CHAPTER

Communicating Pictures—The Future

13

CHAPTER OUTLINE

In this chapter we briefly review the demands of future content types in the context of an *extended video parameter space*, to examine how parameters such as spatial resolution, temporal resolution, dynamic range, and color gamut affect our viewing experience, both individually and collectively. The delivery requirements associated with this extended parameter space are considered, with a focus on how compression might affect and influence the video quality.

Emphasizing the impact of, and the interactions between, the video parameters and their content dependence, the justification is made for an increased use of perceptual models. It is argued that perception-based compression techniques can improve video quality and that a new *rate–quality optimization* approach is required if we are to efficiently and effectively code the immersive formats of the future. In this context, new approaches to compression are proposed based on texture analysis and synthesis, moderated by in-loop quality metrics.

For colour and higher quality versions of Figures 13.3, 13.6, 13.12 and 13.13 please refer to the electronic version or the website.

Communicating Pictures. http://dx.doi.org/10.1016/B978-0-12-405906-1.00013-1

13.1 The motivation: more immersive experiences

The moving image industry is a trillion dollar business worldwide, across film, broadcast, and streaming. And, according to the USA Bureau of Labor Statistics [1] we spend one fifth of our waking lives watching movies, television, or some other form of edited moving image.

Visual experiences are thus important drivers for business and technology development; by 2020 it is predicted that the number of network-connected devices will reach 1000 times the world's population, i.e. there will be 7 trillion connected devices for 7 billion people [2]. YouTube video currently accounts for around 25% of all internet traffic and, by 2016, Cisco predicts that video will account for 54% of all traffic (86% including P2P distribution) with total annual IP traffic rising to 1.3 ZB (zettabyte $= 10^{21}$ bytes $=$ 1000 EB) [3]. Some mobile network operators predict a doubling of wireless traffic every year for the next 10 years driven primarily by video.

Despite the impressive facts and predictions above, for high value content shown in cinemas, attendance and revenues across much of the developed world in 2012 were flat. Although there was an increase of 30% in the number of digital 3-D screens, according to statistics from the British Film Institute, the revenues from 3-D movies actually decreased. 3-D has been at least as much driven by technology capabilities and the need for businesses to add value, as it has been by user demands or indeed because it provides a more engaging experience. It is now clear that the industry emphasis is shifting from stereoscopic 3-D toward UHDTV formats with larger and brighter screens. Audiences *may* still want to watch 3-D, but it seems that many are not willing to pay the premium.

So there is a significant demand for new, more immersive content: from users who want to experience it, from producers who want to add value to their content, and from operators who want to charge you and I more to watch it. The questions we need to ask are: what is the format of such content, how do we assess its quality and how do we preserve its values during acquisition, delivery and display?

13.2 Emerging formats and extended video parameter spaces

The drive for increased immersion, coupled with greater mobility, means that compression efficiency is a priority for the media, communications, and entertainment industries. Two particular recent events—the explosion of video consumption on portable and personal devices, and the increased investment in UHDTV—both demand significant increases in compression efficiency.

However, to produce and deliver more immersive content, it is not sufficient just to increase spatial resolution or screen size. Although these parameters play an important role, other factors such as frame rate, dynamic range, and color gamut are also key to delivering high value experiences. We will examine some of these briefly below.

13.2.1 Influences

There are many things that can influence our perception of video quality and affect the immersiveness of the viewing experience, not least our interest in the subject matter. However, putting the narrative to one side, the main factors are:

- Resolution: temporal, spatial and SNR.
- Motion characteristics: blurring and masking effects.
- Colorimetry: gamut bit depth, tone mapping, and grading.
- Display: size, aspect ratio, dynamic range, and viewing environment.
- Views: monoscopic, stereoscopic, and multiview.
- Cinematography: including set design, lighting, shot length, scene geometry, motion, and color.
- Other psychovisual characteristics: such as peripheral vision, saliency, depth cues, and image noise.
- And of course audio: ambisonics and wavefields are both exciting developments that complement the visual to enhance the experience.

Complex interactions between these exist and they will, of course, be confounded by strong content dependence. Their impact is however highly moderated by the method and degree of compression employed.

13.2.2 Spatial resolution

Why spatial detail is important

Spatial resolution is important as it influences how sharply we see objects. The key parameter is not simply the number of pixels in each row or column of the display, but the angle subtended, θ, by each of these pixels on the viewer's retina. Thought of in this way, increased spatial resolution is particularly important as screen sizes become larger and viewing distances closer. Also, the *effective spatial resolution* will depend on how the compression system interacts with the displayed content; it is very easy to turn HD content into SD through compression. Contrary to popular advertising, there is no such thing as "HD quality." Let us look at this a little closer.

UHDTV and ITU-R REC.2020

The Ultra High Definition Television (UHDTV) standard, finalized recently by ITU-R under the title of Rec.2020 [4], will be a major influence on compression methods in the future. It is intended to deliver the highest quality immersive viewing experience to the home.

Table 13.1 shows the basic parameters for UHDTV and some interesting points can be observed from this table. Firstly, the assumed viewing distance places the viewer very close to the picture at 0.75 times the screen height, compared to $3H$ for HDTV (Rec.709). Obviously large screens are assumed as the viewing angle is stated as 100°, in contrast to that for conventional HDTV of around 33°. This will mean significant changes in the way we interact (psychovisually) with the content through

Table 13.1 Parameter set for UHDTV/ITU-R Rec.2020.

Parameter	Values
Picture aspect ratio	16 × 9
Pixel count ($H \times V$)	7680 × 4320, 3840 × 2160
Sampling lattice	Orthogonal
Pixel aspect ratio	1:1 (square pixels)
Frame frequency (Hz)	120, 60, 60/1.001, 50, 30, 30/1.001, 25, 24, 24/1.001
Color bit depth	10 or 12 bits per component
Scan mode	Progressive
Viewing distance	0.75H
Viewing angle	100°

increased head and eye movements. In practice this author suspects that the viewing distance will be greater than this for most people in many environments. Nonetheless, different cinematographic methods will be required. Secondly, the maximum frame rate is 120 Hz, rather than the 25, 30, 50, or 60 Hz we are familiar with. This is in response to pressure from organizations such as the BBC and NHK who have gained significant experience of spatio-temporal interactions in the Super HiVision project [5]. We will consider this aspect a little more in Section 13.2.3.

Compression performance

Firstly let us take a look at the general interactions between spatial resolution and compression performance and dispel the myth that HD is always better than SD. Considering Figure 13.1, this shows how the RD curves for different formats overlap as the bit rate increases, clearly indicating threshold bit rates where it is better to switch

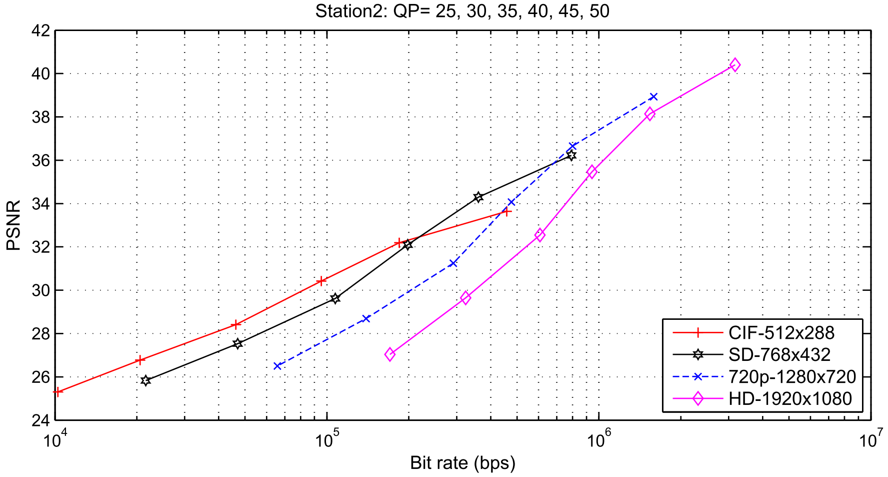

FIGURE 13.1

Does increased resolution mean increased quality?

format in order to ensure the best quality. The problem with this is that it is highly content dependent and most codecs will not provide seamless switching. However, adaptive methods are under investigation in some laboratories and simple approaches have already appeared in some streaming products. The implication of this is quite simple—there is no point increasing spatial resolution unless the compression system can preserve the perceptual benefits associated with that resolution.

A primary and stated goal of HEVC is to provide efficient compression for resolutions beyond HDTV. However, it is interesting to note that the subjective evaluations conducted in HEVC were primarily at HDTV resolution and the metrics used were mainly based on PSNR and subjective tests.

Hanhart et al. [7] have however reported work on evaluation of HEVC at extended spatial resolutions, and the results are encouraging. Subjective tests were conducted at the EPFL laboratory on a limited dataset of 4k content at 24 and 30 fps (one of the issues currently is that data is limited) and the results clearly indicate that greater relative savings are possible for resolutions beyond HDTV. They reported a bit rate reduction of between 51% and 74% for natural content based on MOS scores compared to H.264/AVC. Interestingly, the corresponding PSNR scores were 28% and 38%, highlighting once more the shortcomings of PSNR measurements.

13.2.3 Temporal resolution

Why rendition of motion is important

We saw in Chapter 2 that motion is important. Our retina is highly tuned to motion, even in the periphery, and our visual system is designed to track motion through retinal and head movements. In terms of video content, motion is conveyed through a combination of frame rate and shutter angle. Frame rates for film were standardized at 24 fps in the 1920s and TV frame rates have been fixed at 25 or 30 fps (using 50 or 60 fields) since the BBC adopted Marconi’s 405-line system in 1937. These rates were appropriate at the time, were compatible with mains frequencies and provided an excellent trade-off between bandwidth, perceived smoothness of motion, and the elimination of flicker. However, as screen sizes grow, along with spatial resolutions, there has developed a significant mismatch between frame rate and spatial resolution that few people appreciate [6]. In an attempt to mitigate this, TV manufacturers have introduced sophisticated up-sampling at the receiver to smooth motion at 300 or 600 Hz. While this approach can reduce peripheral flicker for larger screens, it does little or nothing to improve the motion blur.

The portrayal of motion for a given frame rate is a trade-off between long shutters that create motion blur and short shutters that, in cases of lower frame rates, can cause jerky motion and aliasing. We explore this further below.

Frame rates and shutter angles—static and dynamic resolutions

Considering Figure 13.2, this shows, for a fixed frame rate, the effect that shutter angle has on motion blur over a sequence of three frames. The difference between high frame rate capture and conventional rates is also demonstrated by Figure 13.3,

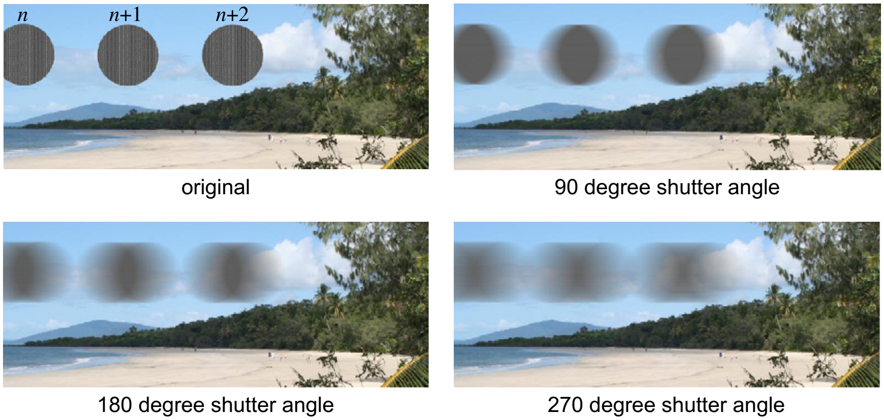

FIGURE 13.2

Frame rate and shutter angle. The figure illustrates the effect of a ball thrown from left to right across a static scene, captured at various shutter angles.

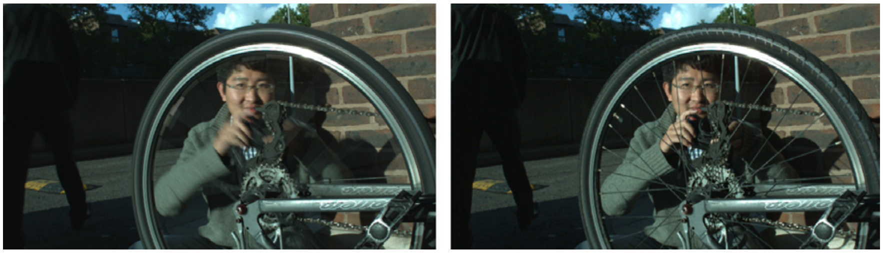

FIGURE 13.3

The influence of frame rate (*Outdoor* sequence). Left: 25 fps. Right: 600 fps.

where the effects of motion blur on the hands and spokes can clearly be seen. To reduce motion blur, we could shorten the shutter time. However, this leads to aliasing effects and jitter, especially in areas that are not being tracked by the viewer.

Let us define dynamic resolution as the effective spatial resolution of a format in the presence of motion (see Example 13.1). The relationship between resolution and frame rate for the case of TV systems is shown in Figure 13.4. This illustrates an important relationship between static and dynamic resolution. The top dashed line shows the frame rate needed to preserve dynamic resolution, extrapolated from a baseline of 50 Hz SDTV, as the spatial resolution is increased. It can be seen that, even at HDTV spatial resolutions, there is a significant reduction in dynamic resolution. Taking this further, to the 8k UHDTV image format, this extrapolation indicates that the corresponding frame rate should be something approaching 600 fps! In contrast,

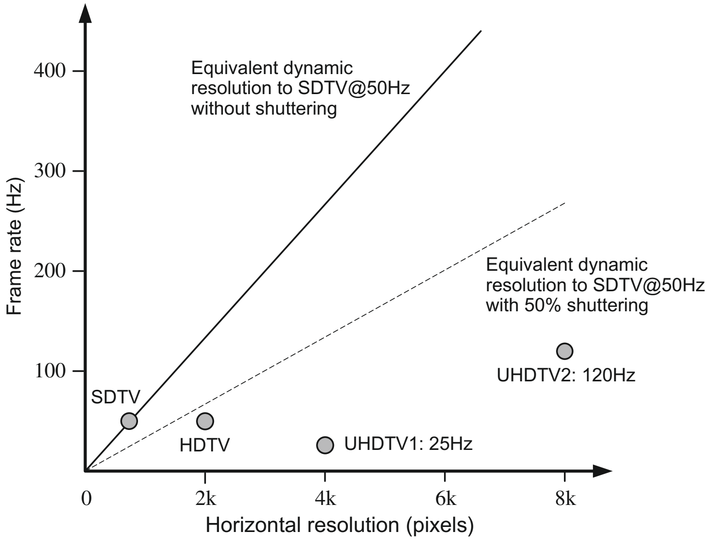

FIGURE 13.4

Static vs dynamic resolution—the relationship between frame rate and spatial resolution (adapted from an original by Richard Salmon, BBC).

UHDTV has been standardized with a maximum frame rate of 120 Hz, which is clearly well below this line.

Shorter shutter times and higher frame rates hold the potential to significantly improve the rendition of motion, especially for large screens where viewers will engage much more in wide angle tracking. They can reduce motion judder and the occurrence of background distractors caused by motion artifacts and flicker; they also provide an increased perception of depth in the image.

Compression methods and performance

The question to ask is how does frame rate interact with compression performance? A number of factors come into play here. Firstly, the temporal correlation between adjacent frames increases with frame rate. This produces smaller motion vector magnitudes and more correlated motion vectors. Secondly, with reduced frame periods, motion will more closely fit the translational model assumed in most codecs. Thus residual energy will be reduced and fewer bits will be required to code it. Higher frame rates, however, can introduce more high frequency spatial detail which can be harder to code in some cases. We can see these effects in Figures 13.5 and 13.6.

We would thus expect that the bit rate increase would be lower than the frame rate increase. Initial results in the Bristol-BBC Immersive Technology Laboratory indicate approximately a 2:1 ratio between the frame rate increase and the bit rate increase using H.264/AVC or HEVC. This is however highly content dependent.

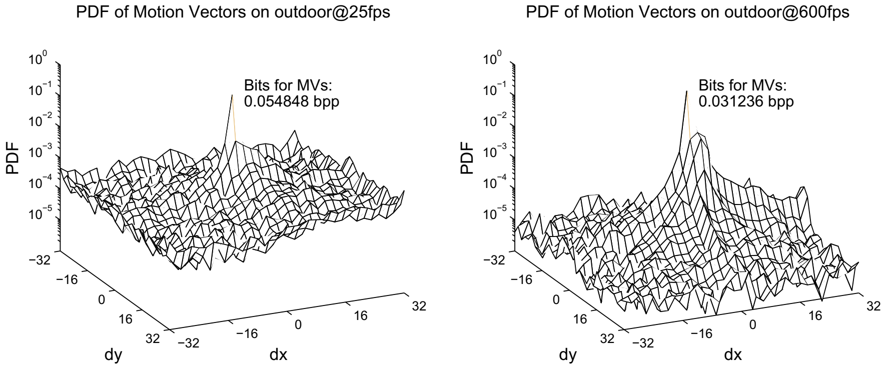

FIGURE 13.5

Motion vector PDF comparisons for 25 fps and 600 fps.

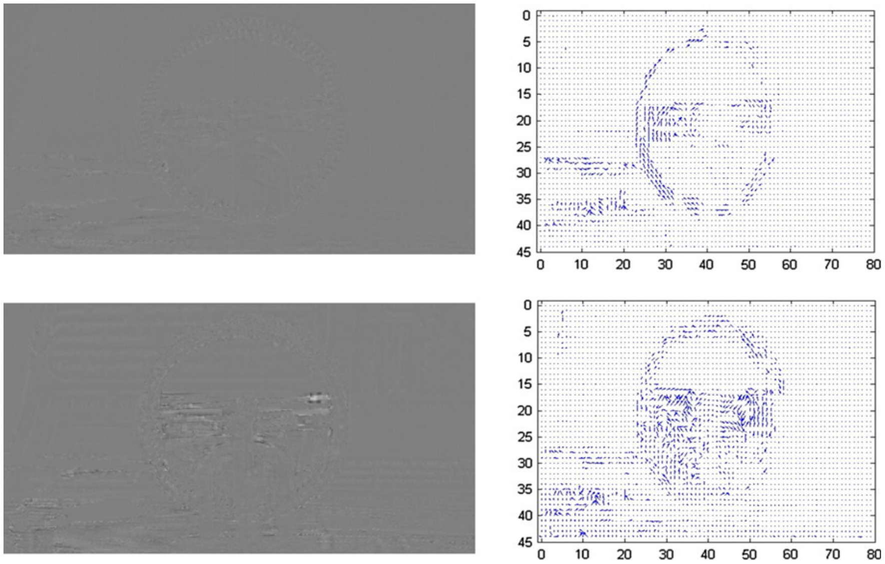

FIGURE 13.6

Motion vectors and residual content for high frame rate content (*Mouse* sequence). Top: 400 fps. Bottom: 25 fps. Left: DFD signal. Right: spatial distribution of motion vectors.

Example 13.1 (Static and dynamic resolution)

Consider the case of an SD format with 720 horizontal pixels, compared with an 8k UHDTV system with 7680 horizontal pixels, both at 30 fps. What is the effective dynamic resolution of both formats for the case of an object traversing the width of the frame in 4 s?

Solution. Clearly the horizontal static resolution of the UHDTV format is over 10 times that of the SD format. However, when an object tracks horizontally across the image frame in 4 seconds, then this corresponds to a blurring across 6 pixels per frame for the SD image, compared with 64 pixels per frame for the UHD image. If the screens are the same size then the dynamic resolution of both formats will be the same (i.e. 120 pixels). However, if the UHD screen is larger and/or the viewer is closer to it, then its perceived dynamic resolution will actually be significantly worse than that for SD, due to the increased angle subtended at the retina by each pixel. Clearly this is not a great situation for our enhanced format.

Let us say, for example, that the viewing angle for the UHDTV screen is 100° and that for the SD screen is 20°. The dynamic resolution for each case is then given by:

$$\begin{aligned} \text{SD}: \ 120/20 &= 6 \text{ pixels per degree} \\ \text{UHD}: 120/100 &= 1.2 \text{ pixels per degree} \end{aligned}$$

The expression for the viewer-screen-normalized dynamic resolution in pixels per degree is thus:

$$r_\theta = \frac{f}{2v_o \tan^{-1}\left(\frac{W}{2D}\right)}$$

where v_o is the object velocity in terms of frame widths per second, f is the frame rate, W is the width of the screen, and D is the viewing distance.

13.2.4 Dynamic range

Why dynamic range is important

It has been shown that increasing the dynamic range of content when displayed, both in terms of bit depth and screen bright to dark range, can make video appear more lifelike, increasing the perceived depth in the image and even imparting a sense of 3-D. The human visual system (HVS) can cover between 10 and 14 stops without adaptation and with adaptation it can accommodate 20 stops for photopic vision, and more for scotopic vision.

A typical modern flat panel TV, however, has a dynamic range capability of only around 8 stops. In order to fully exploit this limited dynamic range, methods such as black-stretch and pre-gamma mappings are used. The grading process in production also lifts detail in darker areas, and compresses highlights into the normal displayed range. New displays such as the SIM2 HDR47E offer the prospect of getting much closer to the capabilities of the HVS—delivering from approximately 0.4 to 4000 cd/m^2. This is achieved using a high dynamic range (HDR) (but low resolution) back-light LED array, in conjunction with a conventional LCD panel. Commercial cameras (e.g. the RED EPIC) are emerging with "HDR" capability but this is achieved using a complementary, delayed exposure for each frame.

The technological challenges of understanding and realizing HDR video are still significant. We must, for example, understand the masking effects of this type of content as well as the impact of the viewing environment. We also need to better understand the impact of compression on the immersive properties of such content.

Perceptual compression methods and performance

Several examples of approaches that exploit human perception during coding have been reported but, while these hold potential for significant coding gains, few have been adopted in practice. Naccari and Pereira [24] propose a perceptual video coding architecture incorporating a JND model based on spatio-temporal masking used for rate allocation and rate–quality optimization. They suggest the use of decoder-based JND modeling to perceptually allocate the available rate. The performance reported for their approach indicates an average bit rate reduction of up to 30% compared to H.264/AVC High profile at the same quality level.

Zhang et al. proposed a perception-based quantization method for high dynamic range content that exploits luminance masking [26] in the HVS in order to enhance the performance of HEVC. Extending the idea in Ref. [25] a profile scaling is proposed, based on a tone-mapping curve computed for each HDR frame. The quantization step is then perceptually tuned on a Transform Unit (TU) basis. The proposed method has been integrated into the reference codec considered for the HEVC range extensions and its performance compared with the case without perceptual quantization. Using HDR-VDP-2 image quality metric, a bit rate reduction of 9% was reported.

An example of the quantization curve is shown in Figure 13.7.

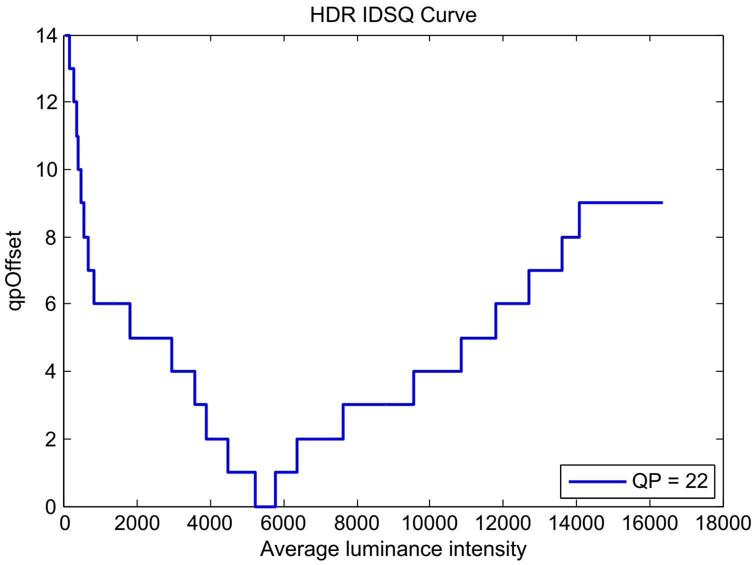

FIGURE 13.7

Intensity-dependent quantization for HDR extensions to HEVC.

13.2.5 Parameter interactions and the creative continuum

So what is required to enable us to create an optimum viewing experience? In essence, we need:

- An understanding of the influence of, and interactions within, the extended visual parameter space.
- Generalized psycho visual measures of quality, separating narrative from format/ medium, and environment.
- Use of these measures to characterize the impact of distortion, dynamic range, color palette, spatial and temporal resolution.
- Representations, formats, acquisition, display, and delivery processes that preserve or enhance immersive properties.
- Parameter optimization and adaptation methods that take account of the influences of content and use-context-dependent manner.
- Compression methods that minimize bit rate while preserving immersive properties.
- An understanding of the impact of cinematographic methods, e.g. shot length, framing, camera placement, camera and subject motion.
- New acquisition, delivery and display technologies.

A question that naturally arises is—should we maintain the idea of a fixed parameter set at all, or should we be looking at new formats that move away from the notion of fixed frame rates and resolutions? Perhaps the most important message is that we should no longer consider acquisition, production, compression, display, and assessment as independent processes, but should instead view these as a *Creative Continuum*. This intimately links production, cinematography, and acquisition to delivery, display, consumption, and quality assessment. Rather than an end-to-end delivery system we should instead consider a set of continuous relationships within an extended parameter space as indicated in Figure 13.8, where:

- The *Experience* must maximize engagement with the displayed content. Measuring it is essential if we are to fully understand the influences of the delivery and display processes.
- The *Delivery* processes must ensure that the content is delivered in a manner that preserves the immersive properties of the format.
- The *Capture* processes must employ formats and parameters that enable immersive experiences.
- The *Creation* processes must be matched to the acquisition formats to ensure optimum exploitation of the format in terms of sets, lighting, etc.

Importantly, an ability to measure the quality of the experience, not simply in terms of video quality, but in terms of how engaging it is, is critical in optimizing other aspects of the delivery chain.

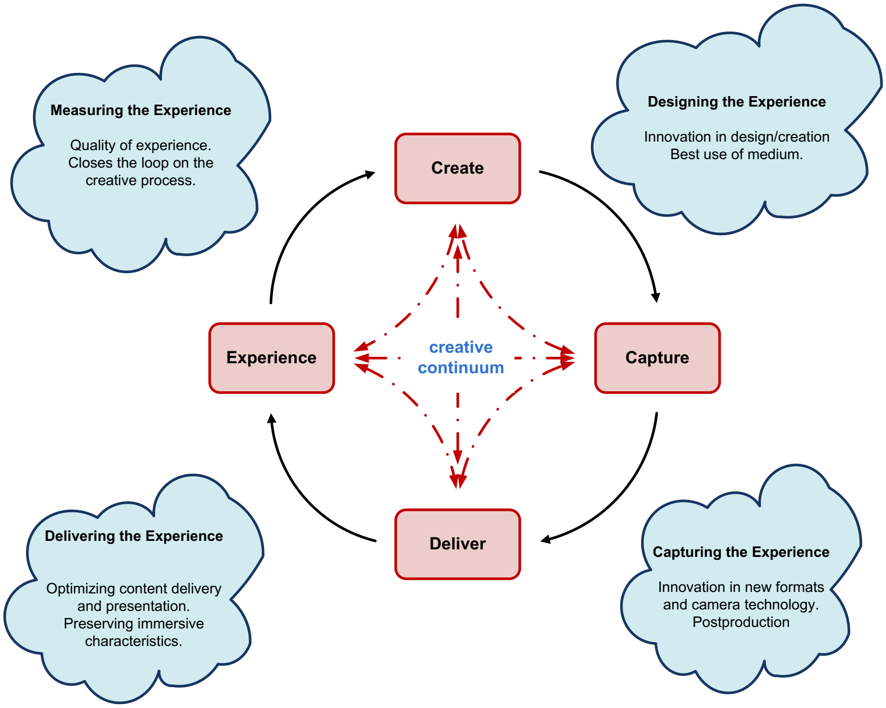
Measuring the Experience
Quality of experience. Closes the loop on the creative process.
Designing the Experience
Innovation in design/creation Best use of medium.
Create
Experience
creative continuum
Capture
Deliver
Delivering the Experience
Optimizing content delivery and presentation. Preserving immersive characteristics.
Capturing the Experience
Innovation in new formats and camera technology. Postproduction

FIGURE 13.8

The creative continuum.

13.3 Challenges for compression

While the demand for new video services will, to some extent, be addressed through efficiency improvements in network and physical layer technology, the role of video compression remains of central importance in ensuring that content is delivered at an acceptable quality while remaining matched to the available bandwidth and variable nature of the channel. In the context of these growing demands for more content, at higher qualities and in more immersive formats, there is an urgent need for transformational solutions to the video compression problem, and these need to go well beyond the capabilities of existing standards.

All major video coding standards since H.261 have been based on incremental improvements to the hybrid motion-compensated block transform coding model. While this approach has produced impressive rate–distortion improvements, this author believes that more disruptive techniques can provide substantial additional gains. As we have seen, H.264/AVC is based on the picture-wise processing and waveform-based coding of video signals. HEVC is a generalization of this approach offering gains through improved intra-prediction, larger block sizes, more flexible ways of decomposing blocks for inter- and intra-coding, and better exploitation of long term correlations and picture dependencies.

Example 13.2 (The delivery challenge)

Let us for now put the above discussion in context by considering the effect that the immersive parameter set might have on bit rate. Let us assume that, for given content, optimum immersion is provided by the following fixed parameter values:

- Frame rate: 200 frames per second.
- Spatial resolution: UHDTV resolution at 7680 × 4320 pixels.
- Dynamic range: requiring 16 bits of dynamic range in *R*, *G*, and *B*.

Calculate the uncompressed bit rate for this format and compare its compression requirements with those for existing HDTV broadcasting and internet streaming systems.

Solution. Assuming no color sub-sampling, the overall uncompressed bit rate would be around 3×10^{11} bps. This is approximately 100 times greater than HDTV at 1080p50, 50,000 times greater than a typical current broadcast compressed bit rate and 100,000 times greater than high quality internet HD delivery.

Consider now Example 13.2. This clearly illustrates that, despite video compression advances providing a 50% reduction in bit rate every 10 years, such immersive parameter sets demand far greater bandwidths than are currently available. We thus not only need to discover the optimum video parameters, but also to understand the perceptual implications of compression in order to specify a compression ratio that, while exploiting psychovisual redundancy, exploits masking effects to minimize bit rate. It is clear that conventional compression techniques are unlikely to meet these requirements and hence new, perceptually driven methods will be necessary.

New approaches should be predicated on the assumption that, in most cases, the target of video compression is to provide good subjective quality rather than to minimize the error between the original and coded pictures. It is thus possible to conceive of a compression scheme where an analysis/synthesis framework replaces the conventional energy minimization approach. Such a scheme could offer substantially lower bit rates through region-based parameterization and reduced residual and motion vector coding. New and alternative frameworks of this type are beginning to emerge, where prediction and signal representations are based on a parametric or data-driven model of scene content. These often invoke a combination of waveform coding and texture replacement, where computer graphic models are employed to replace target textures at the decoder.

Such approaches can also benefit from the use of higher order motion models for texture warping and mosaicing or through the use of contextual scene knowledge. Preliminary work to date has demonstrated the potential for dramatic rate–quality improvements with such methods. It is clear however that a huge amount of research is still needed in order to yield stable and efficient solutions, before we can fully exploit the potential of these new approaches. For example, mean square error is no longer a valid objective function or measure of quality and emphasis must shift from rate–distortion to rate–quality optimization, demanding new embedded perceptually driven quality metrics (see Chapter 10).

The choice of texture analysis and synthesis models, alongside meaningful quality metrics and the exploitation of long-term picture dependencies, will be key if an effective and reliable system is to result. Furthermore, in order to rigorously evaluate these new methods, challenging test data (including high dynamic range, high resolution and high frame rate content) will be required. The hypothesis that underpins this approach is:

> *If consistent, representative spatio-temporal descriptions can be obtained for static and dynamic textures and a valid perceptual distortion measure can be defined, then these descriptions can be employed within a rate–quality optimized parametric framework to dramatically improve video compression performance.*

13.4 Parametric video compression

Parametric video compression falls into the class of perception-based compression methods [16] and uses an analysis/synthesis framework rather than the conventional energy minimization approach. Parametric methods are employed to describe texture warping and/or synthesis as reported by Ndjiki-Nya et al. [8,9], Bosch et al. [10], Byrne et al. [11,12], Stojanovic et al. [22], Zhu et al. [23], and Zhang and Bull [13–15]. Although this approach has significant potential, a number of problems still need to be resolved.

We now focus on the work of Zhang and Bull [15], as an example of this class of codec. Their work combines dynamic and static texture synthesis using robust region segmentation, classification, and quality assessment. In a similar manner to that of other authors, the approach is hosted within a conventional block-based codec which is invoked wherever texture synthesis fails or is inappropriate. Textured regions in images or video frames have arbitrary shapes and these must first be segmented into homogeneous regions, each sharing similar texture properties. The authors employ a spatial texture-based image segmentation algorithm using an enhanced watershed transform [18] to obtain reliable textured regions.

To provide good texture coding performance, textures must be separated into static and dynamic classes. Classification rules are therefore employed to differentiate between them as well as to distinguish non-textured regions. Classification is based on an analysis of wavelet subband coefficients and motion information within each texture region using features from the Dual-tree Complex Wavelet Transform (DT-CWT) [19].

Texture synthesis is commonly employed in computer graphics, where synthetic content is generated using spatial and temporal methods [20]. Zhang and Bull apply different texture synthesis methods to static and dynamic textured regions according to their statistical and spectral characteristics. Static textures are coded using a perspective model, while dynamic textures are coded using a modified version of the synthesis method of Doretto et al. [21]. The primary benefits of this approach are that residual coding is generally not required and only side information (motion parameters and warping/synthesis maps) needs to be encoded.

One of the most challenging problems for synthesis-based coding is to create a reliable in-loop quality assessment measure with which to estimate subjective quality and detect any possible coding artifacts. Existing distortion-based metrics, such as peak signal-to-noise ratio (PSNR) and structural similarity (SSIM), are known to be inappropriate for this type of compression and other perceptual measures are too complex and perform poorly. A meaningful and simple objective video metric (AVM) was thus developed by Zhang and Bull which provides reliable quality estimates of synthesized textures, with low complexity due to extensive reuse of coding parameters (e.g. DT-CWT coefficients and motion vectors).

The overall architecture of the parametric compression scheme is shown in Figure 13.9. Figure 13.10 shows the results of the texture classification stage and

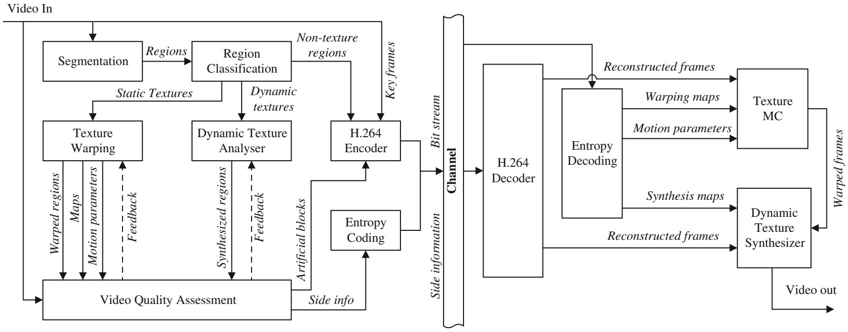

FIGURE 13.9

Parametric video compression architecture (Reproduced with permission from Zhang and Bull [15]).

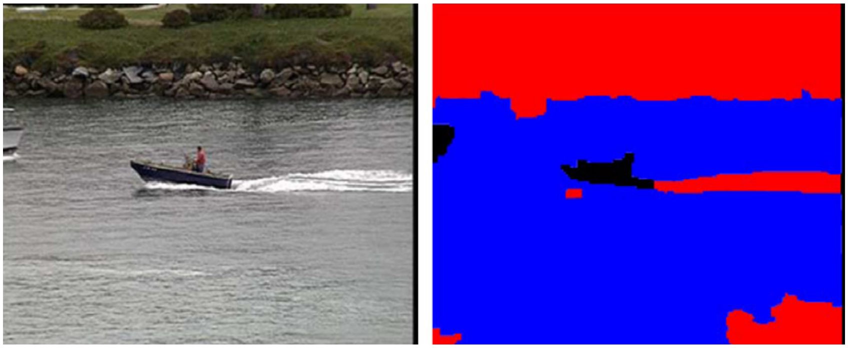

FIGURE 13.10

Texture classification results for parametric video coding. Left: original frame. Right: classified regions for coding: Red—static, blue—dynamic, black—structural, or non-textured (Reproduced with permission from Ref. [15]). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this book.)

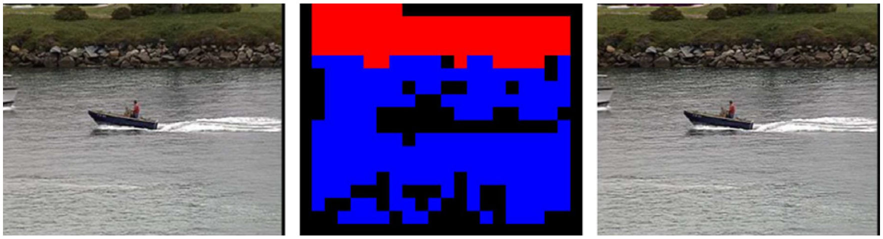

FIGURE 13.11

Coding results from parametric video coding. Left: H.264 frame. Right: proposed method. Center: coding mode employed (from Ref. [15]).

Figure 13.11 gives an example of a synthesized frame compared with H.264/AVC. The results presented in Ref. [15] indicate savings up to 60% over and above those offered by the host codec (H.264/AVC) alone for the same subjective quality.

13.5 Context-based video compression

Certain applications, such as sports broadcasting, are highly demanding due to their activity levels and the perceptual quality levels required. Such content is however often captured in a closed and well-defined environment such as a sports arena. The context-based coding method of Vigars et al. [17] exploits this to produce a hybrid CODEC framework which is able to exploit prior knowledge about the geometry of a scene. It applies a planar-perspective motion model to rigid planar and quasi-planar regions of the video. Unlike mosaic and sprite coding, this enables independent planar regions to be coded accurately without the need for registration, blending, or residual coding. It works well for both global and local perspective distortions. Non-planar or otherwise non-conforming regions are coded as normal by a host CODEC such as H.264/AVC or HEVC.

Figure 13.12 shows the matching process across two frames based on feature point matching. Firstly prior knowledge of the environment in which the video is captured is encapsulated into a scene model. This can be done off- or on-line. Feature matching is then used to detect salient points in each frame, and to compose descriptors of neighborhoods for matching and tracking between images. This is then used to track known planar structures in the video. Vigars et al. use the SURF algorithm for this.

Planar structures in each frame of the video are thus located, facilitating perspective motion estimation between them. Foreground segmentation is then used to separate regions of the video which do not conform to the planar model. These may be foreground objects, reflective surfaces, dynamic textures, or other regions which do not conform. Such foreground regions are processed by the host codec. Vigars et al. report savings up to 50% for this method compared with conventional H.264/AVC encoding (see Figure 13.13).

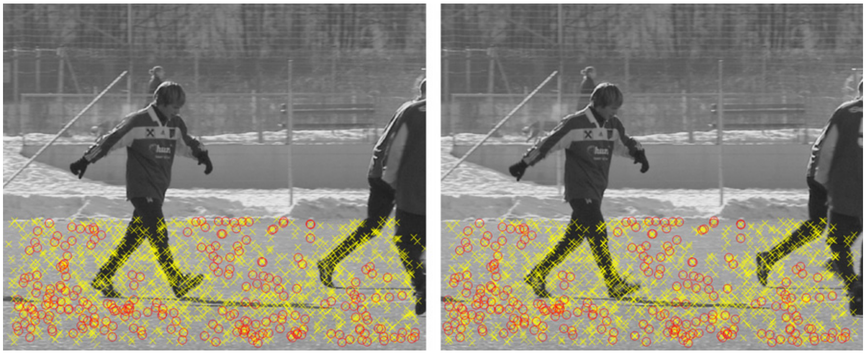

FIGURE 13.12

SURF-based feature matching for context-based coding of planar regions. Left: reference frame. Right: current frame. Circles represent RANSAC inliers used for planar motion modeling.

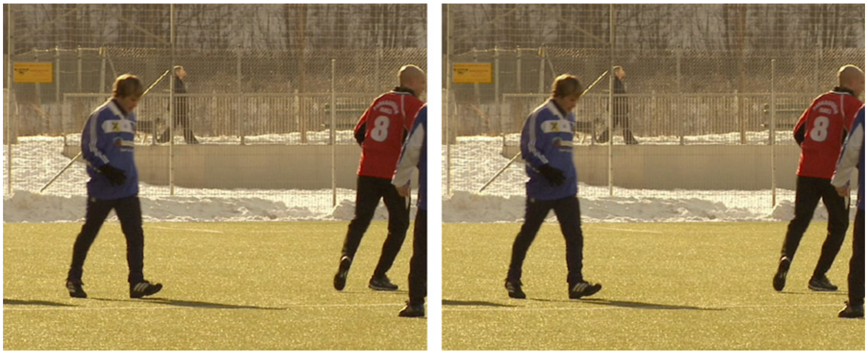

FIGURE 13.13

Context-based video compression results. Left: H.264 coded. Right: context-based coding.

13.6 Summary

This chapter has summarized the likely future demands and challenges for video compression. It has been postulated that the increased bit rates demanded will, for many applications, outstrip those provided by advances in network technology and conventional compression methods. The hybrid codec, that has served us well for the past 30 years, will no doubt continue to dominate for the foreseeable future, but could be enhanced through the exploitation perception-based methods such as those described here.

We have seen that the Human Visual System (HVS) exhibits non-linear sensitivities to the distortions introduced by lossy image and video coding. There are several factors that contribute to this, including luminance masking, contrast masking, and spatial and temporal frequency masking. Coding methods in the future must attempt to exploit these phenomena to a greater extent. It is thus proposed that new approaches to compression, possibly based on increased contextual knowledge and the use of analysis/synthesis models, could provide the next step toward delivering content at the required quality for future immersive applications.

References

[1] Bureau of Labor Statistics, American Time Use Survey, 2010.

[2] <http://www.wireless-world-research.org/fileadmin/sites/default/files/publications/-Outlook/Outlook4.pdf>.

[3] Cisco Visual Networking Index: Forecast and Methodology, 2011–16 (update 2012–17). <http://www.cisco.com/en/US/netsol/ns827/networking_solutions_sub_solution.html>.

[4] Recommendation ITU-R BT.2020 (08/2012), Parameter Values for Ultra-High Definition Television Systems for Production and International Programme Exchange, ITU-R, 2012.

[5] S. Sakaida, N. Nakajima, A. Ichigaya, M. Kurozumi, K. Iguchi, Y. Nishida, E. Nakasu, S. Gohshi, The super HiVision codec, in: Proceedings of the IEEE International Conference on Image Processing, 2007, pp. I-21–I-24.

[6] R. Salmon, M. Armstrong, S. Jolly, Higher Frame Rates for More Immersive Video and Television, BBC White Paper WHP209, BBC, 2009.

[7] P. Hanhart, M. Rerabek, F. DeSimone, T. Ebrahimi, Subjective quality evaluation of the upcoming HEVC video compression standard, in: Applications of Digital Image Processing XXXV, vol. 8499, 2012.

[8] P. Ndjiki-Nya, C. Stuber, T. Wiegand, Texture synthesis method for generic video sequences, in: IEEE International Conference on Image Processing, vol. 3, 2007, pp. 397–400.

[9] P. Ndjiki-Nya, T. Hinz, C. Stuber, T. Wiegand, A content-based video coding approach for rigid and non-rigid textures, in: IEEE International Conference on Image Processing, 2006, pp. 3169–3172.

[10] M. Bosch, M. Zhu, E. Delp, Spatial texture models for video compression, in: IEEE International Conference on Image Processing, 2007, pp. 93–96.

[11] J. Byrne, S. Ierodiaconou, D.R. Bull, D. Redmill, P. Hill, Unsupervised image compression-by-synthesis within a JPEG framework, in: IEEE International Conference on Image Processing, 2008, pp. 2892–2895.

[12] S. Ierodiaconou, J. Byrne, D. Bull, D. Redmill, P. Hill, Unsupervised image compression using graphcut texture synthesis, in: IEEE International Conference on Image Processing, 2009, pp. 2289–2292.

[13] F. Zhang, D. Bull, N. Canagarajah, Region-based texture modelling for next generation video codecs, in: IEEE International Conference on Image Processing, 2010, pp. 2593–2596.

[14] F. Zhang, D. Bull, Enhanced video compression with region-based texture models, in: Picture Coding Symposium (PCS), 2010, pp. 54–57.

[15] F. Zhang, D. Bull, A parametric framework for video compression using region-based texture models, IEEE Journal of Selected Topics in Signal Processing 6 (7) (2011) 1378–1392.

[16] J. Lee, T. Ebrahimi, Perceptual video compression: a survey, IEEE Journal of Selected Topics in Signal Processing 6 (2012) 684–697.

[17] R. Vigars, A. Calway, D. Bull, Context-based video coding, in: Proceedings of the IEEE International Conference on Image Processing, 2013, pp. 1953–1957.

[18] R. O'Callaghan, D. Bull, Combined morphological-spectral unsupervised image segmentation, IEEE Transactions on Image Processing 14 (1) (2005) 49–62.

[19] N. Kingsbury, Complex wavelets for shift invariant analysis and filtering of signals, Journal of Applied and Computational Harmonic Analysis 10 (3) (2001) 234–253.

[20] V. Kwatra, A. Schodl, I. Essa, G. Turk, A. Bobic, Graphcut textures: Image and video synthesis using graph cuts, in: Proceedings of the SIGGRAPH, ACM, 2003, pp. 277–286.

[21] G. Doretto, A. Chiuso, Y. Wu, S. Soatto, Dynamic textures, International Journal of Computer Vision 51 (2) (2003) 91–109.

[22] A. Stojanovic, M. Wien, J.-R. Ohm, Dynamic texture synthesis for H.264/AVC inter coding, in: IEEE International Conference on Image Processing, 2009, pp. 1608–1611.

[23] C. Zhu, X. Sun, F. Wu, H. Li, Video coding with spatio-temporal texture synthesis and edge-based inpainting, in: Proceedings of the ICME (2008) 813–816.

[24] M. Naccari, F. Pereira, Advanced H.264/AVC-based perceptual video coding: architecture, tools, and assessment, IEEE Transactions on Circuits and Systems for Video Technology 21 (6) (2011) 766–782.

[25] M. Naccari, M. Mrak, D. Flynn, A. Gabriellini, Improving HEVC compression efficiency by intensity dependent spatial quantisation, JCTVC-J0076, 10th Meeting, Stockholm, SE, July 2012.

[26] Y. Zhang, M. Naccari, D. Agrafiotis, M. Mrak, D. Bull, High dynamic range video compression by intensity dependent spatial quantization, in: Picture Coding Symposium, December 2013.