Chapter 10

VISION AND VIDEO: MODELS AND APPLICATIONS

Stefan Winkler
Swiss Federal Institute of Technology – EPFL
Signal Processing Laboratory
1015 Lausanne, Switzerland
Stefan.Winkler@epfl.ch

Christian J. van den Branden Lambrecht
EMC Media Group
80 South Street
Hopkinton, MA 01748, USA
vdb@emc.com

Murat Kunt
Swiss Federal Institute of Technology – EPFL
Signal Processing Laboratory
1015 Lausanne, Switzerland
Murat.Kunt@epfl.ch

1. INTRODUCTION

While traditional analog systems still form the vast majority of television sets today, production studios, broadcasters and network providers have been installing digital video equipment at an ever-increasing rate. The border line between analog and digital video is moving closer and closer to the consumer. Digital satellite and cable service have been available for a while, and recently terrestrial digital television broadcast has been introduced in a number of locations around the world.

Analog video systems, which have been around for more than half a century now, are among the most successful technical inventions measured by their market penetration (more than 1 billion TV receivers in the world) and the time span of their widespread use. However, because of the closed-system approach inherent to analog technology, any new functionality or processing is utterly difficult to incorporate in the existing systems. The introduction of digital video systems has given engineers additional degrees of freedom due to the flexibility of digital information processing and the ever-decreasing cost of computing power. Reducing the bandwidth and storage requirements while maintaining a quality superior to that of analog video has been the priority in the design of these new systems.

Many optimizations and improvements of video processing methods have relied on purely mathematical measures of optimality, such as mean squared error (MSE) or signal-to-noise ratio (SNR). However, these simple measures operate solely on a pixel-by-pixel basis and neglect the important influence of image content and viewing conditions on the actual visibility of artifacts. Therefore, their predictions often do not agree well with visual perception.

In the attempt to increase compression ratios for video coding even further, engineers have turned to vision science in order to better exploit the limitations of the human visual system. As a matter of fact, there is a wide range of applications for vision models in the domain of digital video, some of which we outline in this chapter. However, the human visual system is extremely complex, and many of its properties are still not well understood. While certain aspects have already found their way into video systems design, and while even ad-hoc solutions based on educated guesses can provide satisfying results to a certain extent, significant advancements of the current state of the art will require an in-depth understanding of human vision.

Since a detailed treatment of spatial vision can be found in other chapters of this book, our emphasis here is on temporal aspects of vision and modeling, which is the topic of Section 2. Then we take a look at the basic concepts of video coding in Section 3. An overview of spatio-temporal vision modeling, including a perceptual distortion metric developed by the authors, is given in Section 4. We conclude the chapter by applying vision models to a number of typical video test and measurement tasks in Section 5.

2. MOTION PERCEPTION

Motion perception is a fundamental aspect of vision and aids us in many essential visual tasks: it facilitates depth perception, object discrimination, gaze direction, and the estimation of object displacement. Motion, particularly in the peripheral visual field, attracts our attention.

There are many controversial opinions about motion perception. Motion has often been closely linked to the notion of optical flow, particularly in the work on motion prediction for video coding. Sometimes, however, motion can be perceived in stimuli that do not contain any actual movement, which is referred to as apparent motion. In light of these concepts, motion is better defined as a psychological sensation, a visual inference, similar to color perception. The images on the retina are just time-varying patterns of light; the evolution of these light distributions over time is then interpreted by the visual system to create a perception of objects moving in a three-dimensional world.

Extending spatial models for still images to handle moving pictures calls for a close examination of the way temporally varying visual information is processed in the human brain [73]. The design of spatio-temporal vision models (cf. Section 4.) is complicated by the fact that much less attention of vision research has been devoted to temporal aspects than to spatial aspects. In this section, we take a closer look at the perception of motion and the temporal mechanisms of the human visual system, in particular the temporal and spatio-temporal contrast sensitivity functions, temporal masking, and pattern adaptation.

2.1 TEMPORAL MECHANISMS

Early models of spatial vision were based on the single-channel assumption, i.e. the entire input is processed together and in the same way. Due to their inability to model signal interactions, however, single-channel models are unable to cope with more complex patterns and cannot explain data from experiments on masking and pattern adaptation. This led to the development of multi-channel models, which employ a bank of filters tuned to different frequencies and orientations. Studies of the visual cortex have shown that many of its neurons actually exhibit receptive fields with such tuning characteristics [14]; serving as an oriented band-pass filter, the neuron responds to a certain range of spatial frequencies and orientations.

Temporal mechanisms have been studied by vision researchers for many years, but there is less agreement about their characteristics than those of spatial mechanisms. It is believed that there are one temporal low-pass and one, possibly two, temporal band-pass mechanisms [19, 27, 39, 64], which are generally referred to as sustained and transient channels, respectively. Physiological experiments confirm these results to the extent that low-pass and band-pass mechanisms have been found [17]. However, neurons with band-pass properties exhibit a wide range of peak frequencies. Recent results also indicate that the peak frequency and bandwidth of the mechanisms change considerably with stimulus energy [18]. The existence of an actual third mechanism is questionable, though [19, 24].

In a recent study [19], for example, temporal mechanisms are modeled with a two-parameter function and its derivatives. It is possible to achieve a very good fit to a large set of psychophysical data using only this function and its second derivative, corresponding to one sustained and one transient mechanism, respectively. The frequency responses of the corresponding filters for a typical choice of parameters are used and shown later in Section 4.2.2.

2.2 CONTRAST SENSITIVITY

The response of the human visual system to a stimulus depends much less on the absolute luminance than on the relation of its local variations to the surrounding luminance. This property is known as *Weber's law*, and contrast is a measure of this relative variation of luminance. While Weber's law is only an approximation of the actual sensory perception, contrast measures based on this concept are widely used in vision science. Unfortunately, a common definition of contrast suitable for all situations does not exist, not even for simple stimuli.

Mathematically, Weber contrast can be expressed as $C = \Delta L/L$. In vision experiments, this definition is used mainly for patterns consisting of an increment or decrement ΔL to an otherwise uniform background luminance L.

However, such a simple definition is inappropriate for measuring contrast in complex images, because a few very bright or very dark points would determine the contrast of the entire image. Furthermore, human contrast sensitivity varies with the adaptation level associated with the local average luminance. Local band-limited contrast measures have been introduced to address these issues [41, 42, 76] and have been used successfully in a number of vision models [12, 37].

Our sensitivity to contrast depends on the color as well as the spatial and temporal frequency of the stimuli. Contrast sensitivity functions (CSF's) are generally used to quantify these dependencies. Contrast sensitivity is defined as the inverse of the contrast threshold, i.e. the minimum contrast necessary for an observer to detect a stimulus.

Spatio-temporal CSF approximations are shown in Figure 10.1. Achromatic contrast sensitivity is generally higher than chromatic, especially for high spatio-temporal frequencies. The full range of colors is perceived only at low frequencies. As spatio-temporal frequencies increase, sensitivity to blue-yellow stimuli declines first. At even higher frequencies, sensitivity to red-green stimuli diminishes as well, and perception becomes achromatic. On the other hand, achromatic sensitivity decreases slightly at low spatio-temporal frequencies, whereas chromatic sensitivity does not (see Figure 10.1). However, this apparent attenuation of sensitivity towards low frequencies may be attributed to implicit masking, i.e. masking by the spectrum of the window within which the test gratings are presented [78].

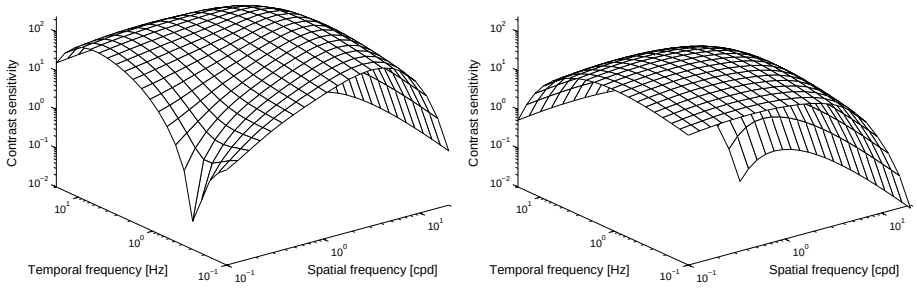

Figure 10.1 Approximations of achromatic (left) and chromatic (right) spatio-temporal contrast sensitivity functions [6, 32, 33].

There has been some debate about the space-time separability of the spatio-temporal CSF. This property is of interest in vision modeling because a CSF that could be expressed as a product of spatial and temporal components would simplify modeling. Early studies concluded that the spatio-temporal CSF was not space-time separable at lower frequencies [34, 47]. Kelly [31] measured contrast sensitivity under stabilized conditions (i.e. the stimuli were stabilized on the retina by compensating for the observers' eye movements) and fit an analytic function to these measurements [32], which yields a very close approximation of the spatio-temporal CSF for counter-phase flicker. It was found that this CSF and its chromatic counterparts can also be approximated by linear combinations of two space-time separable components termed excitatory and inhibitory CSF's [6, 33].

Measurements of the spatio-temporal CSF for both in-phase and conventional counter-phase modulation suggest that the underlying filters are indeed spatio-temporally separable and have the shape of low-pass exponentials [77]. The spatio-temporal interactions observed for counter-phase modulation can be explained as a product of masking by the zero-frequency component of the gratings.

The important issue of unconstrained eye movements in CSF models is addressed in Chapter **??**. Natural drift, smooth pursuit and saccadic eye movements can be included in Kelly's formulation of the stabilized spatio-temporal CSF using a model for eye velocity [13]. In a similar manner, motion compensation of the CSF can be achieved by estimating smooth-pursuit eye movements under the worst-case assumption that the observer is capable of tracking all objects in the scene [70].

2.3 TEMPORAL MASKING

Masking is a very important phenomenon in perception as it describes interactions between stimuli (cf. Chapter **??**). Masking occurs when a stimulus that is visible by itself cannot be detected due to the presence of another. Sometimes the opposite effect, facilitation, occurs: a stimulus that is not visible by itself can be detected due to the presence of another. Within the framework of imaging and video applications it is helpful to think of the distortion or coding noise being masked (or facilitated) by the original image or sequence acting as background. Masking explains why similar coding artifacts are disturbing in certain regions of an image while they are hardly noticeable elsewhere.

Masking is strongest between stimuli located in the same perceptual channel, and many vision models are limited to this intra-channel masking. However, psychophysical experiments show that masking also occurs between channels of different orientations [16], between channels of different spatial frequency, and between chrominance and luminance channels [8, 36, 56], albeit to a lesser extent.

Temporal masking is an elevation of visibility thresholds due to temporal discontinuities in intensity, e.g. scene cuts. Within the framework of television, it was first studied by Seyler and Budrikis [52, 53], who concluded that threshold elevation may last up to a few hundred milliseconds after a transition from dark to bright or from bright to dark. In a more recent study on the visibility of MPEG-2 coding artifacts after a scene cut, significant visual masking effects were found only in the first subsequent frame [57]. A strong dependence on stimulus polarity has also been noticed [7]: The masking effect is much more pronounced when target and masker match in polarity, and it is greatest for local spatial configurations. Similar to to the case of spatial stimulus interactions, the opposite of temporal masking, temporal facilitation, has been observed at low-contrast discontinuities.

Interestingly, temporal masking can occur not only after a discontinuity ("forward masking"), but also before. This "backward masking" may be explained as the result of the variation in the latency of the neural signals in the visual system as a function of their intensity [1].

So far, the above-mentioned temporal masking effects have received much less attention in the video coding community than their spatial counterparts. In principle, temporal masking can be taken into account with a contrast gain control model (cf. Section 4.2.3), as demonstrated in [21]. A video quality metric that incorporates forward masking effects by means of a low-pass filtered masking sequence is described in [66].

2.4 ADAPTATION

Pattern adaptation in the human visual system is the adjustment of contrast sensitivity in response to the prevailing stimulation patterns. For example, adaptation to patterns of a certain frequency can lead to a noticeable decrease of contrast sensitivity around this frequency [22, 55, 71]. Together with masking, adaptation was one of the major incentives for developing a multi-channel theory of vision. However, pattern adaptation has a distinct temporal component to it and is not automatically taken into account by a multi-channel representation of the input; it needs to be incorporated explicitly by adapting the pertinent model parameters. A single-mechanism model that accounts for both pattern adaptation and masking effects of simple stimuli was presented in [49], for example.

An interesting study in this respect used natural images of outdoor scenes (both distant views and close-ups) as adapting stimuli [68]. It was found that exposure to such stimuli induces pronounced changes in contrast sensitivity. The effects can be characterized by selective losses in sensitivity at lower to medium spatial frequencies. This is consistent with the characteristic amplitude spectra of natural images, which decrease with frequency roughly as $1/f$.

Likewise, an examination of how color sensitivity and appearance might be influenced by adaptation to the color distributions of images [69] revealed that natural scenes exhibit a limited range of chromatic distributions, hence the range of adaptation states is normally limited as well. However, the variability is large enough so that different adaptation effects may occur for individual scenes and for different viewing conditions.

3. VIDEO CONCEPTS

3.1 STANDARDS

The Moving Picture Experts Group (MPEG)[1] is a working group of ISO/IEC in charge of the development of international standards for compression, decompression, processing, and coded representation of moving pictures, audio and their combination. MPEG comprises some of the most popular and widespread standards for video coding. The group was established in January 1988, and since then it has produced:

- MPEG-1, a standard for storage and retrieval of moving pictures and audio, which was approved in November 1992. MPEG-1 is intended to be generic, i.e. only the coding syntax is defined and therefore mainly the decoding scheme is standardized. MPEG-1 defines a block-based hybrid

[1] See http://drogo.cselt.stet.it/mpeg/ for an overview of its activities.

DCT/DPCM coding scheme with prediction and motion compensation. It also provides functionality for random access in digital storage media.

- MPEG-2, a standard for digital television, which was approved in November 1994. The video coding scheme used in MPEG-2 is again generic; it is a refinement of the one in MPEG-1. Special consideration is given to interlaced sources. Furthermore, many functionalities such as scalability were introduced. In order to keep implementation complexity low for products not requiring all video formats supported by the standard, so-called "Profiles", describing functionalities, and "Levels", describing resolutions, were defined to provide separate MPEG-2 conformance levels.

- MPEG-4, a standard for multimedia applications, whose first version was approved in October 1998. MPEG-4 addresses the need for robustness in error-prone environments, interactive functionality for content-based access and manipulation, and a high compression efficiency at very low bitrates. MPEG-4 achieves these goals by means of an object-oriented coding scheme using so-called "audio-visual objects", for example a fixed background, the picture of a person in front of that background, the voice associated with that person etc. The basic video coding structure supports shape coding, motion compensation, DCT-based texture coding as well as a zerotree wavelet algorithm.

- MPEG-7, a standard for content representation in the context of audio-visual information indexing, search and retrieval, which is scheduled for approval in late 2001.

The standards being used commercially today are mainly MPEG-1 (in older compact discs), MPEG-2 (for digital TV and DVD's), and H.261/H.263 (which use related compression methods for low-bitrate communications). Some broadcasting companies in the US and in Europe have already started broadcasting television programs that are MPEG-2 compressed, and DVD's are rapidly gaining in popularity in the home video sector. For further information on these and other compression standards, the interested reader is referred to [4].

3.2 COLOR CODING

Many standards, such as PAL, NTSC, MPEG, or JPEG, are already based on human vision in the way color information is processed. In particular, they take into account the nonlinear perception of lightness, the organization of color channels, and the low chromatic acuity of the human visual system.

Conventional television cathode ray tube (CRT) displays have a nonlinear, roughly exponential relationship between frame buffer RGB values or signal

voltage and displayed intensity. In order to compensate for this, *gamma correction* is applied to the intensity values before coding. It so happens that the human visual system has an approximately logarithmic response to intensity, which is very nearly the inverse of the CRT nonlinearity [45]. Therefore, coding visual information in the gamma-corrected domain not only compensates for CRT behavior, but is also more meaningful perceptually.

Furthermore, it has been long known that some pairs of hues can coexist in a single color sensation, while others cannot. This led to the conclusion that the sensations of red and green as well as blue and yellow are encoded in separate visual pathways, which is commonly referred to as the *theory of opponent colors* (cf. Chapter **??**). It states that the human visual system decorrelates its input into black-white, red-green and blue-yellow difference signals.

As pointed out before in Section 2.2, chromatic visual acuity is significantly lower than achromatic acuity. In order to take advantage of this behavior, the color primaries red, green, and blue are rarely used for coding directly. Instead, *color difference* (chroma) signals similar to the ones just mentioned are computed. In component digital video, for example, the resulting color space is referred to as $\mathrm{Y'C'_BC'_R}$, where $\mathrm{Y'}$ encodes luminance, $\mathrm{C'_B}$ the difference between blue primary and luminance, and $\mathrm{C'_R}$ the difference between red primary and luminance (the primes are used here to emphasize the nonlinear nature of these quantities due to the above-mentioned gamma correction).

The low chromatic acuity now permits a significant data reduction of the color difference signals, which is referred to as *chroma subsampling*. The notation commonly used is as follows:

- 4:4:4 denotes no chroma subsampling.
- 4:2:2 denotes chroma subsampling by a factor of 2 horizontally; this sampling format is used in the standard for studio-quality component digital video as defined by ITU-R Rec. 601 [29], for example.
- 4:2:0 denotes chroma subsampling by a factor of 2 both horizontally and vertically; this sampling format is often used in JPEG or MPEG and is probably the closest approximation of actual visual color acuity achievable by chroma subsampling alone.
- 4:1:1 denotes chroma subsampling by a factor of 4 horizontally.

3.3 INTERLACING

As analog television was developed, it was noted that flicker could be perceived at certain frame rates, and that the magnitude of the flicker was a function of screen brightness and surrounding lighting conditions. In a movie theater at relatively low light levels, a motion picture can be displayed at a frame rate

of 24 Hz, whereas a bright CRT display requires a refresh rate of more than 50 Hz for flicker to disappear. The drawback of such a high frame rate is the high bandwidth of the signal. On the other hand, the spatial resolution of the visual system decreases significantly at such temporal frequencies (cf. Figure 10.1). These two properties combined gave rise to a technique referred to as *interlacing*.

The concept of interlacing is illustrated in Figure 10.2. Interlacing trades off vertical resolution with temporal resolution. Instead of sampling the video signal at 25 or 30 frames per second, the sequence is shot at a frequency of 50 or 60 interleaved fields per second. A field corresponds to either the odd or the even lines of a frame, which are sampled at different time instants and displayed alternately (the field containing the even lines is referred to as the top field, and the field containing the odd lines as the bottom field). Thus the required bandwidth of the signal can be reduced by a factor of 2, while the full horizontal and vertical resolution is maintained for stationary image regions, and the refresh rate for objects larger than one scanline is still sufficiently high.

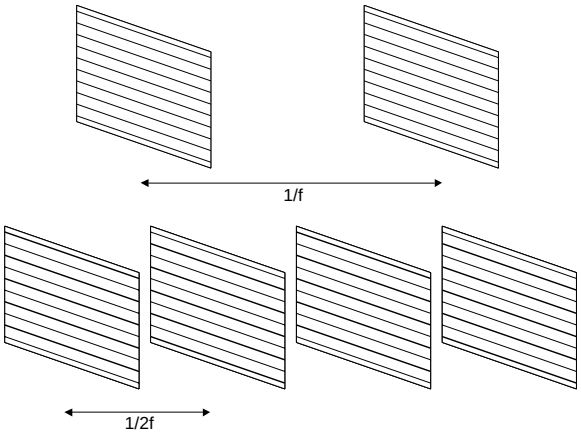

Figure 10.2 Illustration of interlacing. The top sequence is progressive; all lines of each frame are transmitted at the frame rate f. The bottom sequence is interlaced; each frame is split in two fields containing the odd and the even lines (shown in bold), respectively. These fields are transmitted alternately at twice the original frame rate.

MPEG-1 handles only progressive video, which is better adapted to computer displays. MPEG-2 on the other hand was designed as the new standard to transmit television signals. Therefore it was decided that MPEG-2 would support both interlaced and progressive video. An MPEG-2 bitstream can contain a progressive sequence encoded as a succession of frames, an interlaced sequence encoded as a succession of fields, or an interlaced sequence encoded as a succession of frames. In the latter case, each frame contains a top and a bottom field, which do not belong to the same time instant. Based on this, a variety of modes and combinations of motion prediction algorithms were defined in MPEG-2.

Interlacing poses quite a problem in terms of vision modeling, especially from the point of view of temporal filtering. It is not only an implementation

problem, but also a modeling problem, because identifying the signal that is actually perceived is not obvious. Vision models have often overlooked this issue and have taken simplistic approaches; most of them have restricted themselves to progressive input. Newer models incorporate de-interlacing approaches, which aim at creating a progressive video signal that has the spatial resolution of a frame and the temporal frequency of a field. A simple solution, which is still very close to the actual signal perceived by the human eye, consists in merging consecutive fields together into a full-resolution 50 or 60 Hz signal. This is a valid approach as each field is actually displayed for two field periods due to the properties of the CRT phosphors. Other solutions interpolate both spatially and temporally by upsampling the fields. Although the latter might seem more elegant, it feeds into the vision model a signal which is not the one that is being displayed. Reviews of various de-interlacing techniques can be found in [15, 59].

3.4 ARTIFACTS

The fidelity of compressed and transmitted video sequences is affected by the following factors:

- any pre- or post-processing of the sequence outside of the compression module. This can include chroma subsampling and de-interlacing, which were discussed briefly above, or frame rate conversion. One particular example is 3:2 pulldown, which is the standard way to convert progressive film sequences shot at 24 frames per second to interlaced video at 30 frames per second.
- the compression operation itself.
- the transmission of the bitstream over a noisy channel.

3.4.1 Compression Artifacts. The compression algorithms used in various video coding standards today are very similar to each other. Most of them rely on block-based DCT with motion compensation and subsequent quantization of the DCT coefficients. In such coding schemes, compression distortions are caused by only one operation, namely the quantization of the DCT coefficients. Although other factors affect the visual quality of the stream, such as motion prediction or decoding buffer, these do not introduce any distortion per se, but affect encoding process indirectly by influencing the quantization scale factor.

A variety of artifacts can be distinguished in a compressed video sequence:

- *blockiness* or blocking effect, which refers to a block pattern of size 8×8 in the compressed sequence. This is due to the 8×8 block DCT quantization of the compression algorithm.

- *bad edge rendition*: edges tend to be fuzzy due to the coarser quantization of high frequencies.
- *mosquito noise* manifests itself as an ambiguity in the edge direction: an edge appears in the direction conjugate to the actual edge. This effect is due to the implementation of the block DCT as a succession of a vertical and a horizontal one-dimensional DCT [9].
- *jagged motion* can be due to poor performance of the motion estimation. When the residual error of motion prediction is too large, it is coarsely quantized by the DCT quantization process.
- *flickering* appears when a scene has a high texture content. Texture blocks are compressed with varying quantization factors over time, which results in a visible flickering effect.
- *smoothing, loss of detail* are typical artifacts of quantization.
- *aliasing* appears when the content of the scene is above the Nyquist rate, either spatially or temporally.

An excellent survey of the various artifacts introduced by typical compression schemes can be found in [79].

3.4.2 Transmission Errors. A very important and often overlooked source of distortions is the transmission of the bitstream over a noisy channel. Digitally compressed video is typically transferred over a packet network. The actual transport can take place over a wire or wireless, but some higher level protocol such as ATM or TCP/IP ensures the transport of the video stream. Most applications require the streaming of video, i.e. the bitstream needs to be transported in such a way that it can be decoded and displayed in real time. The bitstream is transported in packets whose headers contain sequencing and timing information. This process is illustrated in Figure 10.3. Streams can also carry additional signaling information at the session level. A popular transport protocol at the moment is TCP/IP. A variety of protocols are then used to transport the audio-visual information. The real-time protocol (RTP) is used to transport, synchronize and signal the actual media and add timing information [51]; RTP packets are transported over UDP. The signalling is taken care of by additional protocols such as the H.323 family from the ITU [30], or the suite of protocols (SIP, SAP, SDP) from the Internet Engineering Task Force [50]. A comparison of these schemes is provided in [11].

Two different types of impairments can occur when transporting media over noisy channels. Packets can be lost due to excessive buffering in intermediate routers or switches, or they can be delayed to the point where they are not received in time for decoding. The latter is due to the queuing algorithm in

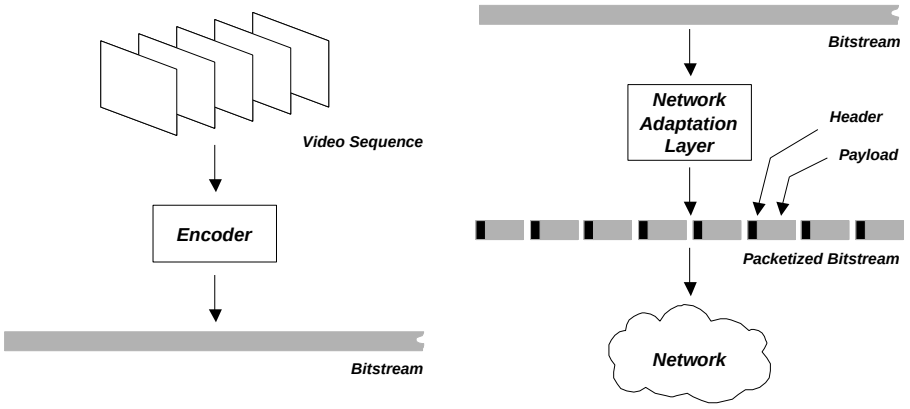

Figure 10.3 Illustration of a video transmission system. The video sequence is first compressed by the encoder. The resulting bitstream is packetized in the network adaptation layer, where a header containing sequencing and synchronization data is added to each packet. The packets are then sent over the network of choice.

routers and switches. To the application, both have the same effect: part of the media stream is not available, thus packets are missing when they are needed for decoding.

Such losses can affect both the semantics and the syntax of the media stream. When the losses affect syntactic information, not only the data relevant to the lost block are corrupted, but also any data that depend on this syntactic information. For example, a loss of packets containing data pertinent to an MPEG macroblock will corrupt all following macroblocks until an end of slice is encountered. This is due to the fact that the DC coefficient of a macroblock is differentially predicted between macroblocks and resets at the beginning of a slice. Also, for each of these corrupted macroblocks, all blocks that are motion predicted from these will be lost as well. Hence the loss of a single macroblock can affect the stream up to the next intra-coded frame. Figure 10.4 illustrates this phenomenon.

The effect can be even more damaging when global data is corrupted. An example of this is the timing information in an MPEG stream. The system layer specification of MPEG imposes that the decoder clock be synchronized with the encoder clock via periodic refresh of the program clock reference sent in some packet. Too much jitter on packet arrival can corrupt the synchronization of the decoder clock, which can result in highly noticeable impairments.

The visual effects of such losses vary a lot among decoders depending on their ability to deal with corrupted streams. Some decoders never recover from certain errors, while others apply clever concealment methods in order to minimize such effects.

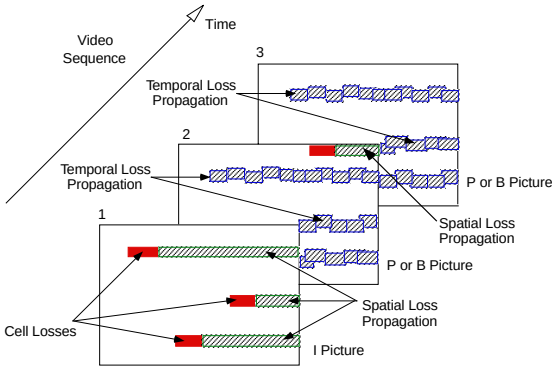

Figure 10.4 Spatial and temporal propagation of losses in an MPEG-compressed video sequence. The loss of a single macroblock causes the inability to decode the data up to the end of the slice. Macroblocks in neighboring frames that are predicted from the damaged area are corrupted as well.

4. VISION MODELS

Modeling the human visual system is a challenging task due to its inherent complexity; many of its properties are not fully understood even today. Its components have been studied in detail, but putting all the pieces together for a comprehensive model of human vision is far from trivial [73]. Quite a few models for still images have been developed in the past; their extension to moving pictures, however, has not received much attention until recently. In this section, we briefly review the development of metrics. We then present a perceptual distortion metric developed by the authors and discuss how the performance of such systems can be evaluated in a meaningful and reliable way.

4.1 MODELS AND METRICS

The objective for any vision model must be good agreement with experimental data. Threshold experiments and preference tests represent some of the most reliable methods available (cf. Chapter **??**). Therefore, an application making use of a vision model to measure perceptual differences in some way provides the most direct evaluation possibility. For this reason, we focus on vision models wrapped into distortion metrics here.

Distortion metrics need not necessarily rely on sophisticated models of the human visual system in order to perform well. They can exploit knowledge about the compression algorithm and the pertinent types of artifacts (cf. Section 3.4). Considering the variety of compression algorithms available and the rapid change of technology in this field, however, a distortion metric that is independent of the particular algorithm is preferable in order to avoid early obsolescence. Metrics based on human vision models are a way to achieve this technology independence, because they are the most general and potentially the most accurate ones [73].

Lukas and Budrikis [38] were the first to propose a spatio-temporal model of the human visual system for use in a video distortion metric. Other models and metrics followed now and then, but only in the past few years has there been an increasing interest in this topic, particularly in the engineering community. This is mainly due to the advent of digital video systems, which have exposed the limitations of the techniques traditionally used for video quality measurement.

For conventional analog video systems there are well-established performance standards. They rely on particular test signals and measurement procedures to determine parameters such as differential gain, differential phase or waveform distortion, which can be related to perceived quality with relatively high accuracy [80]. While these parameters are still useful today, their connection with perceived quality has become much more tenuous: because of compression, digital video systems exhibit artifacts fundamentally different from analog video systems (see Section 3.4). The amount and visibility of these distortions strongly depend on the actual scene content. Therefore, traditional signal quality measurements are inadequate for the evaluation of these compression artifacts.

Given these limitations, the designers of compression algorithms have had to resort to subjective viewing tests in order to obtain reliable ratings for the quality of compressed images or video (see Section 4.3.1). While these tests – if executed properly – certainly are the best measure of "true" perceptual quality, they are complex, time-consuming and consequently expensive. Hence, they are often highly impractical or not feasible at all.

Looking for faster alternatives, researchers have turned to simple error measures such as mean squared error (MSE) or signal-to-noise ratio (SNR), suggesting that they would be equally valid. However, these simple error measures operate solely on a pixel-by-pixel basis and neglect the important influence of image content and viewing conditions on the actual visibility of artifacts. Therefore, they often do not correlate well with perceived quality. These problems prompted the development of distortion metrics based on models of the human visual system.

4.2 A PERCEPTUAL DISTORTION METRIC

We now present the perceptual distortion metric (PDM) developed by the authors [60, 74]. The underlying vision model – an extension of a model for still images [72] – incorporates color perception, temporal and spatial mechanisms, contrast sensitivity, pattern masking, and the response properties of neurons in the primary visual cortex. The PDM works as follows (see Figure 10.5): After conversion to opponent-colors space, each of the resulting three components is subjected to a spatio-temporal perceptual decomposition, yielding a number of perceptual channels. They are weighted according to contrast sensitivity data

and subsequently undergo a contrast gain control stage. Finally, all the sensor differences are combined into a distortion measure.

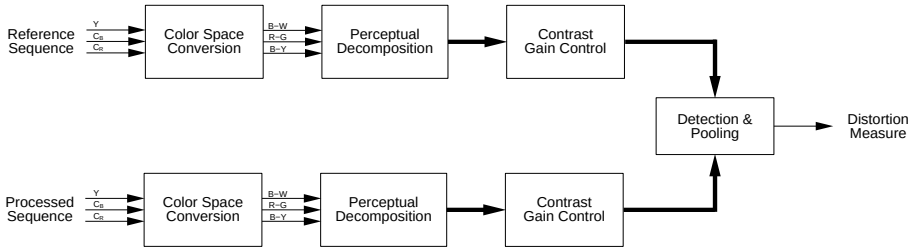

Figure 10.5 Block diagram of the PDM [74].

4.2.1 Color Space Conversion. The first stage of the PDM performs the color space conversion of the video input, usually coded in $\mathrm{Y'C'_BC'_R}$. According to the theory of opponent colors, the human visual system decorrelates the input signals from the cones on the retina into black-white (B-W), red-green (R-G) and blue-yellow (B-Y) difference signals (cf. Section 3.2). The PDM relies on a particular opponent-colors space that is pattern-color separable [43, 44], i.e. color perception and pattern sensitivity can be decoupled and treated in separate stages.

4.2.2 Perceptual Decomposition. The perceptual decomposition models the multi-channel architecture of the human visual system. It is performed first in the temporal and then in the spatial domain. Decomposing the input into a number of spatio-temporal channels is necessary in order to be able to account for the fact that masking is strongest between stimuli of similar characteristics (e.g. similar frequency and orientation) in subsequent stages.

The temporal filters used in the PDM are based on a recent model of temporal mechanisms [19]. The design objective for these filters in the PDM was to keep the delay to a minimum, because in some applications of distortion metrics such as monitoring and control, a short response time is crucial. A trade-off has to be found between an acceptable delay and the accuracy with which the temporal mechanisms ought to be approximated. Recursive infinite impulse response (IIR) filters fare better in this respect than (non-recursive) finite impulse response (FIR) filters [35].

Therefore, the temporal mechanisms are modeled by two IIR filters in the PDM. They were computed by means of a least-square fit to the frequency magnitude responses of the respective mechanisms. A filter with 2 poles and 2 zeros was fitted to the sustained mechanism, and a filter with 4 poles and 4 zeros was fitted to the transient mechanism. This has been found to yield the shortest delay while still maintaining a good approximation of the frequency

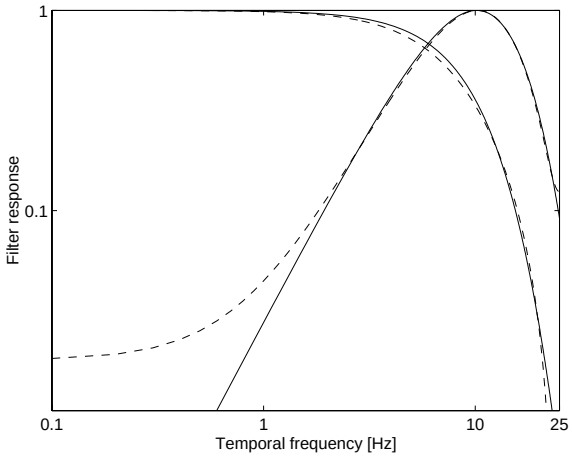

Figure 10.6 Frequency responses of sustained (low-pass) and transient (band-pass) mechanisms of vision according to [19] (solid), and the IIR filter approximations used in the PDM for a sampling frequency of 50 Hz (dashed).

responses, as shown in Figure 10.6. In the present implementation, the low-pass filters are applied to all three color channels, but the band-pass filter is applied only to the luminance channel in order to reduce computing time. This simplification is based on the fact that color contrast sensitivity is rather low for high frequencies (cf. Section 2.2).

The decomposition in the spatial domain is carried out by means of the steerable pyramid transform [54].[2] This transform decomposes an image into a number of spatial frequency and orientation bands. Its basis functions are directional derivative operators. For use within a vision model, it has the advantage of being rotation-invariant and self-inverting, and it minimizes the amount of aliasing in the subbands. In the present implementation, the basis filters have octave bandwidth and octave spacing; five subband levels with four orientation bands each plus one low-pass band are computed (see Figure 10.7 for an illustration). The same decomposition is used for all channels.

4.2.3 Contrast Gain Control Stage. Modeling pattern masking is one of the most critical aspects of video quality assessment, because the visibility of distortions is highly dependent on the local background. Contrast gain control models can explain a wide variety of empirical masking data. These models were inspired by analyses of the responses of neurons in the visual cortex of the cat [2, 25, 26], where contrast gain control serves as a mechanism to keep neural responses within the permissible dynamic range while at the same time retaining global pattern information.

[2] Source code and filter kernels for the steerable pyramid transform are available at http://www.cis.upenn.edu/~eero/steerpyr.html.

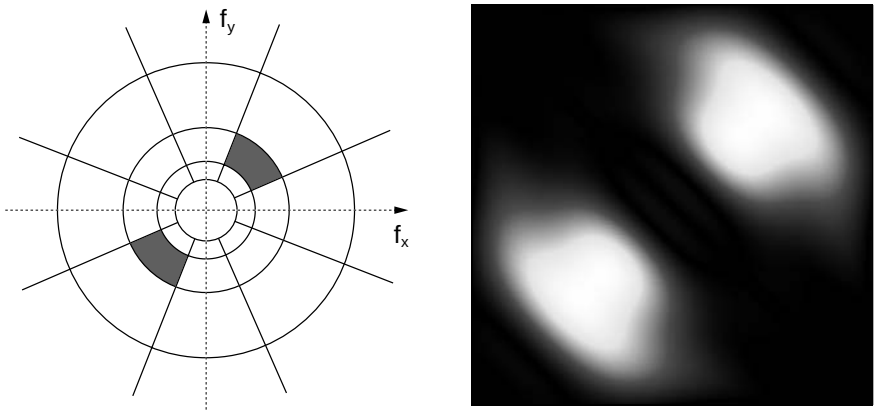

Figure 10.7 Illustration of the partitioning of the spatial frequency plane by the steerable pyramid transform [54]. Three levels and the isotropic low-pass filter are shown. The bands at each level are tuned to orientations of 0, 45, 90 and 135 degrees. The shaded region indicates the spectral support of a single subband, whose actual frequency response is shown on the right.

Contrast gain control can be realized by an excitatory nonlinearity that is inhibited divisively by a pool of responses from other neurons [16, 58]. Masking occurs through the inhibitory effect of the normalizing pool. A mathematical generalization of these models facilitates the integration of many kinds of channel interactions and spatial pooling [67]. Introduced for luminance images, this contrast gain control model can be extended to color and to sequences [72, 74]. In its most general form, the above-mentioned response pool may combine coefficients from the dimensions of time, color, temporal frequency, spatial frequency, orientation, space, and phase; in the present implementation of the PDM, it is limited to orientation.

4.2.4 Detection and Pooling. The information residing in various channels is integrated in higher-level areas of the brain. This can be simulated by gathering the data from these channels according to rules of probability or vector summation, also known as pooling [46].

The pooling stage of the PDM combines the elementary differences between the sensor outputs over several dimensions by means of vector summation. In principle, any subset of dimensions can be used, depending on what kind of result is desired. For example, pooling may be limited to single frames first to determine the variation of distortions over time, and the total distortion can then be computed from the values for each frame.

4.2.5 Model Fitting. The model contains several parameters that have to be adjusted in order to accurately represent the human visual system. Threshold

data from contrast sensitivity and contrast masking experiments are used for this procedure. In the fitting process, the input of the PDM imitates the stimuli used in these experiments, and the free model parameters are adjusted in such a way that the output approximates these threshold curves.

Contrast sensitivity is modeled by setting the gains of the spatial and temporal filters such that the model predictions match empirical threshold data from spatio-temporal contrast sensitivity experiments for both color and luminance stimuli. While this approach may be slightly inferior to pre-filtering the B-W, R-G and B-Y channels with their respective contrast sensitivity functions in terms of approximation accuracy, it is easier to implement and saves computing time. For the B-W channels, the weights are chosen so as to match contrast sensitivity measurements from [32]. For the R-G and B-Y channels, similar data from [33] are used.

The parameters of the contrast gain control stage are determined by fitting the model's responses to masked gratings. For the B-W channel, empirical data from several intra- and inter-channel contrast masking experiments from [16] are used. For the R-G and B-Y channels, the parameters are adjusted to fit similar data from [56].

In the vector summation of the pooling process, different exponents have been found to yield good results for different experiments and implementations. In the PDM, pooling over channels and over pixels is carried out with an exponent of 2, whereas an exponent of 4 is used for pooling over frames.

Our simulation results indicate that the overall quality of the fits to the above-mentioned empirical data is quite good and close to the difference between measurements from different observers. Most of the effects found in the psychophysical experiments are captured by the model. However, one drawback of this modeling approach should be noted: Because of the nonlinear nature of the model, the parameters can only be determined by means of a numerical iterative fitting process, which is computationally expensive.

4.3 EVALUATION

In order to evaluate vision models, subjective experiments are necessary. Subjective ratings form the benchmark for objective metrics. However, different applications may require different testing procedures (cf. Chapter **??**) and data analysis methods.

4.3.1 Subjective Testing. Formal subjective testing is defined in ITU-R Rec. 500 [28], which suggests standard viewing conditions, criteria for observer and test scene selection, assessment procedures, and analysis methods. We outline three of the more commonly used procedures here:

- Double Stimulus Continuous Quality Scale (DSCQS). Viewers are shown multiple sequence pairs consisting of a "reference" and a "test" sequence, which are rather short (typically 10 seconds). The reference and test sequence are presented twice in alternating fashion, with the order of the two chosen randomly for each trial. Subjects are not informed which is the reference and which is the test sequence. They rate each of the two separately on a continuous quality scale ranging from "bad" to "excellent". Analysis is based on the difference in rating for each pair, which is often calculated from an equivalent numerical scale from 0 to 100.

- Double Stimulus Impairment Scale (DSIS). As opposed to the DSCQS method, the reference is always shown before the test sequence, and neither is repeated. Subjects rate the amount of impairment in the test sequence on a discrete five-level scale ranging from "very annoying" to "imperceptible".

- Single Stimulus Continuous Quality Evaluation (SSCQE) [40]. Instead of seeing separate short sequence pairs, viewers watch a program of typically 20-30 minutes duration which has been processed by the system under test; the reference is not shown. Using a slider whose position is recorded continuously, the subjects rate the instantaneously perceived quality on the DSCQS scale from "bad" to "excellent".

4.3.2 Metric Comparisons. The sequences and subjective ratings used in demonstrations of the performance of a particular metric have been mostly proprietary, as hardly any subjectively rated sequences are publicly available. This has made fair comparisons of different metrics difficult.

In order to alleviate this problem, the Video Quality Experts Group (VQEG)[3] was formed in 1997. Its objectives have been to collect reliable subjective ratings for a well-defined set of sequences and to evaluate the performance of different video quality assessment systems with respect to these sequences. The emphasis of the first phase of VQEG was on production- and distribution-class video, i.e. mainly MPEG-2 encoded sequences with different profiles, levels and other parameter variations, including encoder concatenation, conversions between analog and digital video, and transmission errors. A set of 8-second scenes emphasizing different characteristics (e.g. spatial detail, color, motion) was selected by independent labs; the scenes were disclosed to the proponents only after the submission of their metrics. In total, 20 scenes were encoded for 16 test conditions each.

[3] See http://www.crc.ca/vqeg/ for an overview of its activities.

Ten different systems for video quality assessment – among them the PDM from Section 4.2 – were submitted, and their output for each of the 320 sequences was recorded. In parallel, DSCQS subjective ratings for all sequences were obtained at eight independent subjective testing labs. The statistical methods used for the performance analysis were variance-weighted regression, nonlinear regression, Spearman rank-order correlation, and outlier ratio. The results of the data analysis show that the performance of most models as well as PSNR are statistically equivalent for all four criteria, leading to the conclusion that no single model outperforms the others in all cases and for the entire range of test sequences [48, 63]. Furthermore, no objective video quality assessment system was able to achieve an accuracy comparable to the agreement between different subject groups.

4.3.3 PDM Results. Preliminary results for the set of sequences used in the VQEG testing effort are reported here for the perceptual distortion metric (PDM) from Section 4.2. Figure 10.8 shows a correlation plot of the PDM ratings vs. the mean subjective scores (DSCQS ratings) for all 320 test sequences. The metric performs well over all test cases: The overall correlation between the mean subjective scores and the PDM ratings is close to 0.8; for certain subsets of test cases, correlations approach 0.9. The PDM can handle MPEG as well as non-MPEG distortions and also behaves well with respect to sequences with transmission errors. Most of its outliers are due to the lowest-bitrate condition of the test. Such performance degradations for clearly visible distortions are to be expected, because the metric is based on a threshold model of human vision.

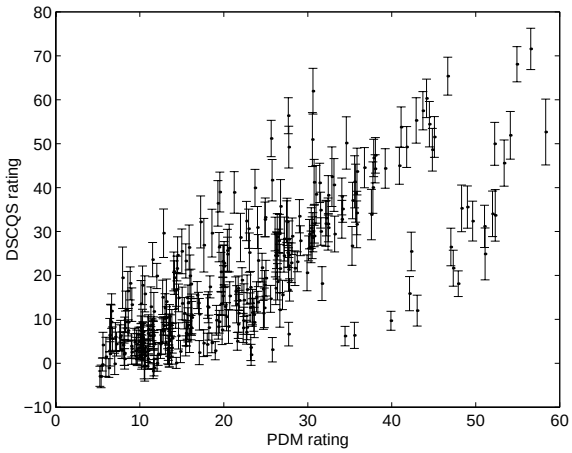

Figure 10.8 Correlation plot of the PDM ratings and the corresponding mean subjective scores for all 320 test sequences from the VQEG effort. The error bars indicate the 95% confidence intervals of the subjective DSCQS ratings.

Further analyses of the PDM with respect to the VQEG sequences also revealed that visual quality metrics which are essentially equivalent at the thresh-

old level can exhibit significant performance differences for complex sequences depending on the implementation choices made for various components of the PDM [75]. In particular, this was found to be true for a comparison of a number of different color spaces, including luminance-only implementations, as well as two pooling algorithms and their parameters.

5. VIDEO APPLICATIONS

There is a wide variety of applications for vision models in video systems, including:

- evaluation, test and comparison of video codecs;
- end-to-end testing of video transmission systems;
- perceptual video compression;
- online quality monitoring;
- encoder regulation and quality control;
- perceptual video restoration.

Coupled with appropriate video segmentation methods, the visual quality of specific features (e.g. contours or textures) or specific compression artifacts (e.g. blockiness) may be evaluated separately, which can be useful to tune certain parameters of the encoder [60]. In a similar fashion, the quality of motion rendition can be assessed [10]. We take a closer look at some of these applications in this section.

5.1 OUT-OF-SERVICE TESTING

In out-of-service testing, the test operation is carried out while the system under test is not performing service. Testing can be done at the box level, where the equipment is disconnected from its operating mode. The test operation typically imposes a given input and compares the output to a reference. The operator feeds a video stream into the system under test; the test equipment synchronizes the output of the system under test with the original signal and applies a metric to it. A generic out-of-service test setup is depicted in Figure 10.9. The metrics developed by Tektronix/Sarnoff [37], KDD [23], or NASA [66] as well as the perceptual distortion metric (PDM) from Section 4.2 are examples of these methods. Many of these metrics require the video stream to be edited in such a way that the output and the reference are aligned spatially and temporally, which can be achieved by means of synchronization markers.

Out-of-service testing can be applied at the system level as well. The system proposed in [61] offers such a solution (see Figure 10.10). The testing

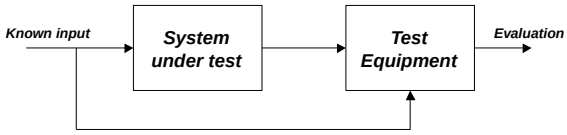

Figure 10.9 Out-of-service testing of a system.

methodology relies on a test pattern generator that creates customizable synthetic video sequences. The synthetic sequences are used as input to a video transmission system. The first frame of the sequence contains synchronization data as well as identification of the test sequence and all configurable parameters. A device connected to the decoder identifies a test sequence from the synchronization data. Based on these it recreates the original sequence at the decoder site, which permits to apply a distortion metric on the decoded video and the original signal.

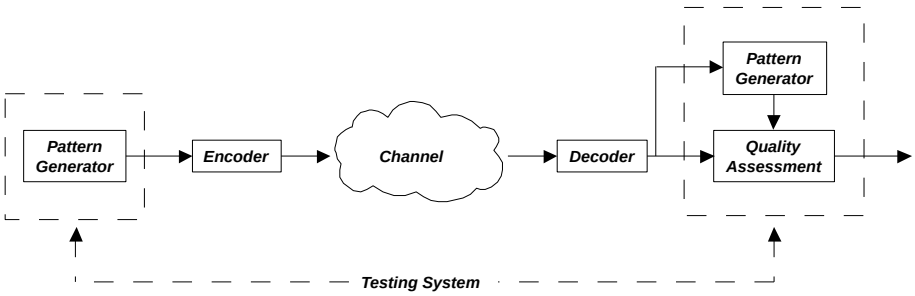

Figure 10.10 Block diagram of the Test Pattern Generator testing system.

Most of the systems submitted to the VQEG evaluation discussed in Section 4.3.2 are designed for out-of-service testing. They are based on a comparison of the distorted sequence with a reference. Such a methodology is aimed at assessing the performance of a system and evaluate it, but it is not meant to be monitoring equipment.

5.2 IN-SERVICE TESTING

In-service testing is aimed at troubleshooting equipment while it is in service. The setup can be intrusive or not, depending on the objective of the test and the nature of the testing methodology. Figure 10.11 illustrates both cases. In many instances, in-service testing of video quality means that the original signal is not available for comparison, which makes an accurate assessment much more difficult. The algorithms are then based on some a priori knowledge about the scene content or on a modeling of the video scene. Several methods have been proposed recently [3, 5]. Most of them aim at identifying certain features in a scene and assessing their distortion.

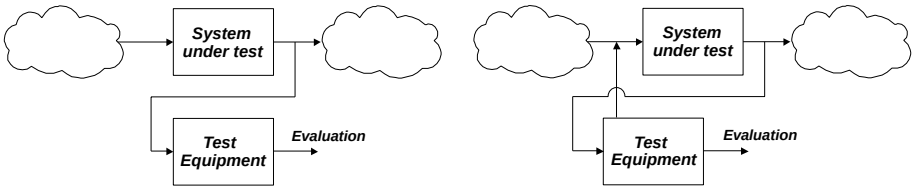

Figure 10.11 Non-intrusive (left) and intrusive (right) in-service testing setup.

Another in-service testing method was developed and implemented at the Hewlett-Packard Laboratories (patent pending). The tool computes a prediction of the mean squared error in an MPEG bitstream and then applies a multi-channel video fidelity metric to predict the visibility of the distortions. The estimation of the MSE is based on an adaptive algorithm in the compressed domain.

5.3 ENCODER REGULATION

An important application for video quality metrics is encoder regulation. Most existing encoders are based on a minimization of the mean squared error. A number of authors have proposed to use a perceptual distortion metric in the encoding process so as to compress the sequence in a way that would minimize the visibility of the distortions. Early work includes the DCTune algorithm [65], which tries to optimize the JPEG quantization matrix for a given picture.

Other methods attempt to jointly optimize encoding and transmission parameters so as to account for losses in the transmission of the stream. Such algorithms determine the optimal rate distribution between MPEG-2 and media-independent forward error correction (FEC) given network loss parameters [20, 62]. The optimality is defined in terms of maximum end-to-end video quality as measured by a vision model. This scheme was shown to outperform classical FEC schemes due to its adaptivity to the video material and the network conditions.

6. CONCLUSIONS

Digital video systems have matured, and their market share has been growing continuously in the last few years. In this chapter we reviewed the current standards of video technology and some of the design issues involved. Vision models are used increasingly in the attempt to analyze the operating behavior of such systems and to overcome their limitations. We discussed the temporal aspects of human vision as well as a variety of modeling approaches, in particular in the domain of perceptual video quality assessment. Several metrics have already been proposed and evaluated in the search for an international

standard. Nevertheless, this is still a relatively young field of research, and many challenging questions remain to be answered.

References

[1] A. J. Ahumada, Jr., B. L. Beard, R. Eriksson: "Spatio-temporal discrimination model predicts temporal masking function." in *Proc. SPIE*, vol. 3299, pp. 120–127, San Jose, CA, 1998.

[2] D. G. Albrecht, W. S. Geisler: "Motion selectivity and the contrast-response function of simple cells in the visual cortex." *Vis. Neurosci.* **7**:531–546, 1991.

[3] V. Baroncini, A. Pierotti: "Single-ended objective quality assessment of DTV." in *Proc. SPIE*, vol. 3845, Boston, MA, 1999.

[4] V. Bhaskaran, K. Konstantinides: *Image and Video Compression Standards. Algorithms and Architectures*. Kluwer Academic Publishers, 2^{nd} edn., 1997.

[5] P. Bretillon, J. Baina: "Method for image quality monitoring on digital television networks." in *Proc. SPIE*, vol. 3845, Boston, MA, 1999.

[6] C. A. Burbeck, D. H. Kelly: "Spatiotemporal characteristics of visual mechanisms: Excitatory-inhibitory model." *J. Opt. Soc. Am.* **70**(9):1121–1126, 1980.

[7] T. Carney, S. A. Klein, Q. Hu: "Visual masking near spatiotemporal edges." in *Proc. SPIE*, vol. 2657, pp. 393–402, San Jose, CA, 1996.

[8] G. R. Cole, C. F. Stromeyer, III., R. E. Kronauer: "Visual interactions with luminance and chromatic stimuli." *J. Opt. Soc. Am. A* **7**(1):128–140, 1990.

[9] S. Comes, B. Macq, M. Mattavelli: "Postprocessing of images by filtering the unmasked coding noise." *IEEE Trans. Image Processing* **8**(8):1050–1062, 1999.

[10] D. Costantini et al.: "Motion rendition quality metric for MPEG coded video." in *Proc. ICIP*, vol. 1, pp. 889–892, Lausanne, Switzerland, 1996.

[11] İ. Dalgıç, H. Fang: "Comparison of H.323 and SIP for internet telephony signaling." in *Proc. SPIE*, vol. 3845, Boston, MA, 1999.

[12] S. Daly: "The visible differences predictor: An algorithm for the assessment of image fidelity." in *Digital Images and Human Vision*, ed. A. B. Watson, pp. 179–206, MIT Press, 1993.

[13] S. Daly: "Engineering observations from spatiovelocity and spatiotemporal visual models." in *Proc. SPIE*, vol. 3299, pp. 180–191, San Jose, CA, 1998.

[14] J. G. Daugman: "Two-dimensional spectral analysis of cortical receptive field profiles." *Vision Res.* **20**(10):847–856, 1980.

[15] G. de Haan, E. B. Bellers: "Deinterlacing – an overview." *Proc. IEEE* **86**(9):1839–1857, 1998.

[16] J. M. Foley: "Human luminance pattern-vision mechanisms: Masking experiments require a new model." *J. Opt. Soc. Am. A* **11**(6):1710–1719, 1994.

[17] K. H. Foster et al.: "Spatial and temporal frequency selectivity of neurons in visual cortical areas V1 and V2 of the macaque monkey." *J. Physiol.* **365**:331–363, 1985.

[18] R. E. Fredericksen, R. F. Hess: "Temporal detection in human vision: Dependence on stimulus energy." *J. Opt. Soc. Am. A* **14**(10):2557–2569, 1997.

[19] R. E. Fredericksen, R. F. Hess: "Estimating multiple temporal mechanisms in human vision." *Vision Res.* **38**(7):1023–1040, 1998.

[20] P. Frossard, O. Verscheure: *Joint Source/FEC Rate Selection for Quality-Optimal MPEG-2 Video Delivery*. Tech. Rep. 1999-04, Signal Processing Lab, Swiss Federal Institute of Technology, Lausanne, 1999.

[21] B. Girod: "The information theoretical significance of spatial and temporal masking in video signals." in *Proc. SPIE*, vol. 1077, pp. 178–187, Los Angeles, CA, 1989.

[22] M. W. Greenlee, J. P. Thomas: "Effect of pattern adaptation on spatial frequency discrimination." *J. Opt. Soc. Am. A* **9**(6):857–862, 1992.

[23] T. Hamada, S. Miyaji, S. Matsumoto: "Picture quality assessment system by three-layered bottom-up noise weighting considering human visual perception." *SMPTE J.* **108**(1):20–26, 1999.

[24] S. T. Hammett, A. T. Smith: "Two temporal channels or three? A re-evaluation." *Vision Res.* **32**(2):285–291, 1992.

[25] D. J. Heeger: "Half-squaring in responses of cat striate cells." *Vis. Neurosci.* **9**:427–443, 1992.

[26] D. J. Heeger: "Normalization of cell responses in cat striate cortex." *Vis. Neurosci.* **9**:181–197, 1992.

[27] R. F. Hess, R. J. Snowden: "Temporal properties of human visual filters: Number, shapes and spatial covariation." *Vision Res.* **32**(1):47–59, 1992.

[28] ITU-R Recommendation BT.500-10: "Methodology for the subjective assessment of the quality of television pictures." ITU, Geneva, Switzerland, 2000.

[29] ITU-R Recommendation BT.601-5: "Studio encoding parameters of digital television for standard 4:3 and wide-screen 16:9 aspect ratios." ITU, Geneva, Switzerland, 1995.

[30] ITU-T Recommendation H.323: "Visual telephone systems and equipment for local area networks which provide a non-guaranteed quality of service." ITU, Geneva, Switzerland, 1998.

[31] D. H. Kelly: “Motion and vision. I. Stabilized images of stationary gratings.” *J. Opt. Soc. Am.* **69**(9):1266–1274, 1979.

[32] D. H. Kelly: “Motion and vision. II. Stabilized spatio-temporal threshold surface.” *J. Opt. Soc. Am.* **69**(10):1340–1349, 1979.

[33] D. H. Kelly: “Spatiotemporal variation of chromatic and achromatic contrast thresholds.” *J. Opt. Soc. Am.* **73**(6):742–750, 1983.

[34] J. J. Koenderink, A. J. van Doorn: “Spatiotemporal contrast detection threshold surface is bimodal.” *Opt. Letters* **4**(1):32–34, 1979.

[35] P. Lindh, C. J. van den Branden Lambrecht: “Efficient spatio-temporal decomposition for perceptual processing of video sequences.” in *Proc. ICIP*, vol. 3, pp. 331–334, Lausanne, Switzerland, 1996.

[36] M. A. Losada, K. T. Mullen: “The spatial tuning of chromatic mechanisms identified by simultaneous masking.” *Vision Res.* **34**(3):331–341, 1994.

[37] J. Lubin, D. Fibush: “Sarnoff JND vision model.” T1A1.5 Working Group Document #97-612, ANSI T1 Standards Committee, 1997.

[38] F. X. J. Lukas, Z. L. Budrikis: “Picture quality prediction based on a visual model.” *IEEE Trans. Comm.* **30**(7):1679–1692, 1982.

[39] M. B. Mandler, W. Makous: “A three-channel model of temporal frequency perception.” *Vision Res.* **24**(12):1881–1887, 1984.

[40] MOSAIC: *A New Single Stimulus Quality Assessment Methodology*. RACE R2111, 1996.

[41] E. Peli: “Contrast in complex images.” *J. Opt. Soc. Am. A* **7**(10):2032–2040, 1990.

[42] E. Peli: “In search of a contrast metric: Matching the perceived contrast of Gabor patches at different phases and bandwidths.” *Vision Res.* **37**(23):3217–3224, 1997.

[43] A. B. Poirson, B. A. Wandell: “Appearance of colored patterns: Pattern-color separability.” *J. Opt. Soc. Am. A* **10**(12):2458–2470, 1993.

[44] A. B. Poirson, B. A. Wandell: “Pattern-color separable pathways predict sensitivity to simple colored patterns.” *Vision Res.* **36**(4):515–526, 1996.

[45] C. Poynton: “The rehabilitation of gamma.” in *Proc. SPIE*, vol. 3299, pp. 232–249, San Jose, CA, 1998.

[46] R. F. Quick, Jr.: “A vector-magnitude model of contrast detection.” *Kybernetik* **16**:65–67, 1974.

[47] J. G. Robson: “Spatial and temporal contrast-sensitivity functions of the visual system.” *J. Opt. Soc. Am.* **56**:1141–1142, 1966.

[48] A. M. Rohaly et al.: “Video Quality Experts Group: Current results and future directions.” in *Proc. SPIE*, vol. 4067, Perth, Australia, 2000.

[49] J. Ross, H. D. Speed: "Contrast adaptation and contrast masking in human vision." *Proc. R. Soc. Lond. B* **246**:61–70, 1991.

[50] H. Schulzrinne, J. Rosenberg: "Internet telephony: Architecture and protocols – an IETF perspective." *Computer Networks and ISDN Systems* **31**:237–255, 1999.

[51] H. Schulzrinne et al.: *RTP: A Transport Protocol for Real-Time Applications*. Tech. Rep. RFC 1889, IETF, 1996.

[52] A. J. Seyler, Z. L. Budrikis: "Measurements of temporal adaptation to spatial detail vision." *Nature* **184**:1215–1217, 1959.

[53] A. J. Seyler, Z. L. Budrikis: "Detail perception after scene changes in television image presentations." *IEEE Trans. Inform. Theory* **11**(1):31–43, 1965.

[54] E. P. Simoncelli et al.: "Shiftable multi-scale transforms." *IEEE Trans. Inform. Theory* **38**(2):587–607, 1992.

[55] R. J. Snowden, S. T. Hammett: "Spatial frequency adaptation: Threshold elevation and perceived contrast." *Vision Res.* **36**(12):1797–1809, 1996.

[56] E. Switkes, A. Bradley, K. K. De Valois: "Contrast dependence and mechanisms of masking interactions among chromatic and luminance gratings." *J. Opt. Soc. Am. A* **5**(7):1149–1162, 1988.

[57] W. J. Tam et al.: "Visual masking at video scene cuts." in *Proc. SPIE*, vol. 2411, pp. 111–119, San Jose, CA, 1995.

[58] P. C. Teo, D. J. Heeger: "Perceptual image distortion." in *Proc. SPIE*, vol. 2179, pp. 127–141, San Jose, CA, 1994.

[59] G. Thomas: "A comparison of motion-compensated interlace-to-progressive conversion methods." *Signal Processing: Image Communication* **12**(3):209–229, 1998.

[60] C. J. van den Branden Lambrecht: *Perceptual Models and Architectures for Video Coding Applications*. Ph.D. thesis, École Polytechnique Fédérale de Lausanne, Switzerland, 1996.

[61] C. J. van den Branden Lambrecht et al.: "Automatically assessing MPEG coding fidelity." *IEEE Design and Test Magazine* **12**(4):28–33, 1995.

[62] O. Verscheure: *User-Oriented QoS in MPEG-2 Video Delivery*. Ph.D. thesis, École Polytechnique Fédérale de Lausanne, Switzerland, 1999.

[63] VQEG: "Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment." 2000, available at http://www.crc.ca/vqeg/.

[64] A. B. Watson: "Temporal sensitivity." in *Handbook of Perception and Human Performance*, eds. K. R. Boff, L. Kaufman, J. P. Thomas, vol. 1, chap. 6, John Wiley & Sons, 1986.

[65] A. B. Watson: “DCTune: A technique for visual optimization of DCT quantization matrices for individual images.” in *SID Symposium Digest*, vol. 24, pp. 946–949, 1993.

[66] A. B. Watson: “Toward a perceptual video quality metric.” in *Proc. SPIE*, vol. 3299, pp. 139–147, San Jose, CA, 1998.

[67] A. B. Watson, J. A. Solomon: “Model of visual contrast gain control and pattern masking.” *J. Opt. Soc. Am. A* **14**(9):2379–2391, 1997.

[68] M. A. Webster, E. Miyahara: “Contrast adaptation and the spatial structure of natural images.” *J. Opt. Soc. Am. A* **14**(9):2355–2366, 1997.

[69] M. A. Webster, J. D. Mollon: “Adaptation and the color statistics of natural images.” *Vision Res.* **37**(23):3283–3298, 1997.

[70] S. J. P. Westen, R. L. Lagendijk, J. Biemond: “Spatio-temporal model of human vision for digital video compression.” in *Proc. SPIE*, vol. 3016, pp. 260–268, San Jose, CA, 1997.

[71] H. R. Wilson, R. Humanski: “Spatial frequency adaptation and contrast gain control.” *Vision Res.* **33**(8):1133–1149, 1993.

[72] S. Winkler: “A perceptual distortion metric for digital color images.” in *Proc. ICIP*, vol. 3, pp. 399–403, Chicago, IL, 1998.

[73] S. Winkler: “Issues in vision modeling for perceptual video quality assessment.” *Signal Processing* **78**(2):231–252, 1999.

[74] S. Winkler: “A perceptual distortion metric for digital color video.” in *Proc. SPIE*, vol. 3644, pp. 175–184, San Jose, CA, 1999.

[75] S. Winkler: “Quality metric design: A closer look.” in *Proc. SPIE*, vol. 3959, pp. 37–44, San Jose, CA, 2000.

[76] S. Winkler, P. Vandergheynst: “Computing isotropic local contrast from oriented pyramid decompositions.” in *Proc. ICIP*, vol. 4, pp. 420–424, Kyoto, Japan, 1999.

[77] J. Yang, W. Makous: “Spatiotemporal separability in contrast sensitivity.” *Vision Res.* **34**(19):2569–2576, 1994.

[78] J. Yang, W. Makous: “Implicit masking constrained by spatial inhomogeneities.” *Vision Res.* **37**(14):1917–1927, 1997.

[79] M. Yuen, H. R. Wu: “A survey of hybrid MC/DPCM/DCT video coding distortions.” *Signal Processing* **70**(3):247–278, 1998.

[80] W. Y. Zou: “Performance evaluation: From NTSC to digitally compressed video.” *SMPTE J.* **103**(12):795–800, 1994.