Eindhoven University of Technology

MASTER

Influence of level of motion on perceived video quality in MPEG-2 coding

van den Berge, B.P.

Award date:
2009

Link to publication

Influence of
level of motion
on perceived video quality
in MPEG-2 coding

Author   Berend P. van den Berge, 455734
                          Faculty of Industrial Engineering & Innovation Sciences
                          Human Technology Interaction
                          Eindhoven University of Technology

University Supervisors   dr. L.M.J. Meesters
                          Faculty of Industrial Engineering & Innovation Sciences
                          Eindhoven University of Technology

                          prof. dr. D.G. Bouwhuis
                          Faculty of Industrial Engineering & Innovation Sciences
                          Eindhoven University of Technology

Company Supervisor   drs. N. Van den Ende
                          Connected Consumer Solutions
                          Philips Research Eindhoven

The World is a book, and those who do not travel read only a page.

~St. Augustine

Acknowledgements

"Potius sero quam numquam"

~Titus Livius

Summary

The number of digital videos transmitted over wireless networks is rapidly increasing. The wireless networks and thus the video signal are vulnerable to interference with external influences such as microwaves and other transmitting devices. As a result of this interference video quality may become impaired. This negative influence on the video quality can be limited with MPEG-2 based adaptive video coding methods such as spatial and temporal scaling.

The perceived video quality is hypothesized to be better when video is coded with a temporal scaling technique than coded with a spatial scaling technique in videos with a low level of motion and the reverse in videos with higher levels of motion. These hypotheses are investigated in the current research by finding an answer to *what the overall effect is of temporal and spatial scaling for different levels of motion in video sequences on perceived video quality*. Results of this research add to obtain a higher end quality for video sent via wireless networks.

A 3 (*Level of motion*) x 3 (*Coding method*) design was used. *Level of motion* is represented by three different video sequences in Standard Definition quality and *Coding method* is represented by a reference version and two coded versions: a temporally scaled version (I-Frame Delay) and a spatially scaled version (transcoding). The ITU (BT.500-11) direct comparison method was used to assess the video pairs.

No significant difference in quality for the reference versions was found. However, the participants preferred temporal scaling over spatial scaling for a low level of motion and the reverse for higher levels of motion. These results help to create content-based decision making mechanisms, able to select the better scaling method (temporal or spatial) to obtain high video quality.

Contents

1. Introduction

In the last few years one notices a huge increase in the availability of digital video for e.g. computer use, home electronics and mobile devices such as PDA's and smart phones. These devices are increasingly used to view video material transmitted via wired and wireless networks. Especially wireless networks are vulnerable to external influences such as microwaves and transmitting devices that might interfere. This interference potentially leads to a decreased video quality when transmitted over wireless networks resulting in fewer details that can be distinguished. Wireless and wired networks are increasingly used in private households.

Households nowadays are well equipped with infrastructures for analog video but these infrastructures are not (standard) suited for digital video. A popular approach is installing a wireless in-home network for data communication between computers, home electronics and mobile devices (Tanenbaum, 2005). The installation of a wireless network is not invasive and does not ask for a major house renovation; it involves little more than installing a wireless router (a little box) to an internet connection and plugging its adapter into a power outlet. Most computers and an increasing number of consumer electronics are equipped with a Wi-Fi antenna. Computers that do not have an integrated Wi-Fi antenna can be connected to a wireless network via a small Wireless Network USB adapter.

The main disadvantage of wireless networks is reliability. Wi-Fi networks are active in the 2.4 GHz band (frequency range of 2.400 to 2.500 GHz) which can be used without a license (ISM Band, 2008). This unlicensed 2.4 GHz frequency is used by several products such as wireless phones, car alarms, microwave ovens, and wireless CCTV cameras which might all interfere with the signals sent over the Wi-Fi network (2.4 GHz frequency, 2008). Microwave ovens are best known for this kind of interference. As soon as it is switched on, it disturbs the signal between devices that are trying to communicate with each other via the wireless network. Although this interference has little influence on the communication between devices for most applications, it does influence video streaming negatively, resulting in bad video quality (Darken et al., 2001). This negative influence can be reduced with adaptive coding methods resulting in a higher video quality than without the use of coding methods in video streaming.

Video streaming is transmitting video by a server that is viewed on a client. No data is stored in between; the signal is shown directly on the television. In data networks (e.g. internet, home network), data is divided in little packets which are sent separately. Packets have the freedom to travel along different routes (and thus routers) depending on the availability of routes (networks switched on and off) and on how busy certain parts in the network are. In addition, a wireless network might loose packets on the way due to interference (e.g. microwave switched on) or packets might be delayed due to heavy network traffic called congestion. This is the unreliability that is mentioned before. For most communication between devices this is no problem, since packets can be sent again to re-deliver any missing parts. However, it impairs a video signal immediately.

Digital video risks being impaired when sent via such data networks because it undergoes the same division into data packets. The difference between video and most other applications or services which make use of wireless data networks is that video is time-based. Packets must arrive on time in order not to impair the video sequence at the client. Packets that are delayed or lost impair the video sequence and therefore lower the perceived quality of the video sequence. To prevent this quality loss several solutions have been found (e.g. Tripathi & Claypool, 2001). The current research focuses on two scaling methods that can help improve the quality of the perceived video in case of any disturbances in the network. Scaling methods (in the current research also named coding methods) are ways to lower the amount of data needed to describe a video. This can be done in different ways. It is important to choose the best scaling method for a (part of a) particular video sequence to keep the amount of distortions in the video sequence as low as possible and in addition, the perceived video quality as high as possible. The scaling methods used in the present research are based on MPEG-2 coded video developed by the Moving Picture Experts Group (MPEG); one method is based on temporal scaling and the other method on spatial scaling.

MPEG-2 coding is a standard for video distribution in home networks (Jarnikov, 2007, p13). It is originally developed to reduce, transport and store digital and analog video (both the video and audio signal). To store an analog video sequence digitally or a

digital video in raw format, a high storage capacity is required (See Section 2.1). By reducing the amount of data needed, not only the necessary amount of storage capacity, but also the network load is reduced. This data reduction can be either perceptually *lossless* or perceptually *lossy*. If a *lossless* coding method is used, this means that the reconstructed signal (data after the process of encoding and decoding) exactly matches the source data. Techniques that are not *lossless* are called *lossy*; reconstructed data encoded with a *lossy* coding technique is not the same as the source data. The above mentioned techniques can be combined with adaptive coding methods that are able to lower the network load even more while video is being sent via a network.

Philips has developed two adaptive coding methods that can adapt video sent via a network on the fly and that are able to limit the loss of video quality when the video quality is (temporary) decreased. One algorithm is based on temporal scaling and the other on spatial scaling. The current study focuses on the use of spatial and temporal coding of an MPEG-2 coded video distributed via a network, depending on the amount of motion of the content. When a different scaling method is used based on the content of the video this is called content-aware scaling.

Content-aware scaling can increase the quality up to 50 % as argued by Tripathi and Claypool (2001). According to other previous studies, motion is expected to have an influence on perceived video quality when the video sequence is scaled using temporal or spatial coding (Tripathi & Claypool, 2001; De Hesselle, 2006; Wijermans, 2007). How this is influenced depends on multiple factors and is still being researched. De Hesselle (2006) found that the length of impairments influences the perceived video quality. Shorter impairments are less perceivable. For this research the total video sequence will therefore be impaired to simulate the worst case situation.

Chapter 2 describes the basics of MPEG-2 coding. Chapter 3 explains perceived video quality, the functioning of the human visual system and the relation between the two. Chapter 3.5 describes how the level of motion in video sequences was estimated. Chapter 4 describes the empirical study that is performed and Chapter 5 its results. The current thesis ends with Chapter 6, to discuss the results and give the conclusions of the present research.

2. Moving Picture Experts Group

The focus of the present investigation is on adaptive coding methods in MPEG-2. Hence, it is important to comprehend the operation of MPEG-2 to understand the coding methods. MPEG-2 coding is used to reduce the amount of information needed to transport video via data networks or to store video sequences. MPEG-2 is one of the methods developed by the Moving Pictures Expert Group (MPEG) responsible for the MPEG standards. The group was founded in 1988 to connect the academic world and commercial companies, and to manage the development of a standard for audio and video. This chapter is devoted to MPEG-2 and its operation. Section 2.1 explains why coding is needed, Section 2.2 explains the operation of MPEG coding. Section 2.3 covers scalability. Section 2.4 explains the basic working of wireless networks to obtain a better understanding why coding methods are needed. Finally, Section 2.5 shows two adaptive coding methods in wireless networks, one temporal oriented and one spatial oriented method.

2.1. Video coding is necessary

A video sequence consists of still pictures sequentially shown, giving the illusion that the pictures in the sequence show a continuous motion. It is important that the sequential pictures are following each other rapidly enough to make sure that the apparent continuous motion does not flicker. This is only possible if the sequence is shown at a rate of at least fifteen pictures per second (Video, 2008; Masri & Memami, 2001). Typically 25 pictures per second are shown on European television which results in a huge amount of data if it will not be compressed. MPEG-2 can reduce video data with a factor 5 to circa 30, depending on the contents and the required end-quality. Coding is always a trade-off between either high quality and large storage space, or reduced quality and reduced storage space.

Take uncompressed (raw) video originally stored with circa 20 MB/s. For a two hour movie, this would imply the need for 144 Gb storage. To put this two hour movie on DVD, we would need more than 30 DVDs of 4.7 Gb each. Moreover, this would require us changing DVD every 4 minutes (30 times in total). Thanks to MPEG-2 coding the amount of data can be reduced to the size of one DVD. The operation of MPEG-2 is explained next.

2.2.   The operation of MPEG-2

The operation of MPEG-2 is important to understand the working of the coding methods later. MPEG-2 is based on its predecessor MPEG-1. The difference between MPEG-1 and MPEG-2 is mainly that MPEG-2 improved the compression factor of MPEG-1 for storing video sequences and that MPEG-2 also allows for broadcasting.

Basically, MPEG-2 is based on the principle that most information in two successive pictures is identical. If the data is stored in one picture and is mainly the same in the successive picture, then the identical information could be stored only once to leave out redundant information. This saves data storage. In order to define which data can be left out, the video has to be looked at in more detail.

2.2.1. Composition of a video sequence

An original video sequence is divided in groups of pictures (GOP). These GOPs normally contain 12 to 15 pictures, which are called frames. Each frame is divided in macroblocks, sized 16 by 16 pixels. These macroblocks in their turn are divided in four blocks of 8 by 8 pixels. Macroblocks are combined into slices during the MPEG-2 encoding. The video sequence is divided in macroblocks and blocks from the moment of encoding until it is decoded again. Figure 1 shows a visual representation of the composition of a video sequence as explained above. The 16x16 macroblocks are the basic building blocks of an MPEG-2 frame, but the smaller 8x8 blocks are ultimately fed to the compression algorithm for encoding (Mitchell et al., 1996). MPEG-2 codes the video sequence using both intra frame coding and inter frame coding.

Figure 1 shows that a video sequence consists of groups of pictures (GOPs). Each GOP consists of at least one frame, but generally of 12 to 15 frames of three different types: I-, P-, and B-frames. I-frames (intra coded frames), are different from P-(predictive) and B- (bidirectional) frames in the sense that independent I-frames are coded using only the data that is present in that specific frame. P- and B-frames are inter- and bidirectional-coded frames that are dependent on one or two frames. I-frames allow for random access points in a video sequence and for functions such as fast forward or reverse.
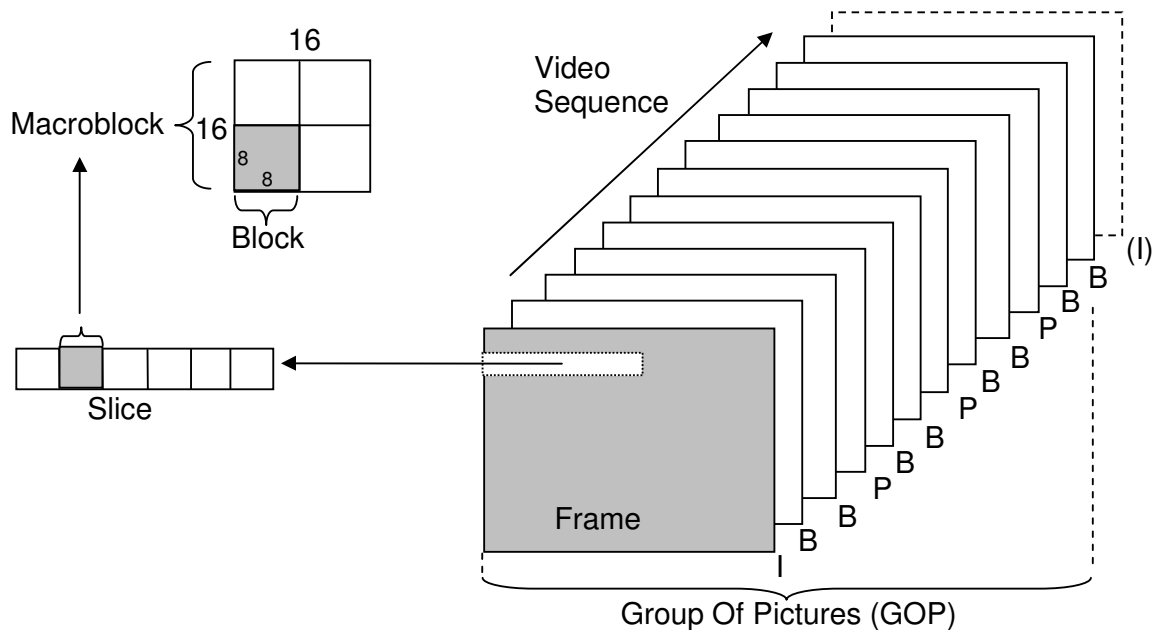
Figure 1: The composition of an MPEG-2 encoded video sequence based on a GOP of 12 frames.

2.2.2. Intra frame coding

Intra frame coding focuses on the spatial redundancy (redundancy within one frame) that might exist between adjacent pixels within a frame. At the basis of intra frame coding lies the discrete cosine transform (DCT). DCT decomposes a block of data into a weighted sum of spatial frequencies or, in other words, it transforms each 8x8 block from the spatial domain into the frequency domain. Figure 2 shows the 64 two-dimensional cosine functions of the 8x8 DCT. This lossless transformation separates the signal into independent frequency bands allowing further analysis of the spatial frequencies. After this transformation the block is represented by DCT coefficients and is ready for the next step: quantization.
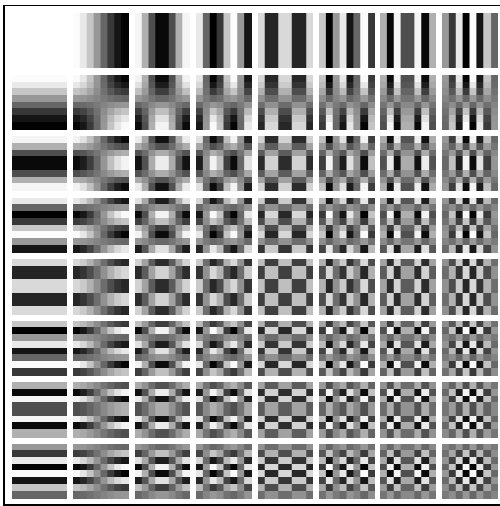
Figure 2: Two-dimensional DCT frequencies.

Quantization, as used for image processing is a lossy compression technique that involves mapping a range of values onto a single value. In this way the number of bits needed to represent the DCT coefficients is reduced. To get a better understanding of this method, it can be compared with the transformation of a continuous analog signal into a discrete digital signal as shown in Figure 3. There, a continuous signal is divided in eight parts (e.g. the analog ranges 0-0.4, 0.5-1.4, 1.5-2.4, …, 6.5-7.4 are translated into the integers 0, 1, 2, …, 7 in the digital domain). This analogy shows what is done in quantization for image processing; in this case it is done with the DCT coefficients.
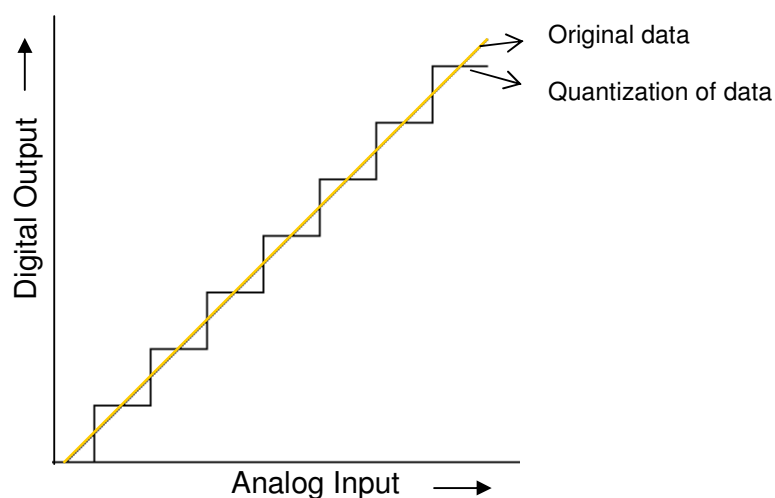
Figure 3: Translation of an analog input signal to a digital output signal is comparable to the quantization of data in image processing.

During the quantization, higher DCT coefficient values usually are quantized more coarsely than lower DCT coefficient values which would affect the quality more. This is because the human eye has a poor response towards high intensity variation. Hence, the eye will not notify the absence of patterns in the high frequency side (patterns in the lower right of Figure 2) which can be used to decrease the data size in compression.

To optimize compression, the resulting data is encoded in a zigzag order as Figure 4 shows. This zigzag ordering orders the two-dimensional array from the most significant element to the least significant element translating them in a one-dimensional array with longer rows of adjacent zeros as a result (Wallace, 1992; Mitchell et al., 1996).
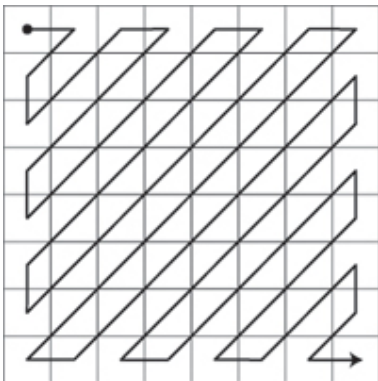
Figure 4: Zigzag ordering to optimize compression.

The one-dimensional array is then transformed using e.g. Run Length Encoding (RLE). RLE can shorten the code by transforming the long rows of zeros that were obtained in the zigzag process into a string that describes the number of zeros (Mitchell et al., 1996). For example, a string of four a's, three b's and twelve c's can be transformed into 4a3b12c using RLE. This is a reduction from 19 to 7 characters (i.e. a data reduction of 63% in this example). In addition, a variable-length coding (VLC) algorithm is used: the Huffman code (VLC, 2008). This code is used to compress the data even further (Mitchell et al., 1996). The Huffman code is based on differentiating between the frequencies of the coefficients; coefficients that are used often are described with fewer bits than the coefficients that are used rarely. In this way, fewer

bits are needed on average to describe the original data. The VLC (Huffman) encoding completes the intra frame encoding.

2.2.3. Inter frame coding

Additional compression to intra frame coding can be achieved using inter frame coding. Inter frame coding involves making use of temporal redundancy (i.e. redundancies between two frames) that might exist between adjacent frames in a video sequence.
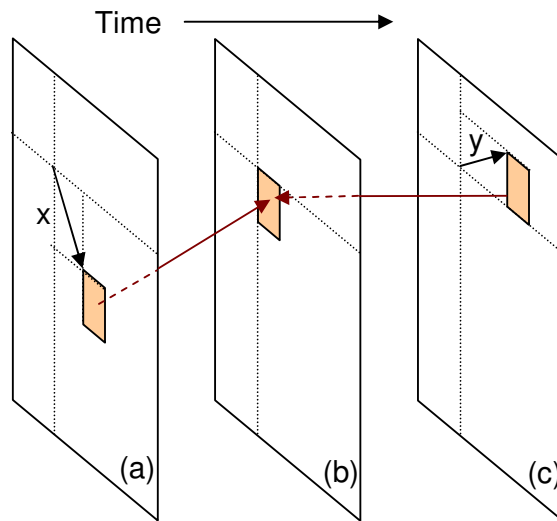
Figure 5: (a) Forward Prediction reference frame; (b) Current (target) frame; (c) Backward prediction reference frame. Motion vector x indicates where the macro block can be found in the previous reference frame and motion vector y indicates where the macro block can be found in the successive reference frame.

P-frames, or Predictive coded frames, are dependent on the previous I- or P-frame for both inter and intra frame coding. This dependency on other P- and I-frames is called forward prediction or unidirectional prediction. A target macroblock in the frame that has to be encoded (target frame) is matched with a set of displaced macroblocks in a previous reference (P- or I-) frame. The macroblock in the reference frame that is best matching the target macroblock in the target frame is used as the prediction macroblock. A motion vector to the reference macroblock indicates where the reference block can be found by describing the displacement from the target macroblock to the prediction macroblock (See frames (a) and (b) in Figure 5). If no macroblock in the reference frame is similar to the one that is needed in the target P-

frame, this macroblock is encoded using intra coding as if it were an I-frame. Any differences between the prediction block and the original block causes a prediction error. This error is then encoded using intra frame coding (see Section 2.2.2).

B-frames, or Bi-directional frames allow for motion compensation and are a key feature of MPEG (Brouwers, 2006). Unlike P-frames, that only make use of the previous reference frame, B-frames make use of the previous and next reference frame for prediction. This combination of forward (a) and backward (c) prediction of the target frame (b) is shown in Figure 5. Any macroblock in the target frame is predicted by either the previous, or the next reference frame, or can be predicted by a combination of both. When a combination is used, this is called interpolation. This involves taking the macroblock of the previous reference frame which is averaged with the macroblock of the next reference frame to create an even more accurate prediction (Brouwers, 2006).

2.2.4. Transportation and dependencies of the different frames

Frames are transmitted (i.e. sent via a data network) in a different order than they are displayed. The transportation starts with an I-frame. Typically, for a GOP (Group Of Pictures) of twelve frames three (four for a GOP of 15) sets of a P- and two B-frames are sent followed by a set of one I- and two B-frames. This is repeated for the whole video sequence. In this way the P- and B-frames that are dependent on other frames are always preceded by the frames they are dependent on. Figure 6 shows graphically the sending order (b) and the dependency relations of the I-, P-, and B-frames (a).

(a)

(b)
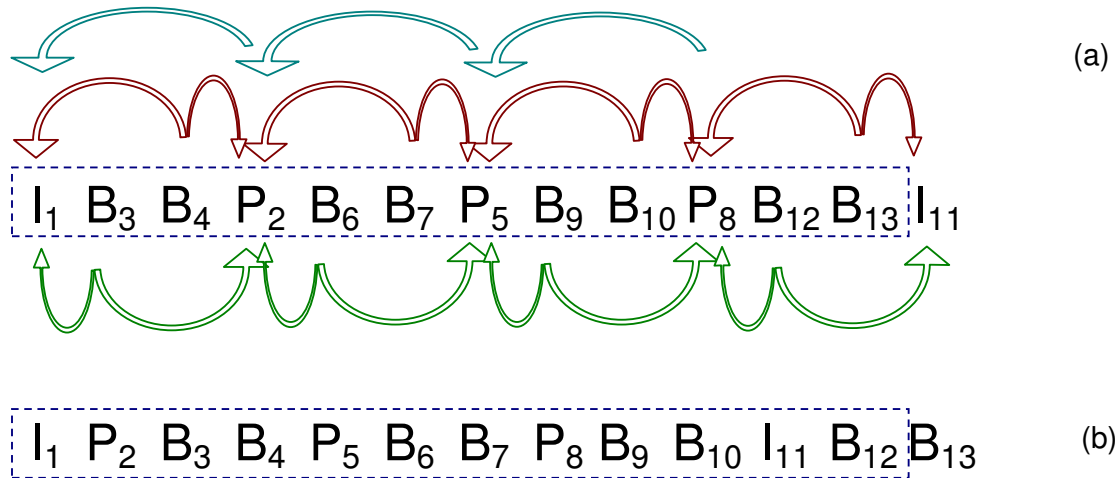
Figure 6: (a) A GOP of twelve frames (within the dashed box) and the interdependencies of frames. I-frames are standalones and are the foundation on which the video sequence is built. P-frames are rebuilt using the previous reference frames (I- or P-frames, see blue arrows). B-frames are reconstructed using the previous and next reference frames (See the red and green arrows). (b) Transport order of the frames of a video sequence.

I-frames are the building blocks and B- and P-frames are based on the assumption that a video sequence does not change drastically between two I-frames. This would make the prediction much more difficult for the B- and P-frames. The video sequence can be improved by adding extra I-frames called I-frame forcing, resulting in a decreased coding efficiency, but an increased video quality.

2.3.   Scalability

Scalable extensions allow the data stream to be partitioned into multiple pieces, or offer temporal flexibility so that not all frames have to be reconstructed (Mitchell et al., 1996). This property of MPEG-2 facilitates the extraction of video in more than one resolution or quality level from a single video stream. The data stream contains a base layer being the lowest layer (e.g. with the lowest resolution) and one or more enhancement layers (e.g. increasing the resolution). In total, MPEG-2 specifies four different scalability modes; temporal scalability, spatial scalability, SNR scalability (i.e. Signal to Noise ratio), and data partitioning (Jarnikov, 2007).

2.3.1. Temporal Scalability

Temporal scalability is originally intended to offer multiple temporal rates. A basic frame rate is stored in the base layer and can be enhanced by adding extra layers contributing to an increased frame rate.

2.3.2. Spatial Scalability

Spatial scalability refers to layered coding in the spatial domain (Brouwers, 2006). Using spatial scalability method, two different spatial resolutions can be encoded in one video stream. At the time of decoding, the base layer can be played with or without the enhancement layer. The signal for example can contain video in SD (SDTV) quality in the base layer and additional information in the enhancement layer to achieve the higher HD (HDTV) quality. Depending on the possibilities of the available hardware the highest possible quality can be selected at the moment of play back. In fact, the enhancement layer contains the coding loss between the base layer and the original signal.

2.3.3. SNR Scalability

In SNR Scalability, the enhancement layer serves to increase the Signal to Noise Ratio (SNR, i.e. how strong a signal is with respect to noise) to achieve a higher quality. Separate bit streams representing individual enhancement layers are generated carrying DCT coefficients that enhance the coefficients of the underlying layer. In general, the enhancement layers carry the coefficients that describe the quantization error (i.e. the difference between the quantized signal and the original signal) of the underlying layer which can improve the signal (Wenger, 1998).

2.3.4. Data Partitioning

Data partitioning is based on splitting the data in critical and non-critical information. For this purpose, the data partitioning approach breaks the block of DCT coefficients into two parts (Jarnikov, 2007). Each part contains different information; the first part contains the lower frequency coefficients and the second part contains higher frequency information that is able to enhance the first part.

2.4. *Wireless networks*

Scaling can be used to improve the quality of video transmitted over wireless networks. Wireless local area networks (WLAN) are standardized by IEEE under the number 802.11 to solve the incompatibility between products of different manufacturers (Tanenbaum, 2005, p.68). Another name for 802.11 is *Wi-Fi*. Wi-Fi has several categories with differences in transmission speed and frequency. Table 1 shows the available and future IEEE standards for 802.11. Bob Metcalfe, inventor of the Ethernet technology on which Wi-Fi networks are based, wrote in 1995 that mobile wireless computers are comparable to mobile toilets that lack a connection to the sewer system. He advised the world to buy cables and stay at home. The future will tell whether this statement falls in the same category as T.J. Watson's statement of 1945, that the world would not need more than 5 computers until the year 2000, in his explanation why IBM was not entering the market of the personal computer (Tanenbaum, 2005, p.23). The use of wireless networks is rapidly increasing and Bob Metcalfe might be proved wrong if he is not already. Table 1 shows that the increase of the number of standards for networks such as a, b, g, n, and y. The specifications of these versions can be found in Table 1.

Table 1: Wireless Networks, IEEE standards.

IEEE standard (802.11)	Frequency (GHz)	Typical throughput (Mbit/s)	Maximal Bandwidth (Mbit/s)
a	5.8	23	54
b	2.4	4.3	11
g	2.4	19	54
n (draft)	2.4 & 5.8	74	248
n (expected Jan 2010)	2.4 & 5.8	130	600
y	3.7	23	54

Table 1 shows that there is a difference between the maximal bandwidth and the typical throughput for the different standards for wireless networks. The difference lies in the fact that most transceivers are half duplex (Tanenbaum, 2005, p.129). A half-duplex system means that in the communication between two devices, only one device can send at a time. This asynchronous data transmission requires additional

data communication taking the place of the actual data that needs to be transmitted; this overhead lowers the throughput. Another negative influence on the throughput is unreliability in networks. This requires extra testing if sent data reached the receiver.

2.5.   Adaptive coding in wireless networks

Wireless networks are unreliable (packets might get lost) and sensitive to interference by other devices (bandwidth might fluctuate due to interference; see Chapter 1). This unreliability shows in fluctuations in the available bandwidth when the actual throughput is lowered. Video streaming is affected by these fluctuations and requires a solution that can be found using adaptive coding methods (Boyce & Gaglianello, 1998). These adaptation methods are able to handle data that is delayed by congestions in the network or by packet losses. This research focuses on two adaptation methods that are in use by Philips Research in Eindhoven. One method is based on temporal scaling (dropping frames) and is called IFD. This stands for I-Frame Delay. The other method is based on spatial scaling and is named TC, which stands for Transcoding.

2.5.1. I-Frame Delay

I-Frame Delay is an adaptation method based on the transmitting order of the frames in a video stream and the difference between transmitting & displaying order (See also Section 2.2.4). If the bandwidth is temporarily decreased, IFD drops the least important frames (Van den Ende, De Hesselle, & Meesters, 2007). The dropped frames are replaced by repeating the last frame that was shown (Kozlov et al., 2005). The first frames to be dropped when necessary are the B-frames since no other frames are dependent on them. If dropping B-frames only is not enough, the P-frames are dropped next. P-frames are the least important frames after the B-frames are taken out, leaving only the I-frames.

The selection of which frames to drop is done by the IFD algorithm. This algorithm tags the packets that contain the frames (data sent via wireless networks is divided in little data packets). The IFD scheduler on the sender site controls the send buffer and drops the packets with the least important frames if the buffer is filling up. The packets that have to be dropped are selected by the tag they received earlier by the IFD scheduler.

Without IFD, all packets are sent and a decrease in bandwidth would result in randomly lost packets. If such a packet would contain I-frame information, its loss would influence the entire GOP (twelve frames) due to the dependencies, resulting in grave distortions as shows Figure 7 (Boyce & Gaglianello, 1998). This could possibly have been reduced by dropping (the packets of) less important B-frames first resulting in artefacts such as jerky movement. This expresses itself in a bit shaky, not continuous video. The IFD algorithm drops as few frames as necessary and focuses on dropping packets that belong to one frame first.

2.5.2. Transcoding (TC)

Transcoding is an adaptation method that can be based on more than one algorithm (See also MCLAB Seminar Series, 2008; Simerly, 2008). The focus in this thesis is on transcoding by re-quantizing the DCT components. Re-quantizing the DCT components might introduce spatial distortions showing as spatial artefacts such as blocking and blurring. By introducing a controlled higher visibility of the spatial artefacts the risk of network induced video stream errors is reduced; decreasing the amount of unpredictable errors that are likely to be worse than the controlled distortions.

Idem to IFD, packets belonging to multiple frames might get lost without TC, including packets of I- and P-frames. If such a packet would contain I-frame information, its loss would influence the entire GOP (twelve frames) due to the dependencies, resulting in grave distortions as shows Figure 7 (Boyce & Gaglianello, 1998). This could possibly have been reduced by applying the TC algorithm (lowering the bit-rate) for a few frames. The algorithm lowers the bit-rate which might lower the quality in the video sequence, however, it prevents the sequence from even worse distortions due to random packet losses.

TC takes the MPEG-2 video sequence as input and adapts it by re-quantizing the discrete cosine transform (DCT) components on the fly, using a feedback loop (Brouwers, 2006). The feedback loop is continuously checking the bandwidth. The transcoding process starts as soon and as long as the feedback loop detects a decrease in bandwidth, which is insufficient for the video sequence. The result of this

Transcoding is that the bit-rate is lowered, but the frame rate remains unchanged. The collateral consequence of the decreased bit-rate is that the image quality will decrease, because the higher frequency DCT components are left out. The result is a decreased amount of detail in the video sequence by introducing blurriness and blockiness and thus a lower video quality.

2.6.   Perceived Artefacts

Artefacts might be introduced by the adaptations methods such as IFD and TC, but also in the encoding process when the MPEG-2 coding is applied due to prediction errors and the quantization process. These errors are less grave than errors resulting from packet loss since the lost information might influence multiple frames. Packets are mostly dropped because of flooded buffers during congestion in a network. Packet loss is extremely undesired because it introduces unpredictable artefacts resulting in significant quality loss. Sometimes it is wiser to prevent packet loss from happening by lowering the bit-rate (and video quality) of a video sequence. This might lower the quality but prevents an even worse quality loss (see Figure 7).

Figure 7: Resulting distortions due to uncontrolled packet loss.
Image source: *http://www.iptv-news.com*

The lowered bit-rate might introduce impairments (artefacts) to the video sequence that can be temporal or spatial. Temporal artefacts occur due to the temporal change in the area between consecutive frames. Spatial impairments are differences with the original image within one frame (Yuen & Wu, 1998). Table 2 and Figure 8 show a short overview of the most common introduced artefacts as explained by Yuen and Wu (1998).

Table 2: Temporal and spatial artefacts that can occur as explained by Yuen and Wu (1998)

Effect	Description
Jerkiness (temporal)	Jerkiness shows up when a drop and a repeat of a frame were used in a video sequence; the video stream does not seem to be continuous anymore, but is a bit shaky.
Motion compensation (MC) mismatch (temporal)	Motion compensation mismatch results in high frequency distortions close to edges of moving objects and are the result of prediction errors during the MPEG-2 encoding.
Stationary area temporal fluctuations (temporal)	Stationary area temporal fluctuations show as flickering in non-moving areas of the video sequence. This is the result of small variations in offset between successive frames.
Smearing (temporal)	Smearing shows up as a loss of spatial resolution and a blurring of features. This effect is only perceived if the object is tracked by the eyes. It is heavily dependent on the speed of motion of the object.
Ghosting (temporal)	Ghosting appears as a blurred trail behind fast moving objects. This is basically achieved when the representation of the current frame is the weighted average of the previous and present frame, thereby averaging out the uncorrelated noise.
Blocking (spatial)	Blocking shows up as little squares in edges inside an image that had a smooth edge in the original video.
Blurring (spatial)	Blurring is the loss of spatial detail and sharp edges. The result is a vaguer image, a reduction in sharpness and a loss of details.
Colour bleeding (spatial)	Colour bleeding often shows up when there is a border between two coloured planes. The result of a border between a white and a red plane with colour bleeding might be that red pixels are wrongly moved into the white plane and vice versa.
Ringing (spatial)	Ringing is seen next to sharp edges where a rippling can be noticed.
False contouring (spatial)	False contouring shows up especially in planes such as pictures of the sky. In stead of a smooth transient sky this is resembled by multiple planes introducing contours that were not present in the original picture.
False edges (spatial)	False edges come forth of blocking artefacts in previous frames that are wrongly detected as edges, although the real edge moved away already (with a moving object).

Figure 8: Examples of artefacts. From left to right:
Blocking – Blurring – Colour Bleeding – Ringing – False Contouring – False Edges

3. Perceived video quality

Business Dictionary (2008) defines perceived quality as the consumer's opinion of a product's (or brand's) ability to fulfil his or her expectations. It adds that it may have little or nothing to do with the actual excellence of the product. Focused on video this means that perceived video quality is not a direct result of the technical quality of the video (although it might be correlated), but whether people like the video quality or not (how it is perceived). How people see the video is important for the quality perception and is handled by the human visual system. Visual perception is a combination of top-down and bottom-up processes in the brain. The operation of this is necessary to understand the results of the study performed in the present research. The main functioning of the human visual system is covered in Section 3.1, including luminance and chrominance sensitivity. Section 3.2 discusses the perception of motion. Finally, Section 3.4 states the research question and hypotheses.

3.1.   Human Visual System

The human visual system is complex and not yet fully understood (Zhong et al., 2004). The present research does not focus on the details of the human visual system but on its main scope. Vision starts with light that travels through the lens of the eye and falls on the retina. Many different cells can be found in the retina all interacting to transport signals to the final destination: our brains. The cells in the retina are named cones and rods. These cells thank their name to the specific form they have. The human eye has circa 8 million cones and 120 million rods. The rods are spread over the retina but the cones are heavily concentrated in the center of the retina (Palmer, 1999). The cones allow us to perceive colour and are most active under high illumination conditions. The rods are much more sensitive to light but not to colour and take over when the illumination is low. This happens especially at night when there is only moonlight or starlight around (Blake & Sekuler, 2006; Tucker, 1997). The image that falls on the retina is transformed into a signal that is transferred via many cells and organs such as the optic chiasm and the lateral geniculate nucleus to the brain where it shows as a mirrored and stretched image topology (Palmer, 1999). The brain processes the signal resulting in a perception of the perceived signal. The process of perceiving visual stimuli is more complex than it is sketched here.

Zhong et al. (2004) mentioned that the complex human visual system is layered. Their theory states that the human visual system responds to movement, colour, texture, orientation, and shape at a low level of operation in ways that can be measured and defined. For a higher level of operation the visual perception is influenced by other factors such as the content of the scene, the task carried out by the observer, and previous experience or learning by the observer. Knowing that the human visual system is able to respond to things such as texture at a low level of operation raises the question how sensitive the human visual system is.

3.1.1. Luminance sensitivity

The human eye is sensitive to luminance intensity (dark and light) thanks to the rods, and to chrominance intensity (colour) thanks to the cones in the retina. Three different cones exist; sensitive to red, green, and blue light. They provide us with two components: one for perceiving red-green and one for perceiving yellow-blue. The sensitivity of the human eye is expressed using *spatial frequency* (cycles / degrees). Blake and Sekuler (2006) explain that spatial frequency refers to the number of light and dark regions (i.e. a grating) imaged within a given distance on the retina. To get an understanding of what one degree is you can look at the fingernail of your index finger at an arm's length which will form an image on the retina of one degree width (O'Shea, 1991). One degree width on the retina is determined to be equal to 0.3 mm (Blake & Sekuler, 2006). Besides spatial frequency, intensity difference between the bars of the grating (contrast) play an important role in our perception.

Contrast is related to the intensity difference between the light and the dark bars of the grating. If the difference is large, this means that the contrast is high; a low contrast means that the difference is small (Blake & Sekuler, 2006). Imagine a stripe pattern with black and white stripes of which we are able to change the intensity of the black stripes. If the contrast is lower than the threshold value the difference is not noticeable anymore and the stripe pattern would be perceived as one uniform plane. Also the direction of the grating is important.

The human eye is quite sensitive to changes in the horizontal and vertical direction. However, it is less sensitive to diagonal patterns (Van den Branden Lambrecht & Verscheure, 1996; Blake & Sekuler, 2006). The human eye is also less sensitive to

higher spatial frequencies than for the lower ones. Discrete Cosine Transform (DCT, See Section 2.2.2) makes clever use of this lack of sensitivity by leaving out the frequencies that the human visual system cannot perceive. This lower sensitivity might result in a masking of spatial artefacts which can be an additional advantage in the MPEG-2 coding process. Another decrease in sensitivity of the human eye can occur because of ageing. Visual acuity – *the spatial resolving capacity of the visual system* – (generally) declines when people grow older. This means that older people in general need a higher contrast to see the same details as they were able to see when they were younger.

The human eye is more sensitive to changes in luminance (brightness) than to changes in chrominance (colour). Usually most information about a scene is therefore not contained in its chrominance, but in its luminance. Because of this property, people have accepted monochrome projections for photo and video for a long time, until the change to colour video and pictures was made (Kingsbury, 2005). Figure 9 shows the contrast sensitivity functions of the human eye for luminance and both red-green chrominance and blue-yellow chrominance. The human eye has a spatial frequency response that is a band pass with a peak frequency around four cycles per degree (See luminance graph in Figure 9). This has been quantized as the contrast sensitivity which means that a signal is only detected by the eye, if its contrast is higher than the threshold defined as the detection threshold (Van den Branden Lambrecht & Verscheure, 1996).
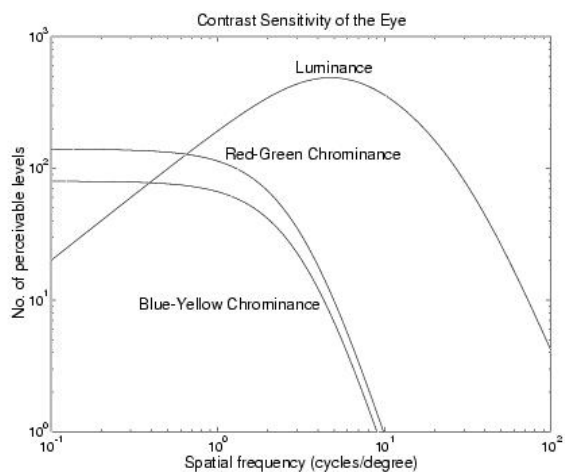
Figure 9: Sensitivity of the human eye to luminance and chrominance intensity changes (image source: Kingsbury, 2005).

3.1.2. Chrominance sensitivity

Several sensitivity levels for chrominance exist. Imagine for example a creature that contains four different types of cones, each with a unique absorption spectrum. This is actually not science fiction; goldfish are an example of such a creature equipped with tetrachromatic vision (Neumeyer, 1992). The extra cone in a goldfish's eye allows it to perceive ultraviolet light waves (Hunt et al., 2001). Hence, they have an even better colour discrimination than humans and other trichromatic species (Blake & Sekuler, 2006). It is possible that a person is not able to differentiate all colours, which means that this person has a colour vision deficiency.

A colour deficiency is sometimes wrongly called colour blindness (Blake & Sekuler, 2006). Physically one of the three cone types is then eliminated. The most common deficiency is red-green deficiency; people that have this deficiency have difficulty to discriminate the colours green and red. Circa 8 out of 100 Caucasian males (e.g. European), 5 out of 100 Asiatic men, and 3 out of 100 African-Americans have a colour deficiency. The occurrence of colour deficiency is higher for men than for women. It is also possible that people have no cones at all, but this is extremely rare and only found in 1 in 10,000 persons (Colour blindness, 2008). The present study focuses on people with normal vision (i.e. vision without any deficiencies).

3.2.   Motion Perception

The human visual system allows us to perceive motion. This is modelled by Van den Branden Lambrecht (1996). This spatio-temporal model incorporated the characteristics of the human visual system that are relevant to image processing applications able to deal with video sequences. Temporal perception studies produced considerable evidence for the existence of two different mechanisms of temporal vision termed *sustained* and *transient*. This is in line with Hammett and Smith (1992) who showed in their research that there can be only these two temporal mechanisms.

The sustained mechanism is sensitive to low temporal frequencies and accounts for perception of single wavelength or single stimuli. This was already discussed in Section 3.1.1. The eye's sensitivity is larger for moving objects than for static ones. In addition, the detection threshold of the foreground will be modified as a function of the contrast of the background, which is also known as masking (Van den Branden

Lambrecht & Verscheure, 1996). For the present research this is important, since any introduced distortion can influence the perception by making the distortion either strongly visible, attenuated or completely invisible by masking it in the scene (Van den Branden Lambrecht & Verscheure, 1996).
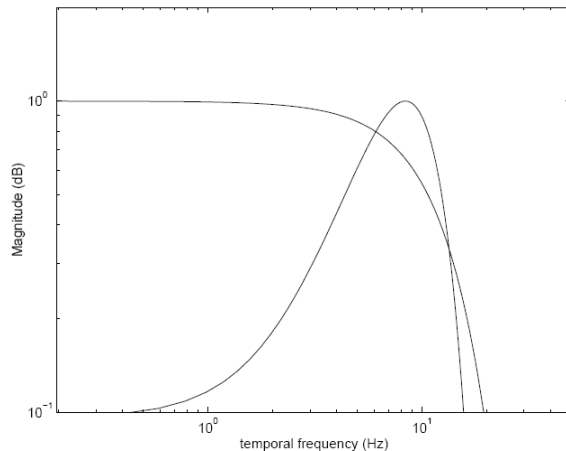
Figure 10: The temporal filter bank accounting for two mechanisms: one low pass (the sustained mechanism) and one band pass (the transient mechanism). The frequency response of the filters is plotted as a function of temporal frequency. (Van den Branden Lambrecht, 1996)

The transient mechanism is sensitive to band pass signals (allowing a range of frequencies to pass and attenuating frequencies outside this range). Physiological and psychological experiments on perception gave evidence of both, a band pass nature of the cortical cells' response in the spectral domain and of the well defined structure of the peak sensitivity positions and the bandwidths of the response of the cells (see Figure 10; Van den Branden Lambrecht & Verscheure, 1996). The sustained and transient mechanisms allow us to perceive motion which is important in daily life.

Imagine that you are walking through the woods and looking around at serene scenery of trees, plants, a little creek and some birds flying from branch to branch. Suddenly you are disturbed by a little creature that bounds away, scared by your footsteps. You never got a good glimpse but still recognized it as a rabbit by the hopping motion of the object. You did not notice the rabbit before, until its movements broke the camouflage.

This scenario illustrates that motion serves multiple perceptual purposes: detection (seeing something hop away), segregation (recognizing something move relative to its

background), guidance (turning toward the source of motion), and identification (the hopping motion identified the object as a rabbit). Blake and Sekuler (2006) explain that the neural processes responsible for perceived motion begin by extracting the spatial displacements of various environmental features over time from the retinal image. This makes it clear that motion is a spatio-temporal event. In the environment, many spatio-temporal events are taking place in the same time; some are in unison (e.g. birds flying back and forth merged into the background) and some are unique (e.g. a disturbed rabbit). On top of this, the body movements add their own spatio-temporal variations to the overall image. The visual system sorts all these variations into sets of events; aggregating local signals and segregating them from other signals (Blake and Sekuler, 2006).

People are capable of perceiving both apparent and actual motion. Apparent motion is motion that is perceived as motion, but the stimulus is not actually moving (Watson & Ahumada, 1985). An example for this apparent motion is a row of light bulbs of which the light is turned on and off in a sequential order. It seems as if the light is 'walking' along the line. That is why motion is a relation between objects and not a property of an object (Wertheim, 1985). Another example of apparent motion is video projection. Although the objects in the video sequence might seem to move, the content actually consists of different static pictures that are shown at a high frame rate. Actual motion is different from apparent motion in the sense that physical movement is involved. An example of actual motion is when you see a person walking on the street (in real life); there is an actual displacement of a person. Two kinds of actual motion are known which are called self-motion and object-motion (Alberta, 2001).

Self motion refers to the motion that is perceived by the observer, when the observer is moving himself (e.g. the observer is sitting in a moving train and watching the scenery passing by outside the window). Object motion refers to the motion that is perceived when the object moves but the observer does not (e.g. the train on the next track starts moving while the observer's train is stopped). The perception of these forms of motion requires the use of our eyes.

Our eyes are perhaps the busiest organs in the entire human body. More than 100,000 eye movements are made daily (Blake & Sekuler, 2006). There are two main

principal types of eye movements; smooth and rapid. If you look across a road you might see all kind of flashes from the cars that pass by. It might be even hard to force your eyes to look at a certain point, because they automatically tend to follow a car. If you would let your eyes follow a car to read the text on its side for example, you would see the car in much more detail. This is called tracking; a moving object is kept in the centre of the view (Blake & Sekuler, 2006). In this case, the eyes will move smoothly, much different from saccades. The eye makes many saccades (jumps) from one point to the next while we are reading a text for example. Saccades are rapid eye movements that serve the purpose to bring potentially interesting objects in the peripheral of our retina (with poor acuity) to the fovea (with good acuity). When focused on an object, the observer is missing out other information as is shown in the following experiment.

Simons and Chabris (1999) asked participants to look at a video and focus their attention on six people in black and white shirts playing with a basketball. The participants' task was to count the number of times that a player with a white shirt passed the ball. While counting and thus focussing on the ball, observers completely miss out a strange object (e.g. woman with umbrella or man in a gorilla suit) walking through the centre of the view. Hence, people easily fail to perceive parts of the video if their attention is focused on something else. The focus of people can be changed by (accidentally) priming participants.

3.3.   Quality Perception

McCarthy et al. (2004) asked participants in their study to tell if the quality was acceptable or not. McCarthy et al. (2004) showed that the perceived video quality of their video sequences with lowered frame rate (down to 6 frames per second) was rated acceptable in most cases (80 %) and claim that this challenges the conventional wisdom that sports coverage with high motion requires a high frame rate to maintain perceived quality of service. However, prior to the experiment, participants were briefed that a Telecom company was testing the minimum acceptable level of video quality. This might have led to an easier acceptance of lower quality since they were primed to tell what was minimally acceptable. This possible effect of priming has to be kept in mind when designing a study. The result might likely have been different if the

participants would have been asked to rate the quality in contrast to the task of finding the lowest *acceptable* quality; the participants might be more strict in the first task.

The fact that instructions are important shows also from the research of Zhong et al (2004), who studied differences in perceived video quality for different instructions. Observers were instructed to especially focus on the foreground or the background in the picture. Their study showed that foreground human figures in a video clip have a significant (perhaps dominant) influence on subjective quality rating, regardless of whether the viewer is concentrating on the human figure. In addition they found that an observer tends to rate subjective quality lower when their task involves concentrating on a distorted area of a scene than when they are not focused on a specific area. Therefore it is important to provide clear instructions and make participants focus on the topic of investigation such as foreground, background, or overall video quality. Thus, instructions and reason to watch a video are important for the quality perception. Video characteristics such as bit-rate and frame-rate are also varied to study influences of perceived video quality.

Hauske et al. (2003) studied the influence of frame rate and image quality on perceived video quality. Their focus was on video displayed on mobile phones with bit-rates of 30 to 128 kbps. Six different video sequences were shown with frame rates between 5 and 15 frames per second. Observers were asked to rate the videos for overall quality, smoothness of movement, quality concerning blocking effects and information value. The research showed that overall quality and quality for blocking effects were highly correlated for indicating the quality of separate frames. Smoothness of movement appeared to be a separate quality measure. An unexpected observation in their research was that an increase of frame rate (at the same bit-rate) does not positively influence the video quality, but the positive correlation between smoothness of movement and frame rate was in line with their expectations.

Haakma et al. (2005) found the same results for an increase of the frame rate (at the same bit-rate); no increase of overall quality. They used a bit-rate of 3 to 6 mbps and a 107 cm wide-screen plasma television. Haakma et al. (2005) suggest that the quality of the video might be more driven by image quality (i.e. quality of separate frames) than frame rate. A few years earlier Masry and Hemami (2001) coded video

sequences in 10, 15, and 30 frames per second using the same bit-rate for each frame rate projecting 352 x 420 video sequences on a 19" CRT monitor. Bit-rates varied from 40 to 800 kbps. Masry and Hemami (2001) found that participants prefer lower frame rates with a special preference for 15 frames per second encoded video sequences. Note that the mentioned studies were done with small displays making it difficult to draw conclusions on video quality perception for larger screens and higher resolution video. It is likely that image quality (i.e. quality of a separate frame) is an important predictor for video sequences' quality but in addition motion is expected to influence the video sequences' quality too as shows the next section.

3.4.   *Research Question and Hypotheses*

Although it is not completely clear what all factors are that influence the perceived video quality, it is clear that many aspects are important such as instructions (or reason to watch a video), temporal and spatial characteristics of the video content, image quality, focus, and it is argued that motion is likely to have an influence on perceived video quality. The influence of motion on perceived video quality has been investigated by researchers such as Tripathi and Claypool (2001), De Hesselle (2006) and Wijermans (2007).

3.4.1. Previous studies

De Hesselle (2006) experimented with short video sequences and impaired sequences of ten seconds with a burst of either one or five seconds, with either temporal or spatial scaling. The amount of motion of the shown video sequences was determined by De Hesselle only, not through a perception experiment or a motion measuring algorithm.

De Hesselle firstly concluded that the length of these impairments influence the final perceived video quality and secondly that preference for either temporal or spatial scaling was dependent on the scene content. In his research he differentiated between in-scene motion and camera motion. However, it was not possible for him to draw conclusions about the influence of the motion of the videos individually. That part was investigated by Wijermans (2007).

Wijermans (2007) argued that it is unclear whether the results of De Hesselle (2006) were the result of motion or of the content of the video. She concluded from her research that the spatially impaired video sequences always were assessed of having better quality than the temporally impaired sequences. Wijermans (2007) employed participants in a test and De Hesselle (2006) assessed the video motion personally to estimate the amount of motion of the video sequences and used these motion measures in later tests. Earlier, Tripathi & Claypool (2001) used an objective motion measure calculated by an automated motion estimation algorithm to calculate the amount of motion in MPEG-1 coded video sequences. Wijermans (2007) did not find a relation however Tripathi and Claypool (2001) did find a relation between amount of motion and temporal and spatial scaling. Also De Hesselle (2006) found evidence for such an influence. Differences between these researches are discussed more thoroughly later.

De Hesselle (2006) and Wijermans (2007) both made use of I-Frame Delay and Transcoding, two coding techniques. Wijermans (2007) showed that Transcoding is always better perceived than I-Frame Delay, but De Hesselle (2006) showed that there might be a trade-off between both coding methods depending on the amount of motion in the video sequences. Both Wijermans (2007) and De Hesselle (2006) could not compare the quality ratings of different stimuli since stimuli of different video sequences were not experimentally compared and thus not relative to each other.

3.4.2. Present research

The present research requires participants to compare stimuli in such way that the quality rating of each stimulus is relative to the quality ratings of all other stimuli. The different video sequences will be selected on the absence of scene changes that account for a masking effect of up to half a second (Seyler & Budrikis, 1965) and based on different levels of motion for each video sequence.

These different levels of motion allow us to study whether the quality of Transcoding is indeed always perceived better than I-Frame Delay as is claimed by Weijermans (2007) or if there is a trade-off between the two coding method as discussed by De Hesselle (2006) and Tripathi and Claypool (2001). I-Frame Delay and Transcoding are

used to represent temporal scaling and spatial scaling respectively. The amount of motion is important and can be assessed objectively or subjectively.

The present study opted for an objective algorithm named 3DRS (De Haan, et al., 1993; see Section 3.5) for the motion estimation, which will assign videos to different levels of motion. The reason to opt for an objective algorithm over assessment of motion by humans is that the latter is time consuming and expensive. Van den Branden Lambrecht & Verscheure (1996) explain that human vision is needed to assess the video quality to find a good correlation between the objective and the subjective judgment. An objective algorithm would allow for the automation of this process. Tripathi and Claypool (2001) stated that content-aware scaling could improve the perceived video quality by as much as fifty percent (compared to content-unaware scaling). Being able to define what coding method should be used in specific cases can contribute to this quality improvement. This approach is summarized in the research question of this research:

What is the overall effect on perceived video quality of temporal and spatial scaling for different levels of motion in video sequences?

It can be expected that the perceived quality for temporally scaled video sequences is higher than spatially scaled video sequences in case of a low level of motion in a video sequence. People are expected to easier notice when video becomes at a high level of motion than at a low level of motion. If the amount of motion is small enough it is expected that viewers might fail to notice the lower quality as a result of the missing frames because video would become little or not shaky at all. However, if the motion is high enough, it is expected that the lower quality is more obvious because the missing frames result in a shakier video sequence.

For a low level of motion, it is expected that the perceived video quality for spatial scaling is lower than for temporal scaling. The reason for this assumption is that the video sequence is likely to be perceived as fluent at a low level of motion for temporal scaling. Because only minor differences are present between the two frames surrounding a lost frame, this is likely to result in a video quality that is perceptually little influenced. The quality of the remaining frames is little or not influenced because

frames that do not influence others are dropped first. Spatial scaling influences the quality of each frame. If the dropped frames in temporal scaling are at a low level of motion indeed little or not noticed, this might lead to lower quality perception for spatial scaling than for temporal scaling.

For a higher level of motion it is expected that the perceived video quality is higher for spatial than for temporal scaling. The reason for this assumption is that shakiness of temporal scaled video sequences will become more noticeable at higher motion levels. It is expected that at a certain high enough level of motion this shakiness will lower the quality of a video more than the quality is lowered by spatial scaled video. These expectations are summarized in the following two hypotheses:

H_I Temporal scaling provides better perceived video quality than spatial scaling for video sequences with low motion levels

H_{II} Spatial scaling provides a better perceived video quality than temporal scaling for video sequences with high motion levels

To test our hypotheses in order to answer the stated research question, Section 3.5 describes an algorithm to classify video sequences according to their degree of motion. Afterwards, the motion classification algorithm will be used to select three video sequences with different levels of motion.

3.5. Level of motion estimated with 3DRS-algorithm

If the influence of artefacts on video sequences could be estimated by an objective measure of motion, this would allow us to assess the video quality technically, without needing human observers. Although subjective evaluations by human observers are the most reliable way of assessing quality of an image or video, they have the disadvantage of being inconvenient, slow and expensive to collect (Wang et al., 2003).

Algorithms for motion estimation are important because motion estimation is a critical component of almost every video coding system (Pan et al., 2007). Motion is a descriptive of the dynamic nature of a motion picture. For example, sports footage would be categorized as "high temporal data" (high level of motion) because of rapid scene changes, while talk shows would be categorized as "low temporal data" (low

level of motion) by the static nature of the talking heads (Apteker et al., 1995). Wijermans (2007) used participants to estimate the level of motion for the stimuli she used. In addition, she investigated whether an objective motion estimation measure could differentiate between the amounts of motion in camera motion and in-scene motion. For this, she used respectively Spatial Information and Temporal Information as defined by Wolf and Webster (1997). The present research makes use of a similar objective motion estimation measure to assign video sequences to different levels of motion. The focus in the present research is on overall motion. Future research may focus more on the different components that contribute to the overall motion in respect to the perceived video quality. For the present study, the algorithm for motion estimation needs to be able to discriminate between different levels of motion. Many papers describe the use of such algorithms but few authors are willing to share the algorithm's source code.

Many algorithms that were considered generally use block matching methodologies that have been popular over a long time and that nowadays still are (Li & Salari, 1995; Cafforio & Rocca, 1976; Ismaeil et al., 1999; Pan et al., 2007). Block matching algorithms divide the current image in (macro) blocks as done in MPEG-2 encoding. Next, a corresponding block is searched for each block in the present frame using search algorithms such as the exhaustive search algorithm, the three-step search (Li & Salari, 1995; Ramachandran & Srinivasan, 2001), or one of the many others. Most of them are used to predict which macro block is most likely to be a good replacement for the present block to achieve better results in the encoding process. However, for the purpose in the present study the most important thing is to obtain a predictor for the overall amount of motion and most algorithms are the same in that sense.

Koivunen and Salonen (1994) used combined shape and edge matching in addition to block matching to achieve a better performance. The algorithms Spatial Information and Temporal Information define the complexity of a scene (Wolf & Webster, 1997; Correia & Pereira, 2003). This can be used to derive a motion measure (Wijermans, 2007). The more complex a scene is, the higher the motion level of the scene.

Tripathi and Claypool (2001) used an algorithm that made use of the number of interpolated macroblocks to assign video to different levels of motion. The less

interpolated macroblocks found in a video sequence, the more changing parts, hence the sequence has a higher level of motion (Tripathi & Claypool, 2001). Hesseller and Eickeler (2005) developed an algorithm that was intended to detect camera motion. This algorithm made use of information such as DCT coefficients and motion vectors. De Haan et al. (1993) introduced a true-motion estimation algorithm with 3D recursive search block matching (3DRS). This algorithm provided a fast convergence and high accuracy

A useful implemented algorithm was found within Philips Research in the 3DRS algorithm that was developed by De Haan et al. (1993). The software made available for the present study made use of the 3DRS algorithm. It was modified to assess the level of motion. The algorithm was modified to take the average of both horizontal and vertical components of the motion vectors that were calculated in the algorithm and built a resulting average motion vector from these values with the length in number of pixels. The absolute value of this resulting average vector was used as an indication of the amount of motion of a video sequence and was named the motion estimation measure. The more motion a video sequence has, the higher this motion estimation measure will be. If there is little or no motion present in the video then the motion estimation measure will be close to zero.

The motion estimation measure indicates thus the average motion of a video sequence as was desired for the present study. However, this measure becomes less indicative for longer video sequences, because a video sequence might consist of faster and slower scenes. A scene is considered to be the part between two scene changes (Seyler & Budrikis, 1965) and does generally not vary much in level of motion (Meng et al., 1995). The present study focused on short, ten second video sequences without scene changes.

4. Method

The method section explains the setup of the present study, describing participants (Section 4.1), stimuli (Section 4.2), the design (Section 4.3), apparatus (Section 4.4), and procedure (Section 4.5). The present research is developed to be in line with the ITU standards (ITU-R BT.500-11, 2002). The ITU standardized methods for subjective video quality assessments (ITU-R BT.500-11, 2002). A typical test scenario according to ITU involves a human observer (e.g. an end-user) watching a series of video clips and rating each clip in terms of its perceived visual quality (Zhong et al., 2004). Before the present study was started an explorative pilot was performed in which the full research cycle was completed. Four major improvements were implemented in the current study based on the results of the pilot study.

One issue was that the pilot study showed that participants had trouble comparing High Definition quality video sequences; only eight of twenty participants showed a significant correlation between two rating sessions. For this reason the current study uses video sequences in Standard Definition quality instead. Another issue was that the experiment lasted one and a half hour including a break, which participants experienced as a long time. Participants scored each stimulus in comparison to each other stimulus twice, resulting in 72 comparisons in total. To shorten the experiment time in the current research only a wise selection of comparisons was shown to the participants: two stimuli were compared only, if the coding method or video sequence matched. A third issue was that the levels of motion were possibly not different enough, which was solved by selecting video sequences that had a larger difference in motion. A final issue was the number of trials in the practice round which needs to be "a few" according to the ITU (ITU-R BT.500-11, 2002). The participants were shown three trial comparisons in the pilot study. The results indicated that this was not enough, because, after leaving out the first 10 comparisons the answers of the participants became significantly more consistent. In the current test the number of trial runs is doubled to six to improve the consistency of the answers.

4.1.   Participants

Participants were mostly students and employees of Eindhoven University of Technology. In total 26 participants (17 males, 9 females) volunteered to participate.

Amongst them were thirteen students and thirteen employed people (mostly TU/e employees). Both gender groups did not differ significantly and are therefore considered as one group in the analysis. Most participants had at least a visual acuity of 1.0 / 1.0 (left / right eye, normal sighted people), only 5 participants showed a lower visual acuity (left / right: 0.8/1.0, 0.25/1.5, 0.8/0.8, 0.8/0.8, 0.65/0.8, less than normal sighted people). Both groups (normal / less than normal sighted people) did not differ significantly and are therefore considered as one group in the analysis. None of the participants had any form of colour deficiency. There was also no one who had participated in a video quality experiment within the last year. The participants varied in age from 21 to 59 (average $M = 34.38$, $SD = 12.43$).

4.2. Stimuli

The video sequences that are used in this study were selected from the VISA database of Philips Research which contains videos that are available for research purposes. Five progressive scan based video sequences were selected from the VISA database of Philips (See Table 3). All video sequences were of Standard Definition quality and had at least a consecutive ten seconds interval without scene changes. The degree of motion in each of the sequences is measured by a motion estimation algorithm and Table 3 shows the stimuli and their level of measured motion.

Table 3: Motion estimation SD video sequences (Numbers are the estimation of the amount of motion)

	Beach	Gladiator	Hospital	Office	Phantom
Motion Estimation Measures					
SD 720x576	10.9	2.0	14.0	22.5	14.0

Based on this motion assessment the video sequences "Gladiator" (2.0), "Beach" (10.9) and "Office" (22.5) were selected (see also Table 4). "Gladiator" and "Office" were chosen because the algorithm showed that these were respectively the slowest and fastest video sequences. "Beach" was selected as the final stimulus because the motion estimation value of that video sequence was closest to the average level of motion of the other two selected video sequences. "Gladiator" is taken from the movie

The Gladiator (2000). "Beach" is taken from the movie Jurassic Park 2 (1997) and

"Office" is taken from the movie High Crimes (2002) (see also Appendix C).

Table 4: SD video sequences selected for the main study assessed by the motion estimation algorithm

1 – Gladiator	2 – Beach	3 – Office
Very slowly moving man, walking forward for a few steps	Normally walking girl and woman, and static people on the beach	Fast walking persons in an office setting
Motion Estimation Measure: 2.0	Motion Estimation Measure: 10.9	Motion Estimation Measure: 23.5

Each video sequence was used in all three MPEG-2 encoded conditions: Reference

condition (REF), I-Frame Delay condition (IFD) and the Transcoded condition (TC).

The reference version (REF) of each video sequence was used with its original bit-rate

of 3 Mbps. Both the IFD version and the TC version of the video sequence had a bit-

rate of 2 Mbps (See Table 5). IFD and TC coding are applied in the complete video

sequence (from start to end).

Table 5: The three conditions for each video sequence REF, IFD, and TC

Condition	Bit-rate	Comments
Reference (REF)	3 mbps	Original
I-Frame Delay (IFD)	2 mbps	Every second B-frame dropped
Transcoding (TC)	2 mbps	DCT components re-quantized

To code the video sequences, a tool named Simulation Toolkit was used which was

especially developed for this kind of research by De Hesselle (2006). For the IFD

coded versions of the video sequences every second B-frame was dropped and for

the Transcoded video sequences the bit-rate was decreased by re-quantizing the DCT

coefficients. The IFD and TC coded versions of a video sequence were created

beforehand and saved as separate video sequences (stimuli). This was done to make

sure that each participant would see exactly the same distortions when these stimuli

were played back. The only difference between participants was the order of the stimuli that was varied.

To correct for possible order effects, participants were randomly assigned to two categories. Group one would compare the comparisons 1 to 18 in Session A and comparisons 18 to 1 in Session B. Group two would compare comparisons 18 to 1 in Session A and comparisons 1 to 18 in Session B. In addition, the order of the comparisons was changed in Session B; Session B showed Stimulus B versus Stimulus A instead of Stimulus A versus Stimulus B in session A.

4.3. Design

Participants were assigned to a 3 (*Level of motion*) x 3 (*Coding method*) within subject design. They had to compare different stimuli to assess the perceived video quality (dependent variable) of the shown video sequences. Each comparison consisted of a comparison number (3 seconds), a comparison of two stimuli as explained below, and a voting screen (5 seconds) allowing the participants to note their rating. The comparison of two stimuli was built up by a sequence showing a screen with letter A (3 seconds), Stimulus A (10 seconds), a screen with letter B (3 seconds), and Stimulus B (10 seconds), which was repeated twice (See also Figure 11). Participants rated the quality of the second stimulus compared to the first stimulus on a seven point category scale (See Figure 12).
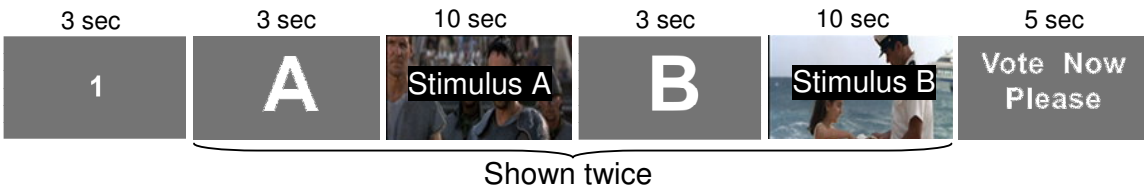
Figure 11: Stimulus order for a comparison between two stimuli. Participants would see the comparison number, twice a sequence of a screen showing an "A", stimulus A, a screen showing B, and Stimulus B, and finally a "Vote Now Please" screen.

B is much worse than A	-3	-2	-1	0	+1	+2	+3	B is much better than A

Figure 12: Rating scale for Direct Comparison as used in the present study where observers had to circle one of the options -3 (Stimulus B was perceived much worse than Stimulus A) to +3 (Stimulus B was perceived much better than Stimulus A).

The level of motion variable is identified by three different video sequences. The first level of motion (the lowest) is represented by video sequence 1, "Gladiator". The middle level of motion is represented by video sequence 2, "Beach". Video sequence 3, "Office", represents the highest level of motion. In addition, each video sequence is varied into three different versions identified by the coding method variable: an I-Frame Delay (IFD) coded version, a Transcoded (TC) version, and an unchanged reference (REF) version. This resulted in a total of 9 stimuli.

The stimuli are named in the format nX with n being the video sequence number 1, 2, or 3 and X being the coding method REF, IFD, or TC. To stay well within the maximum time of thirty minutes for each session as set by the ITU (ITU-R BT.500-11, 2002) only a selection of all possible comparisons between these stimuli was evaluated. Two stimuli were compared only, if the coding method or video sequence matched. This results in 18 comparisons per session. The three different coding methods (REF, IFD, and TC version) were compared with each other per video sequence. In addition, each video sequence's coding version is compared with the equivalent coding version of the other video sequences (e.g. the REF versions between video sequence 1, 2, and 3 were compared). The 'X' marks in Table 6 identify the evaluated comparisons.

Table 6: Comparison of all stimuli; all comparisons marked with "X" are made, a stimulus is not compared with it self

	1REF	1IFD	1TC	2REF	2IFD	2TC	3REF	3IFD	3TC
1REF		X	X	X			X		
1IFD	X		X		X			X	
1TC	X	X				X			X
2REF	X				X	X	X		
2IFD		X		X		X		X	
2TC			X	X	X				X
3REF	X			X				X	X
3IFD		X			X		X		X
3TC			X			X	X	X	

4.4. *Apparatus*

The video sequences were played using a 3 GHz Pentium 4 PC with 1 GB of memory and a 150 GB hard disk. This machine was connected to a television set using a DVI to HDMI cable. The television was a Philips Cineos 52" model (see Appendix A) which was set to show the video sequences with a resolution of 800 x 600. The stimuli

appeared on the screen with black bars on each side and the grey screens in between were shown in full screen. All menus were checked to switch any internal improvement and ambi-light functions off.

The experiment was set up in the UseLab of the IPO building at the TU/e. The room was furnished as a living room to make the participants feel comfortable and to represent a home situation similar to where participants would normally watch television. The television was positioned at a three meters distance from the sofa where the participants would sit and watch the footage (See Figure 13). The viewing distance was five times the picture height (60 cm).

Figure 13: Experimental setup. A comfortable couch was set up in front of the television set to create a home like situation for the participants (Left: view from behind the couch, right: side view)

4.5. Procedure

Maximally two participants were allowed to participate per experiment to assure a good view on the screen for each observer. Participants were asked to enter and were offered a drink. After assigning them to their places, participants were provided with instructions and paper forms to fill in their answers. They were asked to provide their demographic information and to read the instructions on the second page (See Appendix B). Next, their visual acuity was measured using a Landolt C-Chart and the participants were checked for any colour vision deficiencies using the Ishihara test (Ishihara, 1962).

The experiment was started with a test round of six trial runs. All data acquired with these trial runs was discarded; only data acquired during the actual comparison test was used to build up the dataset. The sample set was the same for all participants.

The six test comparisons (pairs: 1REF-1TC, 1IFD-2IFD, 2REF-2IFD, 1REF-3REF, 3IFD-3TC, 2TC-3TC) contained all stimuli and sorts of possible comparisons. This allowed the participants to get used to the stimuli, to the different categories (REF, IFD, and TC), to all sorts of possible comparisons, and to the rating scale. Thus the sample run contained all extremes to give the participants a good impression of what to expect from the upcoming task.

If the participants had any pending questions after the practice trials, these questions were answered if they did not influence the study, otherwise answering questions was postponed until the experiment was completed. Next, the first round of 18 comparisons was started (Session A). After finishing the first round people were invited for a drink to introduce a short break, after which they were presented with the second set of 18 comparisons (Session B). To conclude the experiment participants were asked to write down any remarks. Any pending questions about the experiment were answered. On average, the experiment lasted 45 minutes.

5. Results

An explorative analysis of the data showed that all participants' ratings are significantly correlated between Session A and Session B. The relation between the first and the second trial were all significantly correlated (correlations ranged from $r_s = .548$, p (one-tailed) $< .01$ to $r_s = .900$, p (one-tailed) < 0.001). This is much higher than .40 from which a one-tailed correlation with 16 degrees of freedom and $p < .05$ is significant. This means that participants had a preference for the same stimuli each time they rated a comparison irrespective of the order in which the comparisons were presented and irrespective of the order in which each pair was compared.

After completing the explorative data analysis, the comparison data collected in the present study was analyzed. The analyses of the comparisons are based on Thurstone's law of comparative judgment, which is used because people are not good in directly reporting perceived attribute strengths for presented stimuli on a one-dimensional scale (Torgerson, 1958; Boschman, 2001). One-dimensional scales are used for measuring attributes such as brightness, loudness, and perceived video quality. People are much better in comparing two things such as defining which one of two stimuli is brighter, louder or has higher quality than the other. This sort of comparison is called Direct Comparison.

To analyze comparison data such as acquired using the Direct Comparison method it is possible to use the statistical software program Difscal. Difscal is a program developed by Boschman (2001) to analyze data of comparisons that is based on the Thurstone scaling model. Thurstone's judgment scaling methods can be used to locate stimuli on a psychological continuum and thus make the responses of the participants meaningful. Thurstone stated that a 'discrimination process' mediates each psychological magnitude and every stimulus that is presented to an observer (Torgerson, 1958). Difscal uses the same constraints as in Thurstone Case V of one-way classification. The constraints of Difscal are that attribute values of all stimuli pairs are uncorrelated, that dispersions of all conditions are equal, and that subjects use fixed intervals corresponding with the different categories (Boschman, 2001).

Because of these constraints in Difscal, the standard deviation is independent of the stimulus pairs and thus constant. Difscal produces estimates of Scale Values – the mean of the distribution – and their corresponding S-Estimates. These S-Estimates are estimated errors but can be considered as standard errors and are thus an estimation of the error of the Scale Value. In this way, the difference of two Scale values belonging to two different stimuli can be checked for significance by performing a t-test. Two stimuli are rated significantly different in the experiment if the interval of a stimulus (Scale Value ± S-Estimate) is separated by at least the size of the largest S-Estimate of the two stimuli.

In the experiment, participants are provided with two stimuli on each trial and are asked to define the quality of the second stimulus compared to the first stimulus on a seven point Likert scale. To calculate the abovementioned Scale Values and S-Estimates Difscal accepts frequency distributions per category for each stimulus pair that was presented in the experiment (Boschman, 2001). Finally, Difscal shows whether the data fits the model well (model fit) using three different methods: The rule of thumb for *probability stress*, *Mosteller's Chi-square test*, and *Wilk's likelihood ratio*.
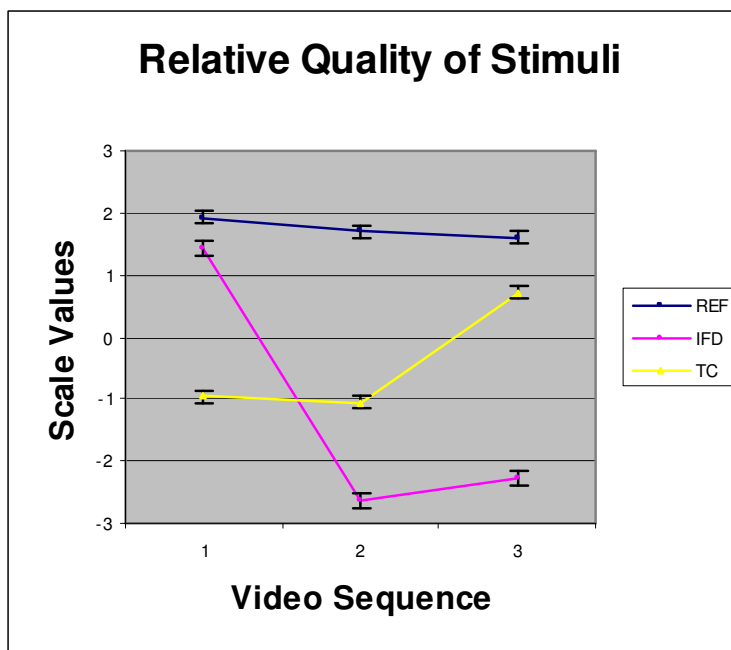

Figure 14: Scale Values with S-Estimates used as vertical error bars plotted for the reference and both coding methods per video sequence (level of motion).

 The analysis with Difscal showed that the data fitted the one-dimensional model well for all three tests. The data from the analysis with Difscal is used to make a graphical representation as shown in Figure 14. The S-Estimates produced by Difscal are measures for the accuracy of the estimated Scale Values. These were used as standard deviation. This shows the quality rating for each video sequence per coding method (REF, IFD, and TC).

Figure 14 shows several interactions. The stimuli 1REF, 2REF, and 3REF are not significantly different from each other. 1IFD was significantly different from both 2IFD and 3IFD, however, there was no significant difference between 2IFD and 3IFD. 1TC and 2TC were not significantly different, but 3TC was significantly different from both 1TC and 2TC.

All IFD and TC stimuli were perceived as having a significantly lower quality than the reference stimulus. The quality of TC coded versions was perceived higher than the quality of IFD coded versions at higher levels of motion (i.e. video sequence 2 and 3). The quality of the IFD coded version was perceived higher than the quality of the TC coded version for a low level of motion (video sequence 1).

6. Discussion & Conclusion

Participants of the current study compared different video sequences (each with a different level of motion) and different coding methods applied to them. They were asked to rate the overall video quality. Because the data of the complete dataset fitted the proposed model well it is possible to compare the quality of different stimuli directly via calculated Scale Values. These Scale Values show that participants perceived the quality of temporal scaling (I-Frame Delay coded video sequences) higher than the quality of spatial scaling (Transcoded video sequences) for a low level of motion. For higher levels of motion this was the reverse.

Section 2.5.1 explained that I-Frame Delay coding makes use of dropping frames (every second B-frame in this study), leaving each remaining frame's quality intact. This generally results in shaky movement within the video sequence that is visible at the places where a frame was dropped. If the level of motion in such a video sequence is low, it is likely that a viewer will not or less likely notice this drop in quality.

For example, in the extreme case of a video sequence without motion (i.e. a series of the same picture) it is possible that people would not notice one missing frame. In an other extreme event of three frames showing a ball moving from the outer left via the centre to the outer right of a television screen in three steps all movement could be lost when the middle frame was dropped. Viewers might perceive this as a ball jumping (instead of moving) from left to right.

The current results show that the participants indeed perceived the quality of the I-Frame Delay version of the video sequence with the lowest level of motion higher than the quality of its Transcoded version. Even more, the participants rated the quality of the I-Frame Delay version almost as high as the quality of the original reference version although the difference in quality is still significant.

For the higher levels of motion the quality of the I-Frame Delay versions drops drastically compared to the reference. This is the area in which the missing (dropped) frames become visible and manifest themselves in coding artefacts such as jerkiness,

blocking, and blurring. Contrary to the I-Frame Delay coded version, the Transcoded versions of the video sequences show a completely different effect.

Transcoding lowers the quality of each frame by introducing artefacts such as blocking and ringing, but does not introduce shakiness as I-Frame Delay does. Participants notice a degradation of the video sequence's quality in the Transcoded version easier than in the I-Frame Delay coded version for the video sequence with the lowest level of motion.

The Transcoded version is perceived significantly better than the I-Frame Delay version for the higher levels of motion. No significant difference is measured between the Transcoded versions of the video sequences with low and middle level of motion. However, the Transcoded version of the video sequence with the highest level of motion is significantly different from the other two Transcoded versions and is rated almost as high as (although still significantly lower than) the reference version.

A possible explanation that the TC versions are rated higher in quality than the IFD video versions at higher levels of motion lies in previous studies. Hauske et al. (2003) showed that overall quality and quality for blocking effects were highly correlated for indicating the quality of separate frames with smoothness of movement a separate quality measure. Haakma et al. (2005) showed that at the same bit-rate, a frame rate increase did not increase the overall quality. Haakma et al. (2005) and Masry and Hemami (2001) suggested that the quality of the video might be more driven by image quality (i.e. quality of separate frames) than by frame rate (which was kept the same for all stimuli in the present study). This could explain that participants rated temporally scaled stimuli higher than spatially scaled stimuli in the low level of motion situation.

For the higher levels of motion the perceived video quality of TC increases. Fischer (2004) argued that the difference in temporal and spatial characteristics might explain why some video sequences can be compressed almost without error while other video sequence show strong blocking effects. Somehow, the effect of image quality is overruled by another influencing factor. It is hypothesized that participants accept the lower quality for spatially scaled video sequences easier than temporally scaled video sequences at higher levels of motion because they are more focussing on what is

happening in the video and are trying to keep up with the actions. This allows participants less time to consider the whole screen because the content changes rapidly. For video sequences at a low level of motion (where fewer things are changing) participants have more time to contemplate the whole screen and rate the quality. Further research is needed to prove this assumption.

Another reason that the participants prefer temporal scaling in a low motion level and spatial scaling in higher motion levels might lie in the two different mechanisms of temporal vision termed sustained and transient (Van den Branden Lambrecht, 1996). The sustained mechanism is responsible for slow or no motion perception. The transient mechanism is responsible for moderate and fast motion perception. The human eye is more sensitive to moving objects which might result in the fact that participants focused on the main figure and thus not seeing the lowered quality in the rest of the screen, leading to the effect called masking (Van den Branden Lambrecht & Verscheure, 1996).

Saturation is another effect that might have influenced the results (Haakma, et al. 2005). This is when image quality reaches the point where observers do not see whether the quality gets worse anymore, because the quality is already too bad to notice these differences. A similar effect occurs when the image quality becomes too well to notice any further quality increase. This is possibly the reason that observers do not perceive differences between the degraded stimuli 2IFD and 3IFD, or between 1TC and 2TC.

Participants in the current research preferred the quality of temporally scaled video sequences over the quality of spatially scaled video sequences. It is likely that there exist a trade-off between spatial and temporal coding for which the preference for both coding methods will be the same by the participants. The present research has not found a specific point or interval where the quality for both temporal and spatial scaling is equal. Future research might focus on finding this point for MPEG-2 coded video sequences, which Tripathi and Claypool (2001) also suggested for MPEG-1 coding.

Tripathi & Claypool (2001) found content-aware scaling could improve MPEG-1 coded video sequences and De Hesselle (2006) was not able to show that participants would

prefer IFD over TC in MPEG-2 coded video sequences. De Hesselle stated that this was content based. The current research showed that the amount of motion in a video sequence is an important predictor whether spatial or temporal coding achieves a higher quality.

Wijermans (2007), however, showed a preference for spatial scaling in all her video sequences. Although she argues that this might be due to a too low level of motion in the sequences, it is more likely due to a too high level of motion based on the current results. The level of motion might have been too high to notice that the quality of temporal scaling is perceived higher for a lower motion level (i.e. higher level of motion than where the two lines of IFD and TC cross; see Figure 14). This means that the situation is little more complex than Wijermans (2007) claims. Transcoding is not always better that I-Frame Delay coding; the present research showed that this depends on the amount of motion in a video sequence.

Perhaps the reason for the different conclusions lies in the motion measures that were used. Tripathi & Claypool (2001) used the number of macro blocks as an indicator of motion per frame in their research. Wijermans (2007) used the less precise differences (temporal and spatial information) between every 25th frame as a predictor of motion. The present research used the motion information of each frame to compute the overall motion indication of the stimuli. It was assumed that the current method was at least as precise as the methods that were used by Tripathi & Claypool (2001) and Wijermans (2007).

The present study assumed no difference between in-scene motion and camera motion but used only the overall motion measure. The overall measure was used because the algorithm for motion estimation used in this study did not allow for splitting different motion sources such as in-scene and camera motion as researched by Wijermans (2007). Zhong et al. (2004) showed that in-scene motion is expected to be dominant over camera motion. Wijermans (2007) used a different motion estimation algorithm than used in the current research which was able to split overall motion in camera and in-scene motion. However, it proved to be not accurate for all used stimuli as Wijermans (2007) discussed.

The present research showed that different levels of motion had an overall effect on perceived video quality. Participants showed a significant preference for temporally over spatially coded video sequences in the low level of motion condition. This was the reverse for video sequences the higher level of motion conditions where the participants showed a preference for spatially over temporally coded video sequences. This implies that both hypotheses need to be accepted.

Hypothesis H_I – *Temporal scaling provides better perceived video quality than spatial scaling for video sequences with low motion levels* – is accepted because the present study showed that temporally (I-Frame Delay) coded video sequences provides better perceived video quality for video sequences with a low level of motion. The second hypothesis H_{II} – *Spatial scaling provides a better perceived video quality than temporal scaling for video sequences with high motion levels* – is accepted because the present research showed that participants preferred video sequences that were coded using spatial scaling (Transcoding) for higher levels of motion.

Future research might focus on searching the exact threshold position (or interval) of the equilibrium between low and high level of motion video sequences. This threshold position will show us at which level of motion the quality for both IFD and TC versions of a video sequence are perceived as identical and might be defined even more precise by looking at different motion components such as in-scene and camera motion. Future research might also focus on the fact that results might be different when not only every second B-frame but all B-frames and perhaps even P- or I-frames were dropped too, due to a severe drop in bandwidth. In addition, the influence of the different GOP compositions can be investigated.

7. References

2.4 GHz frequency, (2008). Wikipedia – Electromagnetic interference at 2.4 GHz. http://en.wikipedia.org/wiki/2.4ghz_Interference Retrieved: 8-10-2008.

Alberta (2001). Perception of Visual Motion, *Department of Psychology, University of Alberta*. http://www.psych.ualberta.ca/~iwinship/vision/index.html Retrieved: 27-10-2008.

Apteker, R.T., Fisher, J.A., Kisimov, V.S. & Neishlos, H. (1995). Video acceptabilitiy and frame rate, *IEEE Multimedia*, 3(3), p. 32-40.

Blake, R. & Sekuler, R. (2006).*Perception* – Fifth Edition, International Edition. McGraw-Hill, 2006. ISBN: 007-111272-3.

Boyce, J.M. & Gaglianello, R.D. (1998). Packet Loss Effects on MPEG Video Sent Over the Public Internet, *Proceedings of ACM Multimedia*, pp. 181–190.

Boschman, M.C. (2001). Difscal: a tool for analyzing difference ratings on an ordinal category scale. *Behavior research methods, instruments and computers. 33*(1), p. 10-20.

Boyce, J.M., & Gaglianello, R.D. (1998). Packet Loss Effects on MPEG video sent over the public internet. *MULTIMEDIA '98: Proceedings of the sixth ACM international conference on Multimedia*, ACM, New York, NY, USA, p. 181—190.

Brouwers, C. J. J. (2006). A real-time SNR scalable transcoder for MPEG-2 video streams. Master thesis, *Eindhoven University of Technology* (2006).

Business Dictionary, (2008). Perceived quality, http://www.businessdictionary.com/definition/perceived-quality.html. Retrieved: 10-10-2008.

Cafforio, C. & Rocca, F (1976). Methods for measuring small displacements of television images, *IEEE transactions on information theory*, *22*, (5), p. 573-579.

Colour Blindness, (2008). Wikipedia – colour blindness. http://en.wikipedia.org/wiki/Colour_blindness, Retrieved: 10-10-2008.

Correia, P.L., & Pereira, F. (2003). Objective evaluation of video segmentation quality, *IEEE Transactions on Image Processing. 12*, (2), p. 186 – 200.

Darken, R., Kempster, K., & Peterson, B. (2001). Effects off streaming video quality of service on spatial comprehension in a reconnaissance task. *Proceedings of I/ITSEC*.

De Haan, G., Biezen, P.W.A.C., Huijgen, H. & Ojo, O.A. (1993). True-motion estimation with 3-D recursive search block matching, *IEEE Transactions on Circuits and Systems for Video Technology. 3*, (5), p. 368-379.

De Hesselle, H. (2006). Toolkit for Perception Experiments: Development, Implementation and Evaluation. *Stan Ackermans Instituut*, 2006.

DivX, (2008). DivX Video Codec. http://www.divx.com/ Retrieved: 6-10-2008.

Elecard Group (2008). Elecard StreamEye Studio, http://www.elecard.com/products/products-pc/consumer/streameye-tools/. Retrieved: 18-10-2008.

FFmpeg, (2008). http://ffmpeg.mplayerhq.hu/. Retrieved: 19-10-2008.

Fischer, W. (2004). Digital Video and Audio Broadcasting Technology: *A Practical Engineering Guide*. Second edition. Publisher: Springer – Verlag Berlin Heildeberger, 2004.

HDTV, (2008). Wikipedia – High Definition Television. http://en.wikipedia.org/wiki/High-definition_television. Retrieved: 7-10-2008.

Haakma, R., Jarnikov, D., & Van der Stok, P. (2005). Perceived quality of wirelessly transported videos. *Philips Research Laboratories, Eindhoven, The Netherlands*. In press.

Hammett, S.T., Smith, A.T. (1992). Two temporal channels or three? A re-evaluation, *Vision Research*. *32*, 2, p. 285-291.

Hauske, G., Stockhammer, T. & Hofmaier, R. (2003). Subjective Image Quality of Low-rate and Low-Resolution Video Sequences, *Proceedings International Workshop on Mobile Multimedia Communication*, Munich.

Hesseler, W. & Eickeler, S. (2005). MPEG-2 compressed-domain algorithms for video analysis, *EURASIP Journal on Applied Signal Processing*, Colume 2006, Article ID 56940, p. 1-11.

Hunt, D. M., Wilkie, S. E., Bowmaker, J. K., & Poopalasundaram, S. (2001). Vision in the ultraviolet. *CMLS Cellular and Molecular Life Sciences*. 58, p. 1583–1598.

IBM (2008), Functions of mobile multimedia QoS control (Version 28 February 2002). http://www.trl.ibm.com/projects/mmqos/system_e.htm. Retrieved: 1-11-2008.

Ishihara, S. (1960). *Tests for colour-blindness*. Fifteenth complete edition. Publisher: Kanehara Shuppan Co., Ltd. Tokyo, Japan. London: H.K Lewis & Co., Ltd.

ISM Band, (2008). Wikipedia – ISM band. http://en.wikipedia.org/wiki/ISM_band Retrieved: 8-10-2008.

Ismaeil, I. Docef, A., Kossentini, F., & Ward, R. (1999). Efficient motion estimation using spatial and temporal motion vector prediction. *International Conference on Image Processing ICIP 99, Proceedings*,*1*, pp.70-74.

ITU-R BT.500-11 (2002). *Methodology for the subjective assessment of the quality of television pictures*. ITU - R Recommendation BT.500-11.

Jarnikov, D. (2007). *QoS framework for video streaming in home networks*. PhD Thesis. Eindhoven, University Press, The Netherlands.

Kingsbury, N. (2005, June 8). *Human Vision*. Retrieved from the Connexions Web site: http://cnx.org/content/m11084/2.5/. Retrieved: 6-10-2008.

Koivunen, T. & Salonen, J. (1994). Motion estimation using combined shape and edge matching, *Signal processing: image communication*, 6, (3), p. 241-252.

Kozlov, S., Van der Stok, P., Lukkien, J., (2005). Adaptive scheduling of MPEG video frames during real-time wireless video streaming. *Sixth IEEE International Symposium on a World of Wireless Mobile and Multimedia Networks, 2005*.

Li, W. & Salari, E. (1995). Successive Elimination Algorithm for motion estimation, *IEEE transactions on image processing*, 4, (1), p. 105-107.

Masry, M. & Hemami, S. S. (2001). An analysis of subjective quality in low bit-rate video. *Proceedings in International Conference on Image Processing.* ICIP, 2001, pp. 465–468.

McCarthy, J.D., Sasse, M.A. & Miras, D. (2004). Sharp or Smooth? Comparing the effects of quantization vs. frame rate for streamed video, *Conference on Human Factors in Computing Systems – Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 535 - 542.

MCLAB Seminar Series, (2008). Introduction to Video Transcoding. http://w7ww.mcl.ie.cuhk.edu.hk/internal/intro_to_transcoding.ppt. Retrieved: 10-10-2008.

Meesters, L.M.J. (2002), *Predicted and perceived quality of bit-reduced gray-scale still images*. PhD thesis, Eindhoven University of Technology.

Meng, J., Juan, Y., Chang, S-F. (1995). Scene detection in a MPEG compressed video sequence. *IS&T/SPIE Syposium Proceedings*, 2419, February 1995, San Jose, California, pp. 1-5.

Mitchell, J.L., Pennebaker, W.B., Fogg, C.E., & LeGall, D.J. (1996). MPEG video compression standard. *Chapman & Hall*, NY, USA.

Neumeyer, C. (1992). Tetrachromatic colour vision in goldfish: evidence from colour mixture experiments. *Journal of Comparative Physiology A, 171,*(5), p. 639-649*.*

O'Shea, R.P. (1991). Thumb's rule tested: visual angle of thumb's width is about 2 deg. *Perception 20*, (3), 415-418.

Palmer, S.E. (1999). *Vision Science: Photons to Phenomenology*. Cambridge, MIT Press.

Pan, W.D., Yoo, S.M., Nalasani, M. & Cox, P.G. (2007). Efficient local transformation estimation using Lie operators, *Information Sciences. 177*, (3), pp. 815-831.

Ramachandran, S. & Srinivasan, S. (2001). FPGA implementation of a novel, fast motion estimation algorithm for real-time video compression, *Proceedings of the 2001 ACM/SIGDA Ninth international Symposium on Field Programmable Gate Arrays* (Monterey, CA, United States). FPGA '01, pp. 213-219.

Run-length encoding, (2008). Wikipedia – Run length coding. http://en.wikipedia.org/wiki/Run-length_encoding. Retrieved: 10-10-2008.

SDTV, (2008). Wikipedia – Standard Definition Television. http://en.wikipedia.org/wiki/SDTV. Retrieved: 10-10-2008.

Sekuler, R., Watamaniuk, S.N.J., and Blake, R. (2002). Perception of visual motion, *Chapter prepared for Stevens' Handbook of Experimental Psychology*, New York: J. Wiley Publishers.

Seyler, A. J., & Budrikis, Z. L., (1965). Detail perception after scene changes in television image presentations. IEEE Transactions on Information Theory *11*,(1), pp. 31–43.

Simerly, T. (2008). Introduction to video transcoding for CE, texas instruments. http://www.embedded.com/design/206905102. Retrieved: 10-10-2008.

Simons, D.J. & Chabris, C.F. (1999). Gorillas in our midst: sustained inattentional blindness for dynamic events, *Department of Psychology, Harvard University*. Perception, *28*, pp. 1059-1074.

Tanenbaum, A.S. (2005). *Computernetwerken* – 4e editie, Vrije Universiteit Amsterdam. Pearson Education Benelux, ISBN: 90-430-0698-X.

Torgerson, W.S. (1958). *Theory and methods of scaling*. London: Wiley.

Tripathi, A. & Claypool, M. (2001). Adaptive content-aware scaling for improved video streaming. in *Proc. 2nd International Workshop on Intelligent Multimedia Computing and Networking (IMMCN), 2002*.

Van den Branden Lambrecht, C.J. (1996). A working spatio-temporal model of the human visual system for image restoration and quality assessment applications, *IEEE International Conference on Acoustics, Speech, and Signal Processing - ICASSP-96, Conference Proceedin*gs. *4*, pp. 2291-2294.

Van den Branden Lambrecht, C. J. & Verscheure, O. (1996). Perceptual quality measure using a spatio-temporal model of the human visual system, *Proceedings of Society of Photo-optical Instrumentation Engineers*. 2668, pp. 450–461.

Van den Ende, N., De Hesselle, H. & Meesters, L.M.J. (2007). Towards content-aware coding: user study. *EuroITV, LNCS 4471*, pp. 185-194.

Video, (2008). Wikipedia – Video. http://en.wikipedia.org/wiki/Video. Retrieved: 26-09-2008.

VirtualDub, (2008). Virtualdub.org. http://www.virtualdub.org/ Retrieved: 19-10-2008.

VLC (2008), Wikipedia – Variable-length code. http://en.wikipedia.org/wiki/Variable-length_code. Retrieved: 8-10-2008.

Wallace, G.K. (1992). The JPEG still picture compression standard, *IEEE Transactions on Consumer Electronics*, *38*, (1), pp. xviii-xxxiv.

Wallach, H., O'Connell, D. N. & Neisser, U. (1953), The memory effect of visual perception of three-dimensional form. *Journal of Experimental Psychology. 45*, (5), pp. 360-368.

Wang, Z., Bovik, A.C., Sheikh, H.R. & Simoncelli, E.P. (2004). Image Quality Assessment: From Error Measurement to Structural Similarity, IEEE transactions on image processing, 13, (1), pp. 1-14.

Wang, Z., Sheikh, H.R. &. Bovik, A.C. (2003). Objective video quality assessment, *Chapter 41 in The Handbook of Video Databases: Design and Applications by Furht, B & Marqure*, O.. CRC Press, pp. 1041-1078.

Watson, A.B. & Ahumada, A.J. Jr (1985). Model of human visual-motion sensing. *Journal of Optical Society of America. 2* (2), pp. 322-342.

Wenger, S. (1998). Temporal Scalability using P-Pictures for Low-Latency Applications, IEEE Signal Processing Society 1998 Workshop on Multimedia Signal Processing. December 7-9, 1998, Los Angeles, California, USA.

Wertheim, A.H. (1985). De wondere waarneming van beweging, *Koninklijke Maatschappij voor Natuurkunde Diligentia, Natuurkundige voordrachten N.R. 63*.

Wijermans, C.G.A. (2007). *Influence of in-scene motion and camera motion on perceived video quality in MPEG-2 adaptive coding*. Master Thesis Technical University of Eindhoven, 2007.

Winkler, S. (1999). Issues in vision modeling for perceptual video quality assessment, *Signal Processing. 78*, (2), pp. 231-252.

Wolf, S. & Webster, A. (1997). *Subjective and objective measures of scene criticality*. Turin ITU Experts meeting.

Yuen, M. & Wu, H.R. (1998). A survey of hybrid MC/DPCM/DCT video coding distortions, *Signal Processing. 70* (3), pp. 247-278.

Zhong, Y., Richardson, I., Sahraie, A. & McGeorge, P. (2004). Influence of task and scene content on subjective video quality. *Lecture Notes in Computer Science*, *3211*/2004, p. 295.

Appendix A – Television

Cineos Flat TV 52" LCD with integrated digital tuner 52pfl9632d/10
Below more information is provided (in Dutch) which can also be found at
http://www.philips.com when searched for "52pfl9632d/10".

Beeldschermresolutie	
Resolutie	1920 x 1080 pixels

Beeldscherm	
Beeldschermdiagonaal	52 "
HD-Ready	✔
SmartContrast-verhouding (dynamisch)	5000:1
Beeldformaat	16:9
Beeldscherm	LCD Full HD W-UXGA Act. Matrix
Contrastverhouding	2000:1
Progressive scan	✔
Helderheid	550 cd/m²
Responstijd	3 ms
Kijkhoek, horizontaal	176 °
Kijkhoek, vertikaal	176 °

Gewicht en omvang	
Breedte	1310 mm
Diepte	125 mm
Hoogte	824 mm
Gewicht met voet	48800 g
Afmetingen (B x D x H) met voet	1310 x 297 x 889 mm
Gewicht	45400 g

Eisen aan de omgeving	
Temperatuur, in bedrijf	5 - 35 °C

Appendix B – Demographic fill-in form and instructions

Participant ID: _____

Name:

Age:

Experience in video quality?   ☐ no   ☐ yes (please elaborate below)

Gender: ☐ male   ☐ female

Visual acuity: __ / __

Colour blindness:

Profession

**Session:**   Date   stimulus set:

Instructions
This study is about the differences in perceived video quality. In this session 18 pairs of short videos will be shown to you. Each pair consists of two videos and each pair is shown twice. Each video will take 10 seconds. Before a video starts, a grey field is shown with a number identifying the sequence number followed by a letter indicating whether the first video, 'A', or the second video, 'B', of the video pair is shown. The pairs of videos are shown twice. **The first presentation of a pair is used to get familiar with the videos, during the second presentation you rate video 'B' compared to video 'A'.** The order for each pair is illustrated in the figure 1 below.
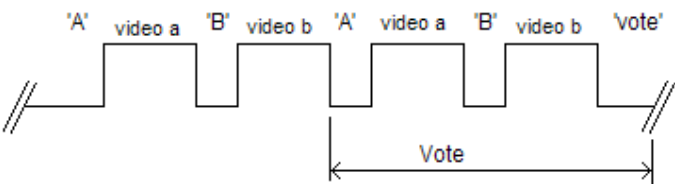
Figure 1: Phases of presentation of test material

Your task is to rate the second video compared to the first video using the scale below. (This means rating 'B' compared to 'A')

Rating scale

Comparison X

B is much worse than A	-3	-2	-1	0	+1	+2	+3	B is much better than A

After the pairs are shown to you for the second time, you will have 5 seconds after the last movie has ended to rate the 'B' video compared to the 'A' video. You can rate the video circling the number on the paper before you, which indicates whether the video quality of video 'B' is better, the same, or worse than video 'A'.

Before the experiment starts, a short trial session is shown. In this trial you have to assess 6 video pairs to get familiar with the footage and the various video quality levels. The data resulting from this trial session will not be used.

The experiment consists of two parts with a break in between. After 18 comparisons (approximately 18 minutes) a short break will be held. After this, the second half of the experiment is continued. The whole experiment will take less than 45 minutes.

Your data will be processed anonymously.

Thank you for participating!

Practice

Please circle the number of your choice for each comparison

Test 1

B is much worse than A	-3	-2	-1	0	+1	+2	+3	B is much better than A

Test 2

B is much worse than A	-3	-2	-1	0	+1	+2	+3	B is much better than A

Test 3

B is much worse than A	-3	-2	-1	0	+1	+2	+3	B is much better than A

Test 4

B is much worse than A	-3	-2	-1	0	+1	+2	+3	B is much better than A

Test 5

B is much worse than A	-3	-2	-1	0	+1	+2	+3	B is much better than A

Test 6

B is much worse than A	-3	-2	-1	0	+1	+2	+3	B is much better than A

If there are any questions please ask them now!

The experiment starts after the next page…

Appendix C – Original movies

Gladiator (2000)

(Source: IMDb, http://www.imdb.com/title/tt0172495/, retrieved 18-07-2009)

Cast:

Russell Crowe	as	Maximus
Joaquin Phoenix	as	Commodus
Connie Nielsen	as	Lucilla
Oliver Reed	as	Proximo
Richard Harris	as	Marcus Aurelius

User Rating:

8.3/10

Director:

Ridley Scott

Writers:

David Franzoni (story)
David Franzoni (screenplay)

Release Date:

18 May 2000 (Netherlands)

Genre:

Action | Adventure | Drama | History

Plot:

Maximus is a powerful Roman general, loved by the people and the aging Emperor, Marcus Aurelius. Before his death, the Emperor chooses Maximus to be his heir over his own son, Commodus, and a power struggle leaves Maximus and his family condemned to death. The powerful general is unable to save his family, and his loss of will allows him to get captured and put into the Gladiator games until he dies. The only desire that fuels him now is the chance to rise to the top so that he will be able to look into the eyes of the man who will feel his revenge.

Awards:

Won 5 Oscars. Another 45 wins & 79 nominations

User Comments:

A nail biting, intense modern epic!

The Lost World: Jurassic Park (1997)

(Source: IMDb, http://www.imdb.com/title/tt0119567/, retrieved 18-07-2009)

Cast:

Jeff Goldblum	as	Dr. Ian Malcolm
Julianne Moore	as	Dr. Sarah Harding
Pete Postlethwaite	as	Roland Tembo
Richard Attenborough	as	John Hammond
Vince Vaughn	as	Nick Van Owen

User Rating:

5.9/10

Director:

Steven Spielberg

Writers:

Michael Crichton (novel)
David Koepp (screenplay)

Release Date:

25 September 1997 (Netherlands)

Genre:

Action | Adventure | Sci-Fi | Thriller

Plot:

After a small girl is attacked by a small group of compsognathus, Ian Malcolm discovers that there is a second island full of a variety of dinosaurs. Dr. John Hammond decides to send four adventure to monitor the dinosaur's lifestyle before INGEN move forward in controlling the island. Ian Malcolm doesn't like the idea and wants to contact the other three members, but before he can contact them, he finds out that his girlfriend, Sarah Harding is already on the island. Now, what was supposed to be a natural viewing of the incredible creatures in their habitats, has turned into a rescue mission with everyones life at danger.

Awards:

Nominated for Oscar. Another 3 wins & 15 nominations

User Comments:

Another Steven Spielberg Classic Comes To Life

High Crimes (2002)

(Source: IMDb, http://www.imdb.com/title/tt0257756/, retrieved 18-07-2009)

Cast:

	Ashley Judd	as	Claire Kubik
	Morgan Freeman	as	Charles W. Grimes
	James Caviezel	as	Tom Kubik / Sgt. Ron Chapman (as Jim Caviezel)
	Adam Scott	as	Lt. Terrence Embry
	Amanda Peet	as	Jackie

User Rating:

6.1/10

Director:

Carl Franklin

Writers:

Joseph Finder (novel)
Yuri Zeltser (screenplay)

Release Date:

22 August 2002 (Netherlands)

Genre:

Crime | Thriller | Drama

Plot:

High powered lawyer Claire Kubik finds her world turned upside down when her husband, who she thought was Tom Kubik, is arrested and is revealed to be Ron Chapman. Chapman is on trial for a murder of Latin American villagers while he was in the Marines. Claire soon learns that to navigate the military justice system, she'll need help from the somewhat unconventional Charlie Grimes; meanwhile, Claire's sister, Jackie, is falling in love with wet-behind-the-ears Lieutenant Embry assigned as the official defense lawyer. And most of the eyewitnesses have rather too conveniently died.

Awards:

1 nomination

User Comments:

Formulaic story saved by strong performances