottom) in video frames to prevent the test subjects from learning the artifact locations. Feature metrics have been used to measure the strength of each of the artifacts. Finally, the weighted Minkowski metric, also referred to as LP-norm of p^{th} order, has been used as a combination rule of the artifacts. It has been observed that the simple linear model for $p = 1$ provides as good correlations as higher order models.

The aforementioned metrics all presume that artifacts in images and video frames are perceived equally annoying no matter in which location they appear. The metric designed in [41], however, besides extraction of blocking, blur and jerkiness, also considers higher order aspects of the HVS in terms of semantic segmentation. This is motivated by the fact that there are usually regions in visual content that are of higher interest and others of lower interest. It is then stated that artifacts in regions of interest (ROI) appear more annoying than in the rest of the image. Of course, the ROI is subject and content dependent but generally two important aspects can be pointed out: the focus of attention and object tracking. The former explains the phenomenon that there are certain objects which attract everyone's attention in an image, for example faces. The latter phenomenon emphasizes that motion attracts peoples attention. Based on these two aspects the image is divided into semantic segments of different importance using a-priori knowledge about the objects to be segmented, for instance face colour or motion information. In the pooling process the features measured in the regions with semantically higher importance are then given higher weights.

Considering the metrics discussed so far, blur and blocking, seem to have received strong attention as perceptually important image and video artifacts. A totally different approach has been examined in [42]. Instead of obtaining structural information as a combination of artifacts, a two state natural scene statistics (NSS) model is proposed for quality evaluation of natural scenes. The authors philosophy is that all images, regardless of content, are initially perfect unless distorted during acquisition, processing, or reproduction. Most distortions that are prevalent in image and video processing systems are not natural in terms of NSS. The method is designed for quality assessment of images compressed with a wavelet based encoder such as JPEG2000. Natural scenes contain nonlinear dependencies which are disturbed by the compression process. This disturbance is quantified based on significance analysis of wavelet coefficient magnitudes and related to human quality perception by conduction of subjective experiments.

C. Metrics incorporating codec parameter settings

In the sequel, metrics are discussed that base quality prediction partly on a set of codec specific objective parameters. This is thought to reduce computational complexity by using readily available information provided by the source encoder.

The goal of modelling a low complexity metric for H.263 encoded video sequences is pursued in [43]. The quality evaluation is based on compression settings and content features. A total of nine features is evaluated regarding their suitability for quality prediction. Five of them are recommended by the American National Standards Institute (ANSI) [47]. All measures were performed on five video sequences representing different content classes. Additionally, subjective experiments have been performed to obtain MOS for the different sequences. In order to reduce the dimension of the parameter space, principal component analysis (PCA) has been used to determine the relationship between MOS and the objective parameters. The result is a reduced set of three parameters frame rate, bit rate, and f_{SI13}, a parameter for overall spatial information. The set represents a trade-off between computational complexity and prediction performance.

A method for objectively evaluating perceived quality of service (PQoS) for MPEG-4 coded video content is reported in [44]. The design is based on observations of data from subjective experiments revealing that over a certain threshold bit rates do not impact on perceived quality (PQ) anymore and below a certain threshold PQ drops drastically. The bit rate thresholds have been found to be highly dependent on the dynamics in the video content. The data from the subjective experiments is used to derive an exponential function which is proposed for objective prediction of PQ. This method was verified to work well on common intermediate format (CIF) and quarter CIF (QCIF) sized sequences.

D. Metrics using data hiding techniques

The following metrics make unconventional use of data hiding procedures by means of watermarking. A watermark is an image or pattern invisibly embedded into a host image and has been traditionally used for purposes such as copyright protection. In the following metrics, however, the watermark is used to assess the quality of its host image based on the assumption, that the host undergoes the same distortions as the watermark. This requires that the transmitted watermark is known at the receiver in order to perform the quality evaluation. Therefore, this type of method is also referred to as a pseudo no-reference method [46] since no information about the reference is needed but instead information about the embedded watermark. The choice of the right watermark plays an important role because it has to be sufficiently robust to be detectable after strong distortions but also fragile enough to be degraded proportionally to the host image. The principle system common to the discussed metrics is illustrated in Fig. 2. Here, h_t and w_t denote the host and watermark to be transmitted, respectively. The received versions are denoted by h_r and w_r. Such a scenario allows for incorporating compression and transmission artifacts in the medium.

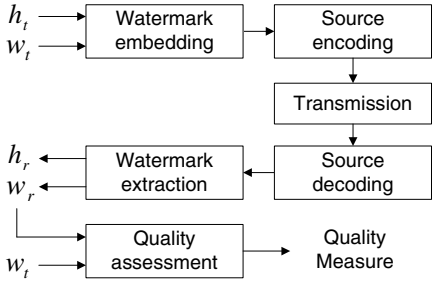

Fig. 2. Quality assessment system utilising watermark based methodology.

In [45] a metric is presented which embeds the watermark in the DCT domain of each frame in a MPEG-2 video sequence. The embedding procedure is summarised as follows. First, a pseudo-noise image $p(n)$ is generated for each frame of the sequence to avoid visual latency. The watermark $w_i(n)$ for a frame $f_i(n)$ at time instant n is then obtained by multiplying $p(n)$ with an image $I(n)$. Finally, the watermark is embedded in the mid-frequencies DCT coefficients of the frame. Embedding in low frequencies would create visible artifacts whereas embedding in high frequencies would cause the watermark to be easily removed. The transmitted sequence is given as $Y_i(n) = DCT\{log[f_i(n)]\} + \alpha \cdot w_i(n)$ where α is a scaling factor varying the strength of the watermark. For the quality assessment the watermark is removed from its host. The quality measure is then calculated as the mean square error (MSE) of the transmitted image $I(n)$ and received image $I_r(n)$. Using this technique enables to use FR methods such as MSE to be used for NR assessment of the host presuming that the embedded image $I(n)$ is available at the receiver.

An empirical approach by means of a psychophysical experiment has been used in [45] to evaluate the visibility of the embedded watermark. To avoid this approach and instead analytically control the watermarks visibility, an embedding method based on a psychovisual model has recently been proposed in [46]. The model provides different frequency and orientation selective subbands. A watermark is then embedded into each subband allowing for the quality metric to have several measuring points on the frequency content. The final quality score Q is attained from averaging of correlation measures in high and middle frequency bands between original and received watermark. A psychometric function is used to translate the objective quality scores into predicted MOS.

IV. Reduced-reference Quality Assessment

The RR approach makes the task of quality evaluation comparably easier to NR techniques by providing information about the reference to the assessment equipment. Therefore, RR methods measure a change in features between the reference and distorted medium which in turn can be used to assess quality degradation. However, this is done at the cost of transmitting the features as side information over the channel which makes the amount of overhead needed for the RR information a crucial aspect of this type of metrics, especially in low-bandwidth wireless channels. In general, RR approaches are based on similar principles to the ones already

TABLE II

OVERVIEW OF REDUCED-REFERENCE QUALITY METRICS

Ref	Features/Artifacts	Domain	Medium	Image size
[48], [49]	Blocking, blur, ringing, masking	Spatial	JPEG	512×512
[50]	Spectral/Temporal content, blocking	Spatial/Temporal	MPEG-2	-
[51]	Motion-related content descriptors	Temporal	MPEG-4	176×144 (QCIF)
[52]	Wavelet-based HVS model	Spatial/Temporal/DWT	H.263	352×240
[53]	Natural image statistics	DWT	JPEG, JPEG2000	768×512
[54]	Temporal and spatial parameters	Temporal/Spatial	Various	525 & 625 line [38]

discussed in Section III but not as many metrics have been proposed yet (see also Table II). Therefore, in this section the metrics are not further classified according to their methods.

In [48] a metric for JPEG coded images is proposed combining five structural features f_i into a hybrid image quality metric (HIQM). In particular, the features are blocking, blur, edge-based image activity, gradient-based image activity, and intensity masking. The overall perceptual quality measure is then computed as a weighted sum of the extracted features

$$HIQM = \sum_{i=1}^{5} w_i \cdot f_i \tag{1}$$

where the weights w_i are derived from subjective experiments and reflect the impact of each of the features on perceptual quality. The quality degradation of a received image as compared to its related reference image can then be obtained as

$$\Delta_{HIQM} = |HIQM_t - HIQM_r| \tag{2}$$

with $HIQM_t$ and $HIQM_r$, respectively, being the $HIQM$ values for the transmitted and received image. The method provides good correlations with perceived quality despite the fact that only a single number needs to be transmitted along with the image. The drawback of this method, however, is a non-uniform range for the different feature measures. This issue has been addressed in [49] by introducing normalised HIQM (NHIQM) which uses an extreme value normalisation [1] of the feature measures in order for them to fall in the interval $[0, 1]$. Similar as in (2) a measure for quality degradation can be obtained. Beside NHIQM the weighted LP-norm has been proposed for quality prediction

$$L_{P,W} = \left[\sum_{i=1}^{5} w_i^P |f_{t,i} - f_{r,i}|^P \right]^{\frac{1}{P}} \tag{3}$$

where $f_{t,i}$ and $f_{r,i}$ are the transmitted and received normalised features, respectively, and P is the order of the norm. The LP-norm provides similar prediction performance as NHIQM with the advantage that the different feature values are available at the receiver as additional information about the structural degradation in the image.

A quality metric for MPEG-2 video streams is proposed in [50] taking into account both chromatic components and the achromatic component of the Krauskopf colour space. A total of four features is extracted on all three components resulting in a set of twelve features for each video frame. In particular, two features related to spectral content and one feature related to temporal content are extracted in addition to the blocking measure in [55]. Data from subjective experiments has been used along with the feature measures to train and test a time delay neural network (TDNN) which is said to preserve the sequential nature of the video stream unlike conventional multilayer perceptrons. Very good correlation of the objectively predicted quality with subjective quality has been shown over a range of different bit rates and video contents.

In [51] the concept of advanced video traces is introduced for MPEG-4 video streams. The key idea is to extend the set of available parameters in conventional video traces, which provide for instance information on frame size (in bits) and frame type (I/P/B), with a set of motion-related content descriptors. These descriptors allow for evaluation on three different temporal granularity levels; frame level, group of pictures (GoP) level (a GoP are the frames between two intra-coded I frames), and shot level. Quality predictors utilising these descriptors are then proposed to quantify the quality degradation due to loss of the different frame types. The performance of the motion based measure has been extensively evaluated with respect to the full-reference metric in [26] which incorporates aspects of different levels of the HVS.

The continuous video quality evaluation (CVQE) metric proposed in [52] is based on a perceptually motivated multi-channel decomposition using the discrete wavelet transform (DWT). A variable amount of coefficients to be transmitted allows for a scalable overhead. A masking model based on the generalised gain control formulation [20] is implemented leading to the channel response

$$r_{k,\Theta}(m,n,t) = \frac{w_k^p (a_{k,\Theta}(m,n,t))^p}{b + w_k^q \sum_{\Theta} (a_{k,\Theta}(m,n,t))^q} \tag{4}$$

where $a_{k,\Theta}(m,n,t)$ are the contrast values, k and Θ, respectively, are scale and orientation of the DWT decomposition, and w_k^p is an excitatory and w_k^q an inhibitory weight. The distortions between reference and distorted frame at time instant t are then computed as the absolute difference between the channel responses

$$d(m,n,t) = \left| r_{k,\Theta}^{ref}(m,n,t) - r_{k,\Theta}^{dist}(m,n,t) \right| \tag{5}$$

To obtain an objective measure of quality degradation, the distortions $d(m,n,t)$ have been converted using a non-symmetric function [10]. Subjective experiments data has been used to evaluate the objective measures of CVQE concluding that the metric performs better the higher the allocated bandwidth (by means of DWT coefficients) for the reduced-reference is.

In [53] a quality metric has been proposed that is based on a natural image statistic model in the wavelet domain. The authors state that many image distortions lead to significant changes in wavelet coefficient histograms. The method utilises a 3-scale and 4-orientation steerable pyramid decomposition [56] for the wavelet decomposition which avoids aliasing between the resulting twelve subbands. The PDF of the wavelet coefficients are analysed in each subband for both reference and received image. The Kullback Leibler distance (KLD) can then be used to quantify the difference of the wavelet coefficients between the two images based on the coefficient histograms. Transmitting the histogram of the reference image, however, might result in a large overhead by means of the reduced-reference. Therefore, the reference histogram is approximated with a two parameter generalised Gaussian density (GGD) model. This means that only two parameters and additionally an approximation error have to be transmitted. At the receiver side the PDF does not get approximated since due to distortions the image might not be natural and therewith not fit the GGD anymore. The final distortion between the received and the transmitted image is then calculated as

$$D = log_2 \left(1 + \frac{1}{D_0} \sum_{k=1}^{K} |\hat{d}^k(p^k \| q^k)| \right) \qquad (6)$$

where the constant D_0 is used as a scaler of the distortion measure, $\hat{d}^k(p^k \| q^k)$ denotes the estimation of the Kullback-Leibler distance between the probability density functions p^k and q^k of the k^{th} subband in the transmitted and received image, and K is the number of subbands. Of the twelve subbands only six are selected (two from each scale) to reduce the overhead for the reduced-reference, which is then composed of 18 different feature measures (162 bits). The metric has been verified to work well on a set of images processed to contain different types of distortions. However, to avoid the problem of sending the features separate from the reference over the channel, the interesting concept of quality aware images has been introduced in [57]. As other methods before, this one makes use of data hiding techniques, but in a different respect. Instead of using the embedded watermark for quality evaluation at the receiver side, the watermark itself contains the quality measure as in (6) and only has to be extracted. Therewith, no overhead is introduced and no ancillary channel for transmission of side information is needed. This solution might also be applicable to other RR metrics previously discussed in this section.

Finally, the General Model of the video quality model (VQM), developed by the National Telecommunications and Information Administration (NTIA), is summarised in [54]. It has been extensively tested by the VQEG [9] and recently been standardised by the American National Standards Institute (ANSI) [47]. The model is said to be general purpose and applicable for various types of coding and transmission systems. The reduced-reference is composed of the VQM and a set of calibration parameters which have to be transmitted over an ancillary channel. The ancillary channel has to provide a total of 14% of the bandwidth of the uncompressed video sequence whereof 9.3% are for the VQM parameters and 4.7% for the calibration parameters. The testing of VQM has been organised by VQEG and performed in three independent laboratories on 525 and 625 line [38] video test material. The General Model of VQM has been found to be comparably better than any other metric in the test. However, this has been achieved at the cost of a high transmission overhead.

V. Summary and Conclusions

Given the growing interest in delivery of multimedia services over wired and wireless networks along with the advent of highly efficient image and video codecs, there is a strong need for metrics being able to measure and quantify transmission and coding quality as perceived by the end-user. A survey and classification of such image and video quality metrics and favorable quality assessment methodologies was presented in this paper.

It has been shown that most of the methodologies proposed over the years are designed to perform best on a certain medium and image size. This approach of application specific quality evaluation is sensible since a "general purpose" metric for various media types might be too complex due to need for scalability. However, more research should be concentrated on quality evaluation of recent image and video codecs, such as H.264. Furthermore, by the number of proposed metrics it can be observed that research in RR quality assessment seems to lack behind the NR quality assessment. The authors opinion, however, is that RR quality assessment methods should receive stronger attention as they are a good compromise between FR and NR methods. Their main advantage over NR methods is the capability of providing a measure of quality degradation instead of an absolute quality measure. On the other hand, the overhead of the reduced-reference becomes a critical issue in metric design. Finally, a fairly unexplored research field is the assessment of combined effects of video quality and audio quality [58]. In order to measure the overall quality of video one needs to define a mutual measure of audiovisual quality.

References

[1] J.-R. Ohm, *Multimedia Communication Technology*. Springer, 2004.

[2] S. Winkler, *Digital Video Quality - Vision Models and Metrics*. John Wiley & Sons, 2005.

[3] D. Soldani, M. Li and R. Cuny (Ed.), *QoS and QoE Management in UMTS Cellular Systems*. John Wiley & Sons, 2006.

[4] F. Pereira, "Sensations, perceptions and emotions towards quality of experience evaluation for consumer electronics video adaptations," in *Proc. of Int. Workshop on Video Processing and Quality Metrics for Consumer Electronics*, Jan. 2005.

[5] ITU, "Method for objective measurements of perceived audio quality," ITU-R, Rec. BS.1387-1, Dec. 2001.

[6] ——, "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow band telephone networks and speech codecs," ITU-T, Rec. P.862, Feb. 2001.

[7] Video Quality Experts Group. (2006) VQEG. [Online]. Available: http://www.vqeg.org/

[8] ——, "Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment," VQEG, Mar. 2000.

[9] ——, "Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment, phase II," VQEG, Aug. 2003.

[10] ITU, "Methodology for the subjective assessment of the quality of television pictures," ITU-R, Rec. BT.500-11, 2002.

[11] ——, "Subjective video quality assessment methods for multimedia applications," ITU-T, Rec. P.910, Sept. 1999.

[12] B. A. Wandell, *Foundations of Vision*. Sinauer Associates, Inc., 1995.

[13] H. R. Wu and K. R. Rao (Ed.), *Digital Video Image Quality and Perceptual Coding*. CRC Press, 2006.

[14] J. Mannos and D. Sakrison, "The effects of a visual fidelity criterion on the encoding of images," *IEEE Trans. on Inf. Theory*, vol. 20, no. 4, pp. 525–536, July 1974.

[15] F. Lukas and Z. Budrikis, "Picture quality prediction based on a visual model," *IEEE Trans. on Commun.*, vol. 30, no. 7, pp. 1679–1692, July 1982.

[16] S. Daly, "Visible differences predictor: an algorithm for the assessment of image fidelity," *Proc. of SPIE Human Vision, Visual Processing, and Digital Display III*, vol. 1666, pp. 2–15, Aug. 1992.

[17] J. Lubin, "A visual discrimination model for imaging system design and evaluation," *in Vision Models for Target Detection and Recognition, World Scientific, E. Peli (Ed.)*, pp. 245–283, 1995.

[18] J. B. Martens, "Multidimensional modeling of image quality," *Proc. of the IEEE*, vol. 90, no. 1, pp. 133–153, Jan. 2002.

[19] P. C. Teo and D. J. Heeger, "Perceptual image distortion," in *Proc. of IEEE Int. Conf. on Image Processing*, vol. 2, Nov. 1994, pp. 982–986.

[20] A. B. Watson and J. A. Salomon, "Model of visual contrast gain control and pattern masking," *Journal of the Optical Society of America*, vol. 14, no. 9, pp. 2379–2391, Sept. 1997.

[21] A. M. Eskicioglu and P. S. Fisher, "Image quality measures and their performance," *IEEE Trans. on Commun.*, vol. 43, no. 12, pp. 2959–2965, Dec. 1995.

[22] H. R. Sheikh, A. C. Bovik, and G. de Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Trans. on Image Processing*, vol. 14, no. 12, pp. 2117–2128, Dec. 2005.

[23] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. on Image Processing*, vol. 15, no. 2, pp. 430–444, Feb. 2006.

[24] C. J. van den Branden Lambrecht and O. Verscheure, "Perceptual quality measure using a spatio-temporal model of the human visual system," in *Proc. of SPIE Digital Video Compression: Algorithms and Technologies*, vol. 2668, Jan. 1996, pp. 450–460.

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

[26] A. B. Watson, J. Hu, and J. F. McGowan, "Digital video quality metric based on human vision," *SPIE Journal of Electronic Imaging*, vol. 10, no. 1, pp. 20–29, Jan. 2001.

[27] T. Yamashita, M.Kameda, and M.Miyahara, "An objective picture quality scale for video images (PQS video) - definition of distortion factors," in *Proc. of SPIE Visual Communications and Image Processing*, vol. 4067, May 2000, pp. 801–809.

[28] S. Liu and A. C. Bovik, "Efficient DCT-domain blind measurement and reduction of blocking artifacts," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 12, no. 12, pp. 1139–1149, Dec. 2002.

[29] L. Meesters and J. B. Martens, "A single ended blockiness measure for JPEG coded images," *Signal Processing*, vol. 82, pp. 369–387, Mar. 2002.

[30] T. Vlachos, "Detection of blocking artifacts in compressed video," *IEE Electronics Letters*, vol. 36, no. 13, pp. 1106–1108, June 2000.

[31] H. R. Wu and M. Yuen, "A generalized block-edge impairment metric for video coding," *IEEE Signal Processing Letters*, vol. 4, no. 11, pp. 317–320, Nov. 1997.

[32] P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi, "A no-reference perceptual blur metric," in *Proc. of IEEE Int. Conf. on Image Processing*, vol. 3, Sept. 2002, pp. 57–60.

[33] X. Marichal, W. Y. Ma, and H. J. Zhang, "Blur determination in the compressed domain using DCT information," in *Proc. of IEEE Int. Conf. on Image Processing*, vol. 2, Sept. 1999, pp. 386–390.

[34] J. Caviedes and S. Gurbuz, "No-reference sharpness metric based on local edge kurtosis," in *Proc. of IEEE Int. Conf. on Image Processing*, vol. 3, Sept. 2002, pp. 53–56.

[35] R. Ferzli, L. J. Karam, and J. Caviedes, "A robust image sharpness metric based on kurtosis measurement of wavelet coefficients," in *Proc. of Int. Workshop on Video Processing and Quality Metrics for Consumer Electronics*, Jan. 2005.

[36] R. R. Pastrana-Vidal and J. C. Gicquel, "Automatic quality assessment of video fluidity impairments using a no-reference metric," in *Proc. of Int. Workshop on Video Processing and Quality Metrics for Consumer Electronics*, Jan. 2006.

[37] M. Caramma, R. Lancini, and M. Marconi, "Subjective quality evaluation of video sequences by using motion information," in *Proc. of IEEE Int. Conf. on Image Processing*, vol. 2, Oct. 1999, pp. 313–316.

[38] ITU, "Encoding parameters of digital television for studios," ITU-R, Rec. BT.601, 1994.

[39] Z. Wang, H. R. Sheikh, and A. C. Bovik, "No-reference perceptual quality assessment of JPEG compressed images," in *Proc. of IEEE Int. Conf. on Image Processing*, vol. 1, Sept. 2002, pp. 477–480.

[40] M. C. Q. Farias and S. K. Mitra, "No-reference video quality metric based on artifact measurements," in *Proc. of IEEE Int. Conf. on Image Processing*, vol. 3, Sept. 2005, pp. 141–144.

[41] A. Cavallaro and S. Winkler, "Segmentation-driven perceptual quality metrics," in *Proc. of IEEE Int. Conf. on Image Processing*, vol. 5, Oct. 2004, pp. 3543–3546.

[42] H. R. Sheikh, A. C. Bovik, and L. Cormack, "No-reference quality assessment using natural scene statistics: JPEG2000," *IEEE Trans. on Image Processing*, vol. 14, no. 11, pp. 1918–1927, Nov. 2005.

[43] M. Ries, O. Nemethova, and M. Rupp, "Reference-free video quality metric for mobile streaming applications," in *Proc. of 8th Int. Symp. on DSP and Communication Systems*, Dec. 2005, pp. 98–103.

[44] H. Koumaras, A. Kourtis, and D. Martakos, "Evaluation of video quality based on objectively estimated metric," *J. of Commun. and Networks*, vol. 7, no. 3, pp. 235–242, Sept. 2005.

[45] M. C. Q. Farias, S. K. Mitra, and M. Carli, "Video quality objective metric using data hiding," in *Proc. of IEEE Workshop on Multimedia Signal Processing*, Dec. 2002, pp. 464–467.

[46] A. Ninassi, P. L. Callet, and F. Autrusseau, "Pseudo no reference image quality metric using perceptual data hiding," in *Proc. of SPIE Human Vision and Electronic Imaging XI*, vol. 6057, Feb. 2006.

[47] ANSI, "American national standard for telecommunications: Digital transport of one-way video signals - Parameters for objective performance assessment," ANSI T1.801.03, 2003.

[48] T. M. Kusuma and H.-J. Zepernick, "A reduced-reference perceptual quality metric for in-service image quality assessment," in *IEEE Symposium on Trends in Communications*, Oct. 2003, pp. 71–74.

[49] U. Engelke and H. J. Zepernick, "Quality evaluation in wireless imaging using feature-based objective metrics," in *Proc. of IEEE Int. Symp. on Wireless Pervasive Computing*, Feb. 2007.

[50] P. Le Callet, C. Viard-Gaudin, and D. Barba, "Continuous quality assessment of MPEG2 video with reduced reference," in *Proc. of Int. Workshop on Video Processing and Quality Metrics for Consumer Electronics*, Jan. 2005.

[51] O. A. Lotfallah, M. Reisslein, and S. Panchanathan, "A framework for advanced video traces: Evaluating visual quality for video transmission over lossy networks," *EURASIP J. on Applied Signal Processing*, vol. 2006, Article ID 42083, 21 pages, 2006.

[52] S. S. Hemami and M. A. Masry, "A scalable video quality metric and applications," in *Proc. of Int. Workshop on Video Processing and Quality Metrics for Consumer Electronics*, Jan. 2005.

[53] Z. Wang and E. P. Simoncelli, "Reduced-reference image quality assessment using a wavelet-domain natural image statistic model," in *Proc. of SPIE Human Vision and Electronic Imaging*, vol. 5666, Mar. 2005, pp. 149–159.

[54] M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Trans. on Broadcasting*, vol. 50, no. 3, pp. 312–322, Sept. 2004.

[55] Z. Wang, A. C. Bovik, and B. L. Evans, "Blind measurement of blocking artifacts in images," in *Proc. of IEEE Int. Conf. on Image Processing*, vol. 3, Sept. 2000, pp. 981–984.

[56] E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Heeger, "Shiftable multi-scale transforms," *IEEE Trans. on Inf. Theory*, vol. 38, no. 2, pp. 587–607, Mar. 1992.

[57] Z. Wang, G. Wu, H. R. Sheikh, E. P. Simoncelli, E. H. Yang, and A. C. Bovik, "Quality aware images," *IEEE Trans. on Image Processing*, vol. 15, no. 6, pp. 1680–1689, June 2006.

[58] S. Winkler and C. Faller, "Perceived audiovisual quality of low-bitrate multimedia content," *IEEE Trans. on Multimedia*, vol. 8, no. 5, pp. 973–980, Oct. 2006.