CHAPTER 9

Conclusion

The increasing demand for compressed digital video and the associated computational complexity, the availability and evolution of video coding standards, the restrictions of low-power computing devices, particularly in mobile environment, the requirements of increased speed and efficient utilization of resources, the desires for the best visual quality possible on any given platform, and above all, the lack of a unified approach to the considerations and analyses of the available tradeoff opportunities—these have been the essential motivating factors for writing this book. In this final chapter, we summarize the book's key points and propose some considerations for future development in the field.

Key Points and Observations

Based on the key points made in this book, the following observations can be made:

- Tradeoffs are possible among the various video measures, including the amount of compression, the visual quality, the compression speed, and the power consumption.
- Tuning of video encoding parameters can reveal these tradeoffs.
- Some coding parameters influence one video measure more than the others; depending on the application, optimization of certain measures may be favored over others.
- Analyzing the impact of various coding parameters on performance, power and quality is part of evaluating the strength of a video coding solution.
- Some video coding solutions are more suitable for certain types of video uses than for others, depending on the optimization performed and the parameters tuned.

- Being able to compare two video coding solutions is not only useful in ranking available solutions but also valuable in making informed choices.
- Sampling and scanning methods, picture rates, color space, chromaticity, and coding formats are among the parameters defined by the ITU-R digital video studio standards in three recommended specifications.
- Visual quality degradation is an option, owing to the natural tolerant characteristics of the human visual system (HVS) and the fact that the HVS is more sensitive to certain types of visual quality loss than others. Many video compression techniques and processes take advantage of this fact and trade quality for compression.
- Chroma subsampling is a common technique to take full advantage of the HVS sensitivity to color information. Many video usages make use of 4:2:0 chroma subsampling.
- Various techniques are available for digital video compression. Most international standards adopt transform-based spatial redundancy reduction, block-matching motion compensation-based temporal-redundancy reduction, and variable-length code-based spectral redundancy-reduction approaches for lossy predictive coding.
- International standards define a range of video applications in the domains of practical compression, communication, storage, broadcast, gaming, and so on. Standard video formats are essential for exchanging digital video among products and applications. The algorithms defined by the standards are implementable in practical hardware and software systems, and are common across multiple industries.
- Video compression is influenced by many factors, including noise present at the input, dynamic range of input pictures, picture resolution, artifacts, requirements for bit rate, frame rate, error resiliency, quality settings (constant or variable from picture to picture), algorithm complexity, platform capabilities, and so on.
- Lossy compression introduces some visual quality impairment, but owing to HVS limitations, a small amount of quality loss is not too objectionable. Common compression artifacts include quantization noise, blurring, blocking, ringing, aliasing, flickering, and so on. Quality is also affected by sensor noise at the video capture device, video characteristics such as spatial and temporal activities, amount and method of compression, number of passes or generations of compression, errors during transmission, and artistic visual effects introduced during post-production.

- The opinions of human viewers are the most important criteria in judging the visual quality of compressed videos, but the opinions are subjective, variable and cannot be repeated reliably.

- Objective measures such as PSNR and SSIM are widey used in the evaluation of video quality. Although they do not correlate perfectly with human experience, they provide a good estimate of visual quality. However, when judging the output of an encoder, such objective measures alone are not sufficient; the cost in terms of bits spent must also be considered.

- Many video quality-evaluation methods and metrics are offered in the literature, with varying levels of complexity. This is an active area of academic research as well as emerging ITU standards.

- Several encoding parameters can be tuned to trade video quality for performance and power consumption. Important parameters here include the bit rate, frame rate, and latentcy requirements; bit-rate control type, available buffer size, picture structure, and picture groups; motion parameters, number of reference pictures, motion vector precision, and motion vector search and interpolation methods; entropy coding type; number of encoding passes; and so on.

- Coding efficiency is determined in terms of the quality achieved in regard to the number of bits used. In the literature, *coding efficiency* is often used to mean performance. However, in this book, *performance* refers to the coding speed.

- Encoding speed is determined by several factors, including the platform and the video characteristics. Platform characteristics include the CPU and GPU frequencies, operating voltages, configurable TDP state, operating system power policy, memory bandwidth and speed, cache policy, disk access speed, I/O throughput, system clock resolution, graphics driver settings, and so on. Video characteristics include formats, resolutions, bit rate, frame rate, group of picture structure, and other parameters. Video scene characteristics include the amount of motion, details, brightness, and so on.

- Various parallelization opportunities can be exploited to increase encoding speed. However, costs of task scheduling and interprocess communication should be carefully considered. Parallelization approaches include data partitioning, task parallelization, pipelining, data parallelization, instruction parallelization, multithreading, and vectorization.

- Faster than real-time encoding is useful in applications such as video editing, archiving, recording, transcoding, and format conversion.

- Visual communication and applications such as screen cast require low-delay real-time encoding, typically on resource-constrained client platforms.
- Performance optimization implies maximal utilization of available system resources. However, power-aware optimization approaches maximize the resource utilization for the shortest possible duration and allow the system to go into deeper sleep states for as long as possible. This is in constrast to the traditional approach of only minimizing the idle time for a given resource.
- Algorithmic optimization, code and compiler optimization, and redundancy removal are noteworthy among the various performance-optimization approaches.
- There are several power-management points in the system, including the BIOS, the CPU, the graphics controller, the hard disk drive, the network, and the display. Memory power management is also possible, but is done infrequently.
- Typically, hardware-based power management involves the various CPU C-states and the render C-states. Software-based power management in the operating system or in the driver includes CPU core offline, CPU core shielding, CPU load balancing, interrupt load balancing, CPU and GPU frequency governing, and so on.
- On low-power platforms, special hardware units are typically needed for power management. Multiple points of power at various voltage levels constitute a complex system, for which fast and precise management of power requirements is handled by these special-purpose units.
- The goal of power management is to allow the processor to go into various sleep states for as long as possible, thereby saving power consumption.
- Total power consumption includes dynamic power and static leakage power; dynamic power depends on the operating voltage and frequency, while static power depends on the leakage current.
- A minimum voltage is required for the circuit to be operational, regardless of frequency change. The maximum frequency at which the processor can operate at minimum voltage ($F_{max}@V_{min}$) is the most power-efficient operating point. Increasing the frequency from this point also increases the dynamic power at a cubic rate and the static power at a linear rate. At this relatively high power, a power reduction can happen with an easy voltage–frequency tradeoff. Reducing the frequency below the most efficient point—that is, into the V_{min} region—reduces the dynamic power linearly while the static power remains constant, drawing a constant leakage current.

- Power optimization can be done at the architecture level, at the algorithm level, at the system integration level, and at the application level.

- On low-power platforms, some practical tradeoffs are possible among processor area, power, performance, visual quality, amount of compression, and design complexity. It may be necessary to sacrifice visual quality in favor of power savings on these platforms.

- The display consumes a considerable portion of the system power for video applications—in many cases, about a third of the system power. There are several display power-management techniques, including panel self-refresh, backlight control using the Intel display power-saving technology, ambient light sensors, and content adaptivity.

- Low-power software design considerations include intelligent power awareness, quality requirements, availability of hardware-acceleration capabilities, energy-efficient UI, code density and memory footprint, optimization of data transfer and cache utilization, parallel and batch processing, and so on.

- Low-power architectural considerations include combining system components on the same chip, optimized hardware-software interaction, workload migration from general-purpose to fixed-function hardware, CPU-GPU power sharing, reduced power core, uncore and graphics units, use of power islands, power-aware simulation and verification, and so on.

- Power-optimization approaches include running fast and turning off the processor, scheduling of tasks and activities, reducing wakeups, burst-mode processing, reducing CPU-GPU dependency and increasing parallelism, GPU memory bandwidth optimization, and power optimization for the display and the storage units.

- To measure power and performance for tradeoff analysis, you calibrate the system and select appropriate settings for operating temperature, voltage and frequency, cTDP, AC or DC power mode, OS power policy, display settings, driver settings, application settings, encoding parameters, and so on.

- The tradeoff analysis discussed in Chapter 8 attempts to fill a void that presently exists in comprehensive analysis methods. Particularly, it is important to examine the impact of tuning various parameters to obtain a better understanding of the costs and benefits of different video measures.

- Understanding the tradeoffs among performance, power, and quality is as valuable to architects, developers, validators, and technical marketers as it is to technical reviewers, procurers, and end-users of encoding solutions.

Considerations for the Future

The topics covered in this book will, I hope, inspire discussions that will take us into the future of video coding and related analysis. Some of the areas where future analysis is likely to extend are the following.

Enhanced Tools and Metrics for Analysis

Although it is possible to look into the details of performance, power, and quality in a given encoding test run, and to understand the relationships between them, it is not easy to determine why there is an increase or decrease for a given metric for that encoding run. This is especially difficult when comparing the results of two tests, likely generated by two different encoding solutions with different capabilities. Similarly, when comparing two metrics for the same run, it is not always obvious why there is an increase or decrease relative to each other. The complexity arises from the presence of many variables that react non-deterministically to changes in the system or video coding parameters, and that affect one another. Also, those influences are different for different video contents, applications, and usage scenarios. There needs to be study, as well as careful and time-consuming debugging, so we can understand these complex relationships.

Researchers are trying to come up with better video metrics, indices, and scores, particularly for visual quality, compression, performance, and power consumption. The analysis techniques are expected to adapt to more comprehensive future metrics. Eventually, there will be a single measure for all the benefits to weigh against a single measure for all the costs for video coding, and that this measure will be universally accepted for evaluation and ranking purposes. With the availability of the new metrics, enhanced benchmarking tools that consider all aspects of video coding are also expected.

Improved Quality and Performance

Techniques to improve visual quality with the same amount of compression will follow a path of continuous improvement. In the past couple of decades, this trend was evident in algorithms from MPEG-2 to AVC, and from AVC to HEVC. Similarly, optimization techniques for performance and power are improving at a rate even faster than that for quality improvement. Every generation of Intel processors is producing roughly 30 to 200 percent performance for the same power profile as compared to the previous generation for GPU-accelerated video coding and processing. Even low-power processors today are capable of supporting video applications that were only matters of dreams a decade ago. It is not far fetched to think that, with appropriate tradeoffs and optimizations, everyday video applications will have better visual quality despite platform limitations.

Emerging Uses and Applications

Wearables pose unique challenges when it comes to power consumption and performance, yet new uses on these emerging computing platforms are appearing every day. The role of video here is an open area of research. It embraces the notions of how to determine measures of goodness in video coding for these uses, how to quantify them, and which metrics to use.

The capabilities, uses, and requirements of video coding in driverless cars and radio-controlled drones are being assessed and developed. With their increasing processing abilities operating on resource-constrained systems, tradeoff analysis and optimization will play major roles in design and application. However, the methodologies and metrics for these uses are still open to definition.

Telemedicine, too, is in its infancy. Compression and communication technologies for high-resolution video are maturing to eventually reach flawless execution on handheld devices that can be used in remote surgical operations. Performance, power, and quality will be factors requiring tradeoffs in these scenarios as well.

Beyond Vision to the Other Senses

Of the five human senses, vision is considered the most important, but human experience is not complete with vision alone. Consequently, video is not the only data type for digital multimedia applications. Typically, audio and video are experienced together; touch and gestures are also rapidly evolving. So their measurement, understanding, and tuning will include audio, touch, and gesture. The relationships among these sense-based data types are complex and will require deep analysis, detailed study, and—ultimately—tradeoffs. This remains another active area of research.

As the challenges of the future are resolved, we will experience the true, full potential of the human senses.