Signal Processing: *Image Communication* 25 (2010) 535–546

Contents lists available at ScienceDirect

Signal Processing: *Image Communication*

journal homepage: www.elsevier.com/locate/image

Modelling of spatio–temporal interaction for video quality assessment

Quan Huynh-Thu [a,*], Mohammed Ghanbari [b]

[a] *Psytechnics Ltd, Fraser House, 23 Museum Street, Ipswich IP1 1HN, UK*
[b] *University of Essex, Wivenhoe Park, Colchester CO4 3SQ, UK*

ARTICLE INFO

Article history:
Received 30 April 2009
Accepted 25 March 2010

Keywords:
Video
Subjective quality
Objective model
Spatial quality
Temporal quality
Spatio–temporal interaction

ABSTRACT

Video services have appeared in the recent years due to advances in video coding and convergence to IP networks. As these emerging services mature, the ability to deliver adequate quality to end-users becomes increasingly important. However, the transmission of digital video over error-prone and bandwidth-limited networks may produce spatial and temporal visual distortions in the decoded video. Both types of impairments affect the perceived video quality. In this paper, we examine the impact of spatio–temporal artefacts in video and especially how both types of errors interact to affect the overall perceived video quality. We show that the impact of the spatial quality on overall video quality is dependent on the temporal quality and vice-versa. We observe that the introduction of a degradation in one modality affects the quality perception in the other modality, and this change is larger for high-quality conditions than for low-quality conditions. The contribution of the spatial quality to the overall quality is found to be greater than the contribution of the temporal quality. Our results also indicate that low-motion talking-head content can be more negatively affected by temporal frame freezing artefacts than other general type of content with higher motion. Based on the results of a subjective experiment, we propose an objective model to predict overall video quality by integrating the contributions of a spatial quality and a temporal quality. The non-linear model shows a very high linear correlation with subjective data.

1. Introduction

There is a growing number of video quality metrics proposed in the literature. Since a video can be thought as a motion sequence of still images, one typical approach is to extend models designed for still images [1–3] to video. A single overall quality score for a video is typically obtained by applying an image quality metric individually on each frame of the video and by temporally pooling the individual frame quality values. The simplest form of temporal pooling is an average calculation that equally weights all frames. Alternatively, additional information about motion or motion models can be used to weight differently the quality values calculated on each frame [4–6]. More recently, models taking into account the spatio–temporal characteristics of a video to compute an overall video quality prediction have been recommended by the International Telecommunication Union [7].

The most important coding distortions are the blockiness and blurriness. Blockiness (or blocking distortion) is a spatial degradation characterised by the appearance of an underlying block structure in the image. This block structure is a common feature to all DCT-based video compression techniques. It is caused by coarse quantisation of the spatial frequency components during the

[*] Corresponding author.
E-mail addresses: qht@ieee.org (Q. Huynh-Thu), ghan@essex.ac.uk (M. Ghanbari).

0923-5965/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.image.2010.03.006

encoding process. In practice, blockiness appears as high data compression ratios are used in order to transmit video contents (especially those with high level of motion) on low bandwidth. Blurriness (or blurring distortion) is a spatial degradation that can also be caused by the encoding. It is characterised by a reduced sharpness of edges and loss of spatial detail. Compression algorithms often cause blurriness by trying to trade-off bits to code resolution and motion. Blurring is also a common feature of wavelet-based video compression techniques. In practice, blurriness appears as an attenuation of high spatial frequencies in the image. A variation of blurring, termed ringing, is caused by the quantisation of high frequency coefficients in transform coding and is characterised by ripples around sharp edges. Examples of video quality assessment models designed to address the impact of coding distortions in video sequences can be found in [8–14].

Digital video transmitted over networks of limited bandwidth (e.g. mobile video broadcast) is also often compressed in the temporal domain using frame rate decimation. Furthermore, besides video coding, another very important source of video impairments comes from the transmission of the video stream over an error-prone channel. Transmission errors (e.g. packet loss) can have a serious impact on video quality. Packets can be lost or they can be delayed to the point where they are not received in time for decoding. Both result in packets being discarded by the decoder at the receiving end. A single discard or loss of packet can corrupt a macroblock or motion vectors. Because of the differential prediction of frequency coefficients and the motion estimation/compensation techniques used in video coding, this corruption of information can then spread spatially within a slice of a frame and cascade temporally to adjacent frames up until the next intra-coded frame. The visual impact of such losses varies between video decoders depending on their ability to deal with corrupted streams.

Some decoders hardly recover from certain distributions of errors (e.g. bursts of packet loss). Others will try to apply more or less complex error concealment mechanisms in order to recover missing or corrupted frame information and display a partially damaged frame. The severity of the visual distortions depend on the amount and distribution of lost information, the video content and the method for error concealment. However, more recent video decoders adopt the strategy of entirely discarding a frame that has corrupted or missing information and repeat the previous video frame instead, until the next valid decoded frame is available. In this scenario, additional spatial degradations are not introduced by the transmission channel but full frame repetition and frame drop occur. This creates irregular events of frame freezing and skipping in the video stream that can be visually annoying.

For a given source video content and transmission bandwidth, a critical issue remains to choose the adequate encoding parameters to achieve the best trade-off between picture quality and motion fluidity in order to optimise the overall perceived video quality. When encoding is performed at a constant bit rate, the quantiser step (QS) size will vary accordingly depending on the frame rate for the given bit rate. This will affect the ratio of bits per picture used by the encoder. For a higher frame rate, the encoder must lower the ratio of bits per picture by increasing the QS size and this can cause stronger spatial distortions (e.g. blocking). Conversely, video compression algorithms can allocate a higher ratio of bits per picture at lower frame rates. However, this comes at the expense of a degradation of the motion fluidity, i.e. jerky motion, in the coded video.

Masry and Hemami examined the relationship between frame rate and perceived quality when videos are encoded at fixed bit rates and different frame rates [15]. They found that the perceived quality was higher at lower frame rates for some high-motion action sequences, but that quality was lower at lower frame rates in the case of some low-motion sequences (i.e. videoconferencing).

Speranza et al. investigated the effects on subjective quality of a reduction of temporal resolution either alone or combined with moderate levels of quantisation [16]. They observed that the reduction in frame rate produced a significant loss of quality and that quality decreased with increasing level of quantisation. However, the differential effect of quantisation on subjective quality was more pronounced at higher frame rates than at lower frame rates. Their results indicated that frame rate resolution produced a significant loss of quality that increased with increasing level of quantisation.

Brun et al. also examined perceived video quality as a function of frame rate and quantiser step size [17]. They conducted a subjective quality assessment experiment using QCIF video sequences encoded with H.264. Video encoding was applied using a fixed frame rate and fixed quantiser step. For each of the source sequences, the frame rate and quantiser step were jointly adjusted to achieve a target bit rate between 8 and 64 kbps. The authors reported that, for a target bit rate, a content-dependent optimal frame rate exists from the point of view of perceived quality.

Watson examined the influence of the spatial configuration of a stimulus on the temporal sensitivity and discussed the separability of the spatial and temporal dimensions in the modelling of visual sensitivity [18]. Different aspects of the spatial stimulus (size, surround, edges and spatial frequency) were shown to influence the temporal sensitivity.

Subjective quality of a video degraded by spatial distortions has been extensively studied and different quality metrics have been proposed. On the other hand, the characterisation of temporal errors and the understanding of their impact on perceived quality is a less well-understood area. Although the studies referenced above are limited to temporal degradations introduced by frame rate decimation, they suggest that the overall perceived video quality may be the result of an interaction between the contributions from a spatial quality and a temporal quality axes. However, the interaction between perceived spatial and temporal qualities has not been much documented. In order to investigate further the interaction between spatial and temporal qualities, we conducted a quality assessment experiment. We introduced a temporal frame freezing impairment both in non-coded videos to study the effect on temporal quality alone and in coded videos to examine the interaction between spatial and temporal errors. Based on the experimental results, we propose an objective model to predict overall video quality by combining the contributions of a spatial quality and a temporal quality.

The remainder of the paper is organized as follows. Section 2 describes our subjective quality assessment experiment. Section 3 provides an analysis of the experimental results. Section 4 presents an objective video quality assessment model integrating the interaction between spatial and temporal qualities. Finally, our conclusions are provided in Section 5.

2. Quality assessment experiment

2.1. Test environment

The subjective quality assessment experiment was conducted in a soundproof test room with controlled lighting. The viewing conditions conformed to international recommendations [19]. Background room illumination was set to 20 Lux using lighting providing daylight colour temperature. A calibrated 17-inch computer LCD monitor was used to display the video sequences. The monitor was selected to have a low response time (<15 ms) and its colour temperature was set to 6500 K. The display had a native resolution of 1280×1024 pixels and a dot pitch of 0.264 mm.

2.2. Methodology

We conducted our experiment using the single-stimulus Absolute Category Rating (ACR) method with a 5-point discrete quality scale [19]. We selected the ACR method based on the results of previous studies, which have shown that this method is reliable for assessing the quality of low-resolution videos encoded at low bit rates and exhibiting both spatial and temporal distortions [20–23]. The ACR method with a 5-point discrete scale was used by the Video Quality Experts Group (VQEG) to conduct their evaluation of objective quality assessment models for low-resolution videos [24].

With the ACR method, each test video sequence is presented one at a time and rated individually. After each video presentation, participants are asked to judge its overall quality.

Voting period was not time-limited. Participants had to select one of the five on-screen labelled buttons indicating their opinion of the quality. The buttons were labelled with Excellent, Good, Fair, Poor and Bad, respectively. Then participants had to confirm their choice by clicking on an “OK” button. After each vote, a neutral grey background was displayed on the screen during one second before the next sequence was presented.

Participants provided their quality ratings electronically using the computer mouse. The presentation order of the test videos was randomised between participants such that each of them viewed the test sequences in a different presentation order.

The viewing distance was set to be 8*H* from the screen (where *H* is the physical height of the picture). The chosen viewing distance follows the guidelines set-up by VQEG [24] as well as ITU-T Recommendation P.910 [19]. Brotherton et al. examined the effect of viewing distance in subjective quality assessment of low-resolution video [25]. In their study, two laboratories conducted two experiments with QCIF videos and one with CIF videos. For each subjective test, one experimental set-up enforced a fixed viewing distance using a chin rest and another set-up did not strictly fix the viewing distance. The data from one of the two laboratories showed smaller standard deviations of subjective ratings when stabilised viewing distance was employed using a chin rest. However, the subjective results for test conditions with and without a chin rest were statistically equivalent. In the scenario where a constant viewing distance was not strictly fixed by the chin rest, they found that subjects showed a tendency to find their own preferred viewing distance and that this distance remained relatively stable during the test. They concluded that stabilising the viewing distance by use of a chin rest had no significant effect on subjective quality ratings in their experiments. Based on the conclusions from Brotherton et al., we left the participants to adjust to their most comfortable viewing distance in the sense that a constant viewing distance was not strictly enforced using a device such as a chin rest. However, participants were instructed to keep their back in contact with the chair, which was initially placed at a distance of 8*H* from the screen, so to avoid important variation of the viewing distance during the experiment.

After receiving the instructions in written form and before starting the experiment, each participant had to go through a series of practice trials. Video clips used in the practice trials were representative of the range of quality and types of degradations used in the actual test but used different video scenes than those included in the test. Ratings given during practice trials were not used in data analysis.

Viewers were not provided with information about the nature of the degradations occurring in the test sequences they had to assess. They did not know in advance if a test sequence would contain only spatial, only temporal or a mix of spatio–temporal degradations.

2.3. Viewers

A total of 16 people participated in the experiment, 4 males and 12 females. Participants were recruited from the public. They did not have expertise in video processing or quality assessment. None of them had previously participated in a subjective quality assessment experiment. All subjects assessed all processed video sequences in the test.

2.4. Source material

Since quality judgement can be affected by video content, a variety of video scenes was included in the test material. Seven source videos were used in the test. They had a duration of 8 seconds, did not contain any audio track and were in progressive format. The full frame rate of the reference videos was 25 fps.

The image resolution of the test material was QCIF (176 × 144 pixels) corresponding to a picture dimension of 4.6 cm × 3.8 cm on the screen. QCIF resolution was selected as it is currently widely used in mobile applications such as mobile broadcasting.

Reference video material at QCIF resolution was generated from original uncompressed standard-definition or high-definition video material by spatially down-sampling the original video to QCIF. Down-sampling was performed using a Lanczos filter. Before down-sampling, cropping was applied where necessary to keep the original content aspect ratio when going from high-definition or standard-definition format to QCIF.

Table 1
Description of source video material.

Name	Category	Description
Drink	Advert	Waitress bringing a drink to a customer; camera forward-travelling
Mountain	Documentary	Aerial views of a mountain; very slow camera panning
Ski	Home video	Skier going downhill; slight camera shaking
Music	Music video	Women singing and dancing; focus mostly on faces; plain background; contains 3 scene cuts
Boxing	Movie/advert	Boxer punching training bag; some camera panning
Amfoot	Sports	American football sequence; very fast motion; highly textured areas; some camera movement
Talk	Talking-head	Close-up view of a man speaking

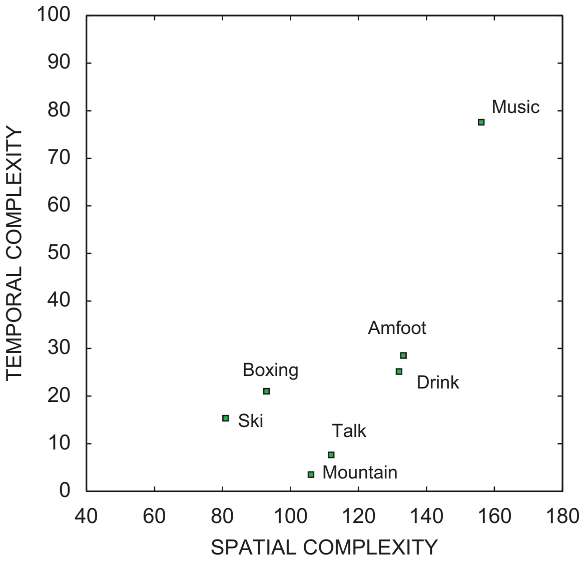

Fig. 1. Spatial and temporal complexity information of the source material.

Table 2
Description of error conditions.

HRC #	Bit rate	Freezing duration (ms)
1	N/A	0
2	N/A	t1 = 160
3	N/A	t2 = 400
4	N/A	t3 = 800
5	N/A	t4 = 1200
6	N/A	t5 = 2000
7	N/A	t6 = 4000
8	BR1	0
9	BR1	t1 = 160
10	BR1	t2 = 400
11	BR1	t3 = 800
12	BR1	t4 = 1200
13	BR1	t5 = 2000
14	BR1	t6 = 4000
15	BR2	0
16	BR2	t1 = 160
17	BR2	t2 = 400
18	BR2	t3 = 800
19	BR2	t4 = 1200
20	BR2	t5 = 2000
21	BR2	t6 = 4000
22	BR3	0
23	BR3	t1 = 160
24	BR3	t2 = 400
25	BR3	t3 = 800
26	BR3	t4 = 1200
27	BR3	t5 = 2000
28	BR3	t6 = 4000

A brief description of the content is provided in Table 1. Video content covered a variety of spatial (SI) and temporal (TI) complexities. Fig. 1 shows the SI and TI values calculated according to ITU-T Recommendation P.910 [19].

2.5. Error conditions

Each source video (SRC) was processed through 28 error conditions (HRC) to generate the processed video sequences (PVS) used in the experiment. A hidden reference condition was also included in the test design.

A description of each error condition is provided in Table 2. HRCs 2–7 included only a temporal impairment, i.e. videos were not encoded. HRCs 8–28 included both spatial and temporal impairments. Spatial artefacts were generated using H.264 [26] and encoding at constant bit rate. An open source implementation of the codec was used.[1] The Baseline profile was used with the default encoder's settings.

Three levels of coding degradations were considered: low, medium and high. The bit rate was carefully selected for each content such that the encoded video would not exhibit any dropped or frozen frames. For each level, a different bit rate was selected for each content to produce a similar level of spatial degradations between the different videos in that level. Bit rates are provided in Table 3. Since the different video scenes had a wide range of content complexity, different bit rate values were used

[1] Available at http://www.videolan.org/developers/x264.html.

Table 3
Coding bit rates for HRCs8–28.

Content	BR1	BR2	BR3
Drink	256	128	56
Mountain	32	24	16
Ski	128	56	24
Music	180	96	56
Boxing	180	96	56
Amfoot	320	180	96
Talk	32	24	16

for the different videos. Indeed, if identical bit rates had been selected for all of the video contents, quality would have saturated towards the top of the quality scale at high bit rates for scenes with lowest complexity, whilst it would have saturated towards the bottom of the quality scale at low bit rates for scenes with highest complexity. Although the severity of the spatial degradations in each level may not be absolutely identical across the video sequences, the bit rates were selected to create enough obvious differences in distortion severity to constitute three levels.

The temporal frame freezing in the video was simulated separately from the encoding. The temporal frame freezing impairment was generated such that the SRC and PVS temporally realigned after the end of the freezing. The temporal freezing was simulated by discarding n frames in the reference video from the frame located at the temporal index p and repeating n times the frame corresponding to the temporal index $p-1$. The p value was not identical between the source videos but was selected so that the frame freezing would occur roughly in the middle of the video sequence. The freezing duration varied from 160 to 4000 ms.

3. Results

Fig. 2 shows that the subjective test design produced a balanced distribution of subjective votes over the entire rating scale, with an overall average Mean Opinion Score (MOS) of 2.91.

Fig. 3 shows the variation of the average quality as a function of the freezing duration. Average quality is computed by averaging MOS of all PVSs through the same HRC. Each of the four curves indicates the quality variation for a different level of coding distortions. We can make the following observations:

- The top curve corresponds to quality in the absence of any spatial distortion. Consequently, MOS reflects the perceived temporal quality. Average quality decreases as the duration of the impairment increases following an inverse-logarithmic or logistic regression. Interestingly, the impact of the temporal impairment on perceived quality is not extensive. Condition MOS drops from 4.68 for no temporal impairment (hidden reference condition) to 3.22 for a temporal frame freezing duration of 4000 ms (i.e. equal to 50% of the length of the video). For that impairment duration, condition MOS still remains in the fair quality category.
- For a given level of spatial impairment, overall quality decreases as the duration of the impairment increases. Data suggest that MOS decreases with increasing freeze duration following an inverse-logarithmic or logistic regression for all three levels of spatial impairment considered. The rate of the quality decrease is faster for lower impairment durations and tends to slow down (regression curve becomes flatter) as the freezing duration increases. This is the case for all levels of spatial impairment.
- Quality decreases at a slower rate when there are stronger spatial impairments already present in the videos. At the smallest bit rates (bottom curve), the overall quality deteriorates only slightly as the freezing duration increases. On the other hand, at the highest bit rates (second curve from the top), the overall quality varies over a much wider range. The variation of overall quality between no freezing and maximum

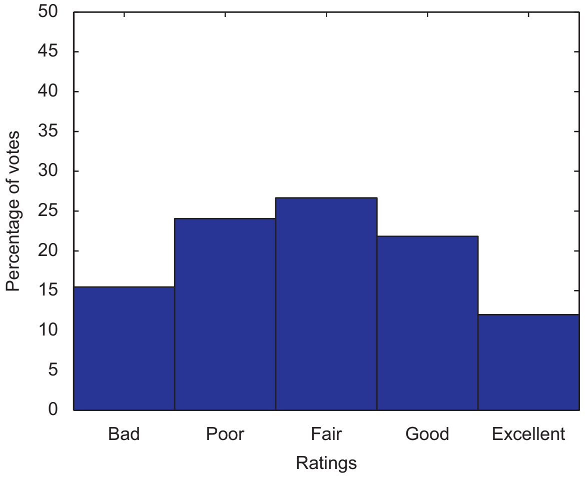

Fig. 2. Distribution of subjective quality ratings.

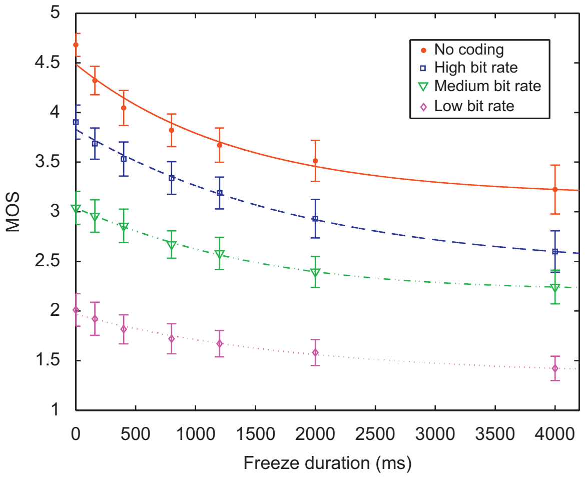

Fig. 3. Average quality variation with freeze duration for different levels of spatial quality. Each curve shows a logistic fit.

freezing duration is 1.46, 1.30, 0.80 and 0.59, for the cases of highest spatial quality (top curve) to lowest spatial quality (bottom curve), respectively.

- For a given freeze duration, average quality is always higher for a higher level of spatial quality.

Fig. 4 shows the influence of content on the quality variation. When we examine the temporal quality shown in plot (a), unsurprisingly we observe a certain content dependency in the MOS value for a specific freeze duration. More interestingly, for a freezing of 4000 ms, the ski sequence stands out with a significantly worse quality than most of other videos, which seem to form a tighter group. The ski sequence is characterised by a very regular and almost periodic motion as it shows a person skiing downhill in a slalom or zig-zag trajectory. This might explain the much higher level of annoyance when motion disruption occurs. As the motion trajectory is very periodic, viewers tend to anticipate where the object of interest will be located in the next frames. So when a freezing occurs, motion is clearly disrupted and highly noticeable or annoying.

The general trend for all videos is that quality decreases as the freeze duration increases. This is also the case when spatial distortions are also present as shown in plots (b)–(d). Some data points seem to indicate an occasional inversion, i.e. MOS becomes higher for higher freeze duration, but these points indicate in reality statistically equivalent MOS values with overlapping confidence intervals (not plotted for graph clarity). The difference of behaviour (i.e. the way that MOS decreases as the freeze duration increases) between video sequences across the freezing durations is highest for cases of the source (no coding) and high bit rate encoding. At this coding level, the sequences exhibit no or very small spatial distortions. The dominant

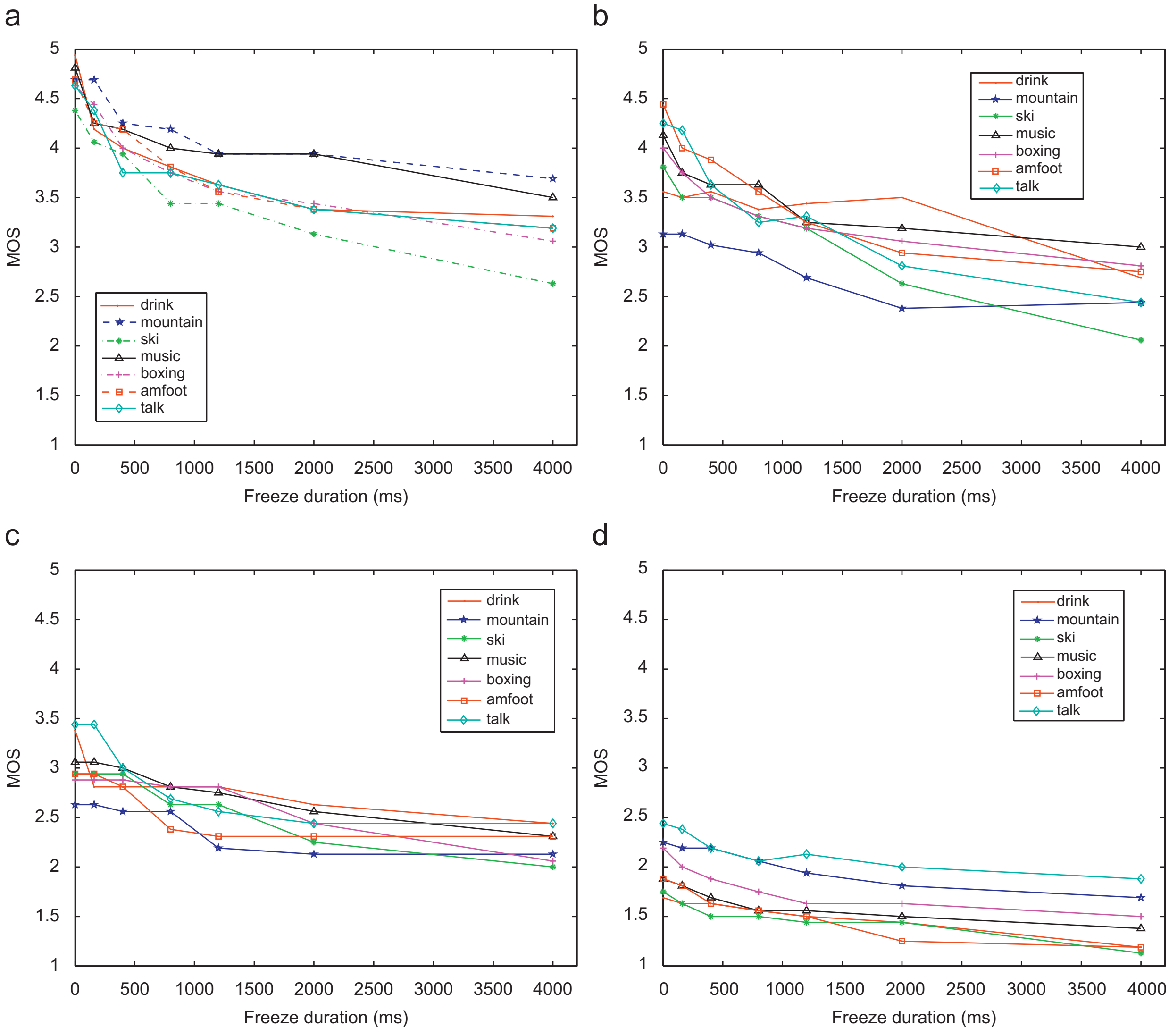

Fig. 4. Quality variation per content for different levels of spatial quality: (a) no coding, (b) high bit rates, (c) medium bit rates and (d) low bit rates.

distortion is therefore expected to be the temporal frame freezing, especially as the duration of the freezing increases. However, the effect of the temporal freezing is highly dependent on the spatio–temporal characteristics of the video sequence so differences in the way that MOS decreases with increasing freezing duration can be observed between the sequences. As explained in Section 2.5, different bit rates were selected for the video sequences in each level of spatial quality. Bit rates for the high spatial quality were selected to produce high quality at $t=0$ (no freezing). Bit rates were selected subjectively but it can be seen in Fig. 4(b) that the sequence Mountain actually received a much lower MOS for $t=0$. Consequently, as the freeze duration increases, the slope of MOS decrease is flatter than for the other sequences as this coded sequence was already exhibiting annoying spatial artefacts. Therefore, the addition of the freezing impairment did not have an important effect on overall quality as the dominant type of distortion may have remained the spatial distortion. Likewise, at medium and low bit rates, the difference of behaviour across the video sequences tends to disappear across the freeze durations as the severity of the spatial distortions has increased. In other words, as the bit rate decreases, the spatial distortions may become the more dominant type of artefact. Consequently, if temporal freezing occurs, the overall quality is not much more affected, especially at low bit rates. Finally, the fact that the absolute MOS values of the coded sequences at $t=0$ may be closer to each other in a given coding level may simply be due to subjective selection of the bit rates.

Fig. 5 shows the overall quality against each of the quality axes. Each data point represents the average (spatial or temporal or overall) quality across the PVSs that have been processed through the same HRC. The figure illustrates how the introduction of a degradation in one modality (spatial or temporal domain) affects the quality perception in the other modality, and that this change is much larger for high-quality conditions than for low-quality conditions.

Fig. 5(a) shows four groups of points, where each group indicates the overall quality for a given spatial quality condition. The top group of points in the graph corresponds to degradations introduced only in the temporal domain since videos are not encoded. The overall quality is therefore equal to the temporal quality. From top to bottom, the other three groups correspond to a variation of the overall quality as a function of the temporal quality for a constant spatial quality of 3.90, 3.04 and 2.01, respectively. For the highest temporal quality condition (no freeze), overall quality dropped from 4.68 to 2.01 (decrease of 2.67) between the highest and lowest spatial quality condition. For the lowest temporal quality condition (freeze duration of 4000 ms), overall quality dropped from 3.22 to 1.42 (decrease of 1.80) between the highest and lowest spatial quality condition. These results show that a temporal impairment affects the overall quality differently depending on the spatial quality. Overall quality is less affected by a temporal frame freezing degradation as the spatial quality decreases (the curve becomes flatter as spatial quality decreases).

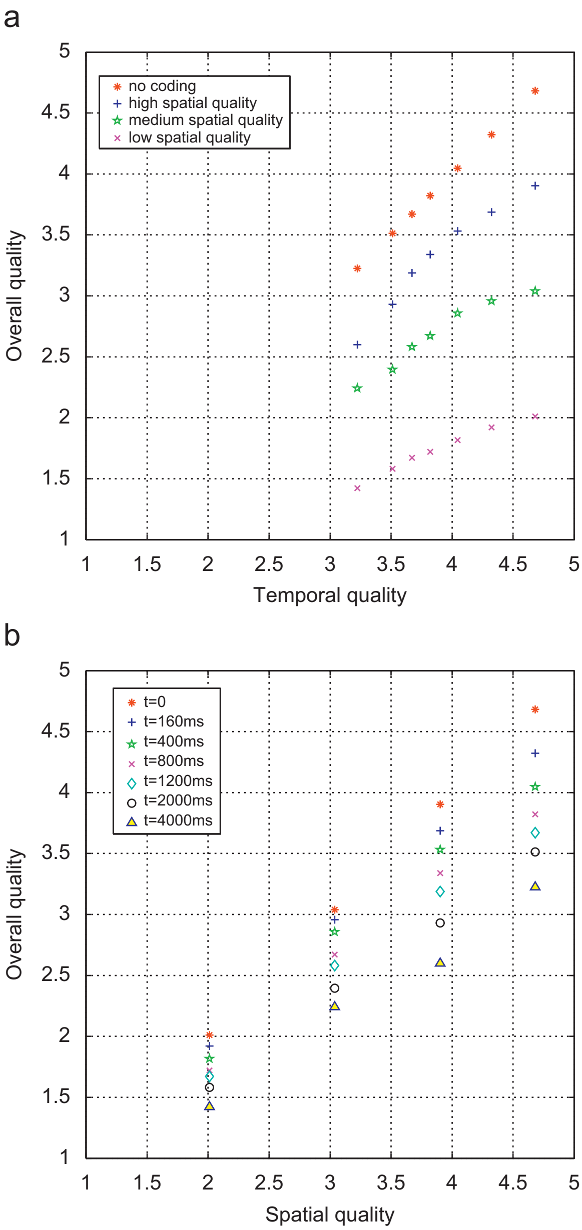

Fig. 5. Overall quality as a function of the (a) temporal quality and (b) spatial quality.

Fig. 5(b) shows seven groups of points, with each group indicating the overall quality for a given temporal quality condition. The top group of points in the graph corresponds to degradations introduced only in the spatial domain since no freezing is present. The overall quality is therefore equal to the spatial quality. From top to bottom, the other six groups of points correspond to a variation of the overall quality as a function of the spatial quality for a constant temporal quality of 4.32, 4.05, 3.82, 3.67, 3.51 and 3.22, respectively. For the highest spatial quality condition (no coding), overall quality dropped from 4.68 to 3.22 (decrease of 1.46) between the highest and the lowest temporal quality condition. For the lowest spatial quality (coding with lowest bit rates), overall quality dropped from 2.01 to 1.42 (decrease of 0.59) between the highest and the lowest temporal quality condition. These results show that spatial artefacts affect overall quality differently depending on the temporal quality of the video. Overall quality is less affected by the spatial artefacts as the temporal quality decreases.

We observe that the decrease of overall quality due to a decline of temporal quality is smaller than in the case of a decline in spatial quality. For the lowest level of spatial quality, overall quality declines very slowly as temporal quality decreases (flat curve). On the other hand, for the lowest temporal quality (corresponding to a freezing of 50% of the video), overall quality declines greatly as the spatial quality decreases. This suggests that spatial quality contributed more than temporal quality to the overall quality in our test. Finally, we see that the overall quality varies monotonically with both spatial quality and temporal quality.

Table 4 shows the Pearson correlation coefficient (R) between overall quality (MOS) and spatial quality (SQ), temporal quality (TQ) and product of the two (SQxTQ). SQ (TQ) is the subjective quality of the video in absence of temporal (spatial) degradations. Correlation values are calculated separately per video content and for all videos. The results show that SQ has a much greater correlation with MOS than TQ with MOS. This is the case for all seven video contents. For the test sequences used in the experiment, we therefore find that spatial errors have a much greater impact on the overall quality than the temporal frame freezing distortion. This finding differs from the results of a previous research study [27] in which it was found that a fluidity impairment, mainly in the form of several small bursts of dropped frames, had a stronger effect on the overall quality. This difference between our results and the ones reported in [27] suggests that the importance of the contribution of either modality (spatial or temporal) may be influenced by the subjective test design. In other words, the type and severity of the spatio–temporal distortions considered in the data will influence the integration function that combines the spatial quality and temporal quality into a single overall quality.

An analogy can be made with the field of audio–visual quality assessment, where audio–visual quality is modelled by combining the separate video and audio qualities. Indeed, different past studies on audio–visual quality prediction models have proposed different integration functions, depending on the scope of the study (e.g. types of video and audio degradations) or the audio–video material considered (e.g. talking-head or more general content) [28,29]. However, a major difference between studies addressing audio–visual quality and the present study on video quality concerns the methodology to collect the quality scores for the different modalities. Audio–visual quality prediction models are built by collecting subjective quality scores in three separate (audio, video and audio–video) experiments or in one audio–video experiment asking three ratings (audio quality, video quality and overall quality ratings). However, in the present study, the quality scores for the different modality were obtained in one experiment and using only one question asking overall quality but by combining the spatial–temporal errors such that some stimuli only exhibited distortions in one of the two modalities, whilst others exhibited distortions in both modalities. Nonetheless, clear similarities exist with the field of audio–visual quality modelling.

We note that the talking-head content stands out from all other videos with the lowest correlation (0.81) between SQ and MOS. Also, that content shows a much higher correlation (0.54) between TQ and MOS compared to the other videos. In other words, the correlation between temporal quality and overall quality was higher for this very low-motion content than for all other videos with higher motion. Data therefore indicate that viewers were more responsive to a temporal freezing in this type of content, although it is characterised by very low motion. The fact that the talking-head video contains mainly natural motion such as human head or lip movement to which viewers are familiar in real life may explain that a sudden temporal impairment of such content creates a stronger negative impact on quality judgement. The ski sequence had the second highest correlation between TQ and MOS. Again, this was the case although other videos are characterised by higher motion magnitude or complexity. This confirms earlier observations made in Fig. 4 that the nature of the motion in the video rather than the magnitude of the motion is an influential factor in the perceptual impact of a frame freezing impairment. Our results seem to defeat the common belief that content with higher motion magnitude should be more negatively affected by a temporal distortion. Similar observations were made in a past study [30] addressing the impact of jerkiness on video quality.

The best variables to be used in an objective video quality assessment model are those that correlate as linearly as possible with the target (i.e. the subjective quality). Furthermore, it is desirable to use independent variables in the sense that one cannot be represented by a function of the other ones. Otherwise, that variable is likely to represent redundant information in the model. The Pearson correlation coefficient between SQ and TQ is close to zero for each content and across all videos. This suggests that the two perceptual quality axes contribute independently to the perceived quality in the sense that they can be independent variables in an objective model predicting overall quality.

Table 4
Correlation between spatial (SQ), temporal (TQ) quality and overall quality (MOS).

Content	R(MOS,SQ)	R(MOS,TQ)	R(MOS,SQxTQ)	R(SQ,TQ)
Drink	0.91	0.28	0.96	< 0.0001
Mountain	0.95	0.29	0.99	< 0.0001
Ski	0.87	0.44	0.99	< 0.0001
Music	0.95	0.28	0.99	< 0.0001
Boxing	0.89	0.39	0.98	< 0.0001
Amfoot	0.90	0.39	0.99	< 0.0001
Talk	0.81	0.54	0.98	< 0.0001
All	0.90	0.36	0.98	0.0065

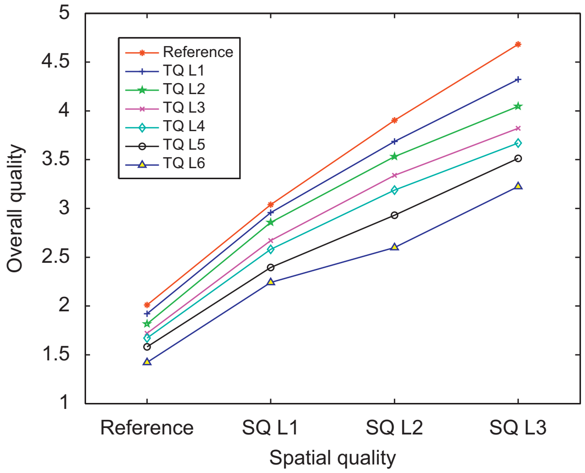

Fig. 6. Factorial graph indicating a linear fan response.

Theory of cognition has established algebraic rules to model how humans integrate information. The three basic rules of cognitive algebra are addition, multiplication and averaging rules [31]. If two variables are integrated by some algebraic rule, the pattern of the response plotted in a factorial graph can reveal the form of that rule. Addition rules correspond to a pattern of parallelism; multiplicative rules correspond to a linear fan pattern; averaging corresponds to a parallelism with cross-over pattern. For two stimulus variables, the factorial design is constructed by a Row × Column matrix. Each row of the matrix corresponds to one level of the row variable. In our case, there are four levels of spatial quality (Reference+three coding levels indicated by SQ L1-L3). For each row or level of spatial quality, there are seven columns or levels of temporal quality (Reference+six freezing conditions indicated by TQ L1-L6). Each cell of the matrix represents one experimental observation defined by the corresponding levels of row and column. From this design, the factorial graph is obtained as follows. The vertical axis represents the response (overall quality). The column stimuli are placed at equal intervals on the horizontal axis. Each row of data points are plotted above the corresponding column stimuli and are connected to form a curve. For clarity, data were re-plotted in a factorial graph as shown in Fig. 6. The graph shows that the interaction between the quality modalities approaches a linear fan response, indicating a multiplication integration rule. The linear fan can be seen as the data points for the reference column stimuli are tighter than the data points for the SQ L3 column stimuli. The multiplication rule is also supported by the very high correlation between the product of the two qualities (SQxTQ) and MOS as shown in Table 4.

We performed an analysis of variance (ANOVA) on the data. Table 5 reports the results of the two-way ANOVA using content (*Content*) and freeze duration (*FDuration*) as independent variables. MOS was the dependent variable. Data for the ANOVA exclude files with coding degradations. Results therefore refer to the influence of the variables on the temporal video quality only. There are main effects of content and freeze duration, as well as interaction effect on the temporal quality.

Table 6 reports the results of the three-way ANOVA using content (*Content*), coding level (*CLevel*) and freeze duration (*FDuration*) as independent variables. MOS was the dependent variable. Results therefore refer to the influence of the variables on the overall video quality. Since different bit rates were used depending on the content, only a coding level (corresponding to BR1, BR2 and BR3 in Table 2) can be used in the ANOVA. There are main effects of all variables, two-way interaction effects between all variables and three-way interaction effect on overall video quality.

4. Model of spatio–temporal quality interaction

The most general functional form of a model of first order that represents the overall video quality (VQ) as a function of the spatial quality (SQ) and temporal quality (TQ) is:

$$VQ = \alpha * SQ + \beta * TQ + \gamma * SQ * TQ + K \tag{1}$$

where K is a constant offset. This model takes into account the separate contribution from each variable (expressed by the two additive terms) as well as the interaction between the two variables (expressed by the multiplicative term).

The drawback of such a functional form (with α, β, and $\gamma \neq 0$) is the fact that if one of the two modalities is characterised by a very low quality whilst the other modality has a very high quality, then overall quality prediction may result in value that remains high because both parameters represent a quality value and are added together in Eq. (1). However, experimental data have indicated that overall subjective quality can degrade very quickly with the decrease of either of the two modalities even if the other modality is characterised by a high quality. Furthermore, correlation values presented in Table 4 have indicated that the strongest contribution to the overall quality was from the interaction between the two modalities in the sense that this multiplicative term had the highest linear correlation with MOS.

Based on the findings from the experimental data presented in Section 3, another functional form is therefore proposed to model the overall video quality from the combination of separate spatial and temporal qualities:

$$VQ = 1 + \left(\frac{TQ - 1}{MOS_{max} - 1}\right)^{\alpha} (SQ - 1)^{\beta} \tag{2}$$

Table 5
Two-way ANOVA using data excluding coding degradations.

Source	Sum Sq.	df	Mean Sq.	*F*-value	*p* -value
Content	26.6939	6	4.4489	6.0332	< 0.0001
FDuration	146.3367	6	24.3894	19.4727	< 0.0001
Content * FDuration	30.0383	36	0.8344	1.7519	0.005098
Error	257.1862	540	0.4763		

Table 6
Three-way ANOVA using all data.

Source	Sum Sq.	df	Mean Sq.	*F*-value	*p*-value
Content	32.7723	6	5.4620	4.3850	0.000624
CLevel	2008.1515	3	669.3838	112.6726	<0.0001
FDuration	290.1875	6	48.3646	21.6738	<0.0001
Content $*$ CLevel	96.8374	18	5.3798	7.9847	<0.0001
Content $*$ FDuration	34.0402	36	0.9456	1.7371	0.00573
CLevel $*$ FDuration	55.6543	18	3.0919	5.9831	<0.0001
Content $*$ CLevel $*$ FDuration	70.1537	108	0.6496	1.4828	0.001306
Error	709.6626	1620	0.4381		

where:

- The offsets are all set to 1 to ensure that quality of 1 in any modality (i.e. severe degradations in that modality) produces an overall quality of 1.
- MOS_{max} is the highest possible quality (reference condition) and is used to map the quality prediction to the 1–5 range.
- α and β are weighting powers applied to take into account the difference in contributions of the temporal and spatial qualities.

Eq. (2) can be viewed as a simplification of Eq. (1) where only the multiplicative term is retained and individual contributions of SQ and TQ are set to 0; the constant K is set to 1 so that minimum quality is 1; γ is set to 1; exponents α and β are added on SQ and TQ respectively to allow the different contributions of spatial and temporal qualities to the overall quality as indicated by the experimental data and discussion in Section 3.

The optimum values of the different parameters in the model were obtained using the Nelder–Mead simplex search method [32]. A subset of the data was used for model training and the rest of the data was used as unknown data for model validation. One file was randomly selected for each condition to give a set of 28 files for model validation. The training data set included the remaining 168 of the 196 files. The optimum values of α and β were determined to be respectively 0.89 and 0.98. A value $\alpha<\beta$ reflects the previous observation that temporal quality (TQ) contributed less to overall quality than spatial quality (TQ). We investigated the effect of the training data set by repeating the random selection of training files several times and by re-calculating the optimum values of α and β for each different training set. Even if the training data changed, the resulting optimum values of α and β only marginally changed. Furthermore, these small changes in values of the model's parameters did not actually affect the correlation with subjective data.

Fig. 7 shows the resulting scatter plot of subjective data against model prediction. The correlation per file was 0.96 on both the training and validation data. The correlation per condition was 0.99 on the training data and 0.96 on the validation data. Note that in this case each test condition is actually represented by one file for the validation data.

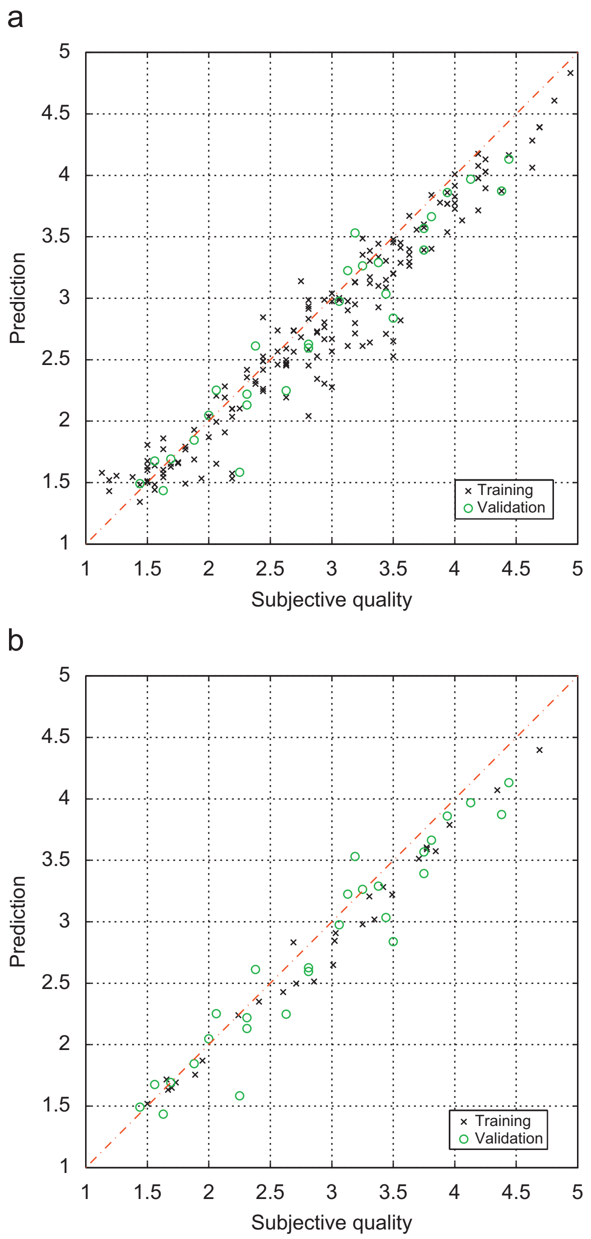

Fig. 7. Scatter plot of overall subjective video quality against quality prediction from model: (a) per file and (b) per condition. A subset of the data was used as unknown data for model validation and the rest was used for model training.

Although the correlation value obtained for the model of Eq. (2) may not seem like a major improvement compared to the one reported in Table 4 when using only the multiplicative term SQxTQ, the introduction of the parameters α and β makes the model more generally applicable. Indeed, the observations made in Section 3 indicate different contributions from the spatial and temporal qualities. A more accurate prediction model can therefore be obtained by tuning the parameters to specific content or specific types of distortions or specific applications, whereas a simple multiplicative model (without exponent parameters) cannot be optimised. It was also discussed in Section 4 that the exact form of the integration function, i.e. the values of α and β, may be dependent on the dominant distortions considered in the study.

Note that the results shown in Fig. 7 were obtained by using the subjective values of SQ and TQ in Eq. (2). In other words, the subjective values of SQ and TQ were used in Eq. (2) to obtain the prediction values of overall video quality (VQ). In order to produce a complete objective model, the values of SQ and TQ would therefore also need to be computed (predicted) using objective metrics or computing algorithms. Many existing objective models measuring the impact of spatial distortions have been proposed in the literature. Examples can be found in [5,33] and can be used to model SQ. An objective temporal quality metric such as the one in [34] can be used to model TQ.

5. Conclusions

A subjective experiment was conducted to study the interaction between the spatial and temporal errors occurring in a video stream. Spatial artefacts were generated by encoding videos at different bit rates. Temporal artefacts were generated by introducing a single event of frame freezing of different durations. The temporal impairment was introduced both in the non-coded videos to examine the variation of the temporal quality alone and in the coded videos to examine the interaction between spatial and temporal qualities.

Data have indicated that the variation of the temporal quality followed an inverse-logarithmic or logistic regression with the freezing duration. Although an expected content dependency was found in the variation of temporal quality for a given temporal impairment severity, our results indicate that low-motion talking-head content can be more negatively affected by temporal frame freezing artefacts than other general type of content with higher motion.

The overall video quality was affected by both spatial and temporal qualities. It was found that the introduction of a degradation in one modality affected the quality perception in the other modality, and this change was larger for high-quality conditions than for low-quality conditions. It was found that the contribution of the spatial quality to the overall quality was greater than the contribution of the temporal quality. However, this last observation may be due to the range of error conditions used in the test design.

Overall video quality can be modelled by the interaction between two perceptual axes: spatial and temporal quality axes. Variation of spatial quality is influenced by the temporal quality and vice-versa. However, our results show that this dependency can be modelled by an interaction function such that the quality in each modality can be studied and modelled separately, whilst their dependence can be modelled by the interaction function.

A non-linear model integrating the separate contributions of the spatial and temporal qualities was proposed to predict overall video quality. The model showed a very high linear correlation with subjective data. This model requires the ability to represent each of the quality modalities by a single parameter.

Because of the limited amount of available data, we divided the files of the same experiment into a training and validation set to assess the performance of the model. Validation files were selected randomly and we ensured that the random selection of the validation set did not influence significantly the performance of the model. Although, the validation data contained test sequences degraded by different error conditions than those used in the training data, they were derived from the same original content. Further work would be necessary to validate the proposed model using a more extensive dataset and in particular using validation data including different video contents than those used in the training data.

References

[1] T.N. Pappas, R.J. Safranek, Perceptual criteria for image quality evaluation, in: A. Bovik (Ed.), Handbook of Image and Video Processing, Academic Press, 2000, pp. 669–684.

[2] Z. Wang, H.R. Sheikh, A.C. Bovik, Objective Video Quality Assessment, The Handbook of Video Databases: Design and Applications, CRC Press, 2003, pp. 1041–1078, (Chapter 41).

[3] H.R. Sheikh, M.F. Sabir, A.C. Bovik, A statistical evaluation of recent full reference image quality assessment algorithms, IEEE Transactions on Image Processing 15 (11) (2006) 3441–3452.

[4] Z. Wang, L. Lu, A.C. Bovik, Video quality assessment based on structural distortion measurement, Signal Processing: Image Communication 19 (2) (2004) 121–132.

[5] S. Winkler, Digital Video Quality: Vision Models and Metrics, John Wiley and Sons, 2005.

[6] K. Seshadrinathan, A.C. Bovik, A structural similarity metric for video based on motion models, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 1, Honolulu, 2007, pp. 869–872.

[7] International Telecommunication Union, Objective perceptual multimedia video quality measurement in the presence of a full reference, ITU-T Rec. J.247 (August 2008).

[8] Z. Yu, H.R. Wu, S. Winkler, T. Chen, Vision-model-based impairment metric to evaluate blocking artifacts in digital video, Proceedings of the IEEE 90 (1) (2002) 154–169.

[9] M.H. Pinson, S. Wolf, A new standardized method for objectively measuring video quality, IEEE Transactions on Broadcasting 50 (3) (2004) 312–322.

[10] F. Yang, S. Wan, Y. Chang, H.R. Wu, A novel objective no-reference metric for digital video quality assessment, IEEE Signal Processing Letters 12 (10) (2005) 685–688.

[11] G. Ginesu, F. Massidda, D.D. Giusto, A multi-factors approach for image quality assessment based on a human visual system model, Signal Processing: Image Communication 21 (4) (2006) 316–333.

[12] E.P. Ong, X. Yang, W. Lin, Z. Lu, S. Yao, X. Lin, S. Rahardja, B.C. Seng, Perceptual quality and objective quality measurements of

compressed videos, Journal of Visual Communication and Image Representation 17 (4) (2006) 717–737.

[13] I.P. Gunawan, M. Ghanbari, Reduced-reference video quality assessment using discriminative local harmonic strength with motion consideration, IEEE Transactions on Circuits and Systems for Video Technology 18 (1) (2008) 71–83.

[14] I.P. Gunawan, M. Ghanbari, Efficient reduced-reference video quality meter, IEEE Transactions on Broadcasting 54 (3) (2008) 669–679.

[15] M. Masry, S.S. Hemami, A.-M. Rohaly, W. Osberger, Subjective quality evaluation of low bit rate video, in: Proceedings of SPIE Human Vision and Electronic Imaging, San Jose, 2001, pp. 102–113.

[16] F. Speranza, A. Vincent, D. Wang, A. Mainguy, P. Blanchfield, R. Renaud, Rate control for improved picture quality in low-bit rate video coding, in: Proceedings of SPIE Visual Communications and Image Processing, vol. 4671, San Jose, 2002, pp. 722–733.

[17] P. Brun, G. Hauske, T. Stockhammer, Subjective assessment of H.264-AVC video for low-bitrate multimedia messaging services, in: Proceedings of IEEE International Conference on Image Processing (ICIP), vol. 2, Singapore, 2004, pp. 1145–1148.

[18] A.B. Watson, Temporal sensitivity, in: K. Boff, L. Kaufman, J. Thomas (Eds.), Handbook of Perception and Human Performance, Wiley, New York, 1986 (Chapter 6)..

[19] International Telecommunication Union, Subjective video quality assessment methods for multimedia applications, ITU-T Rec. P.910 (April 2008).

[20] Q. Huynh-Thu, M. Ghanbari, A comparison of subjective video quality assessment methods for low-bit rate and low-resolution video, in: Proceedings of the IASTED International Conference on Signal and Image Processing, vol. 479, 2005, pp. 70–76.

[21] Q. Huynh-Thu, M. Ghanbari, D.S. Hands, M. Brotherton, Subjective video quality evaluation for multimedia applications, in: Proceedings of SPIE Human Vision and Electronic Imaging XI, vol. 6057, San Jose, 2006, pp. 464–474.

[22] M.D. Brotherton, Q. Huynh-Thu, D.S. Hands, K. Brunnström, Subjective multimedia quality assessment, IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences (Special Section on Image Media Quality) 89 (11) (2006) 2920–2932.

[23] C. Lee, H. Choi, E. Lee, S. Lee, J. Choe, Comparison of various subjective video quality assessment methods, in: Proceedings of SPIE Conference on Image Quality and System Performance III, vol. 6059, San Jose, 2006, pp. 60590601–07.

[24] Video Quality Experts Group, VQEG multimedia test plan (September 2007).

[25] M.D. Brotherton, D.S. Hands, K. Brunnström, J. Jonsson, O.A. Soysuren, Stabilising viewing distances in subjective assessment of mobile video, in: Proceedings of SPIE Human Vision and Electronic Imaging XI, vol. 6057, San Jose, 2006, pp. 268–275.

[26] International Telecommunication Union, Advanced video coding for generic audiovisual services, ITU-T Rec. H.264 (November 2007).

[27] R.R. Pastrana-Vidal, J.-C. Gicquel, J.L. Blin, H. Cherifi, Predicting subjective video quality from separated spatial and temporal assessment, in: Proceedings of SPIE Human Vision and Electronic Imaging XI, vol. 6057, San Jose, 2006, pp. 276–286.

[28] J.G. Beerends, F.E. de Caluwe, The influence of video quality on perceived audio quality and vice versa, Journal of Audio Engineering Society 47 (5) (1999) 355–362.

[29] D.S. Hands, A basic multimedia quality model, IEEE Transactions on Multimedia 6 (6) (2004) 806–816.

[30] Q. Huynh-Thu, M. Ghanbari, Temporal aspect of perceived video quality in mobile video broadcasting, IEEE Transactions on Broadcasting 54 (3) (2008) 641–651.

[31] N.H. Anderson, A Functional Theory of Cognition, Lawrence Erlbaum Associates Inc., 1996.

[32] J.C. Lagarias, J.A. Reeds, M.H. Wright, P.E. Wright, Convergence properties of the Nelder–Mead simplex method in low dimensions, SIAM Journal of Optimization 9 (1) (1998) 112–147.

[33] I.P. Gunawan, Reduced-reference impairment metrics for digitally compressed video, Ph.D. Thesis, University of Essex, June 2006.

[34] Q. Huynh-Thu, M. Ghanbari, No-reference temporal quality metric for video impaired by frame freezing artefacts, in: Proceedings of IEEE International Conference on Image Processing (ICIP), Cairo, 2009.