1

Introduction

1.1 Why 4G?

Before we get into too much technical jargon such as 4G and so on, it would be interesting to take a moment to discuss iPhone, which was named *Time* magazine's Inventor of the Year in 2007, and which has a significant impact on many consumers' view of the capability and future of mobile phones. It is amazing to see the enthusiasm of customers if you visit Apple's stores which are always crowded. Many customers were lined up at the Apple stores nationwide on iPhone launch day (the stores were closed at 2 p.m. local time in order to prepare for the 6 p.m. iPhone launch) and Apple sold 270 000 iPhones in the first 30 hours on launch weekend and sold 1 million iPhone 3G in its first three days. There are also many other successful mobile phones produced by companies such as Nokia, Motorola, LG, and Samsung and so on.

It is interesting to observe that these new mobile phones, especially smart phones, are much more than just phones. They are really little mobile PCs, as they provide many of the key functionalities of a PC:

- A keyboard, which is virtual and rendered on a touch screen.
- User friendly graphical user interfaces.
- Internet services such as email, web browsing and local Wi-Fi connectivity.
- Built-in camera with image/video capturing.
- Media player with audio and video decoding capability.
- Smart media management tools for songs, photo albums, videos, etc..
- Phone call functionalities including text messaging, visual voicemail, etc.

However, there are also many features that some mobile phones do not yet support (although they may come soon), for example:

- Mobile TV support to receive live TV programmes.
- Multi-user networked 3D games support.
- Realistic 3D scene rendering.

4G Wireless Video Communications Haohong Wang, Lisimachos P. Kondi, Ajay Luthra and Song Ci

- Stereo image and video capturing and rendering.
- High definition visuals.

The lack of these functions is due to many factors, including the computational capability and power constraints of the mobile devices, the available bandwidth and transmission efficiency of the wireless network, the quality of service (QoS) support of the network protocols, the universal access capability of the communication system infrastructure and the compression and error control efficiency of video and graphics data. Although it is expected that the mobile phone will evolve in future generations so as to provide the user with the same or even better experiences as today's PC, there is still long way to go. From a mobile communication point of view, it is expected to have a much higher data transmission rate, one that is comparable to wire line networks as well as services and support for seamless connectivity and access to any application regardless of device and location. That is exactly the purpose for which 4G came into the picture.

4G is an abbreviation for Fourth-Generation, which is a term used to describe the next complete evolution in wireless communication. The 4G wireless system is expected to provide a comprehensive IP solution where multimedia applications and services can be delivered to the user on an 'Anytime, Anywhere' basis with a satisfactory high data rate and premium quality and high security, which is not achievable using the current 3G (third generation) wireless infrastructure. Although so far there is not a final definition for 4G yet, the International Telecommunication Union (ITU) is working on the standard and target for commercial deployment of 4G system in the 2010–2015 timeframe. ITU defined IMT-Advanced as the succeeding of IMT-2000 (or 3G), thus some people call IMT-Advanced as 4G informally.

The advantages of 4G over 3G are listed in Table 1.1. Clearly 4G has improved upon the 3G system significantly not only in bandwidth, coverage and capacity, but also in many advanced features, such as QoS, low latency, high mobility, and security support, etc.

Table 1.1 Comparison of 3G and 4G

	3G	4G
Driving force	Predominantly voice driven, data is secondary concern	Converged data and multimedia services over IP
Network architecture	Wide area networks	Integration of Wireless LAN and Wide area networks
Bandwidth (bps)	384K–2M	100 M for mobile 1 G for stationary
Frequency band (GHz)	1.8–2.4	2–8
Switching	Circuit switched and packet switched	Packet switch only
Access technology	CDMA family	OFDMA family
QoS and security	Not supported	Supported
Multi-antenna techniques	Very limited support	Supported
Multicast/broadcast service	Not supported	Supported

1.2 4G Status and Key Technologies

In a book which discusses multimedia communications across 4G networks, it is exciting to reveal part of the key technologies and innovations in 4G at this moment before we go deeply into video related topics, and before readers jump to the specific chapters in order to find the detail about specific technologies. In general, as the technologies, infrastructures and terminals have evolved in wireless systems (as shown in Figure 1.1) from 1G, 2G, 3G to 4G and from Wireless LAN to Broadband Wireless Access to 4G, the 4G system will contain all of the standards that the earlier generations have implemented. Among the few technologies that are currently being considered for 4G including 3GPP LTE/LTE-Advanced, 3GPP2 UMB, and Mobile WiMAX based on IEEE 802.16 m, we will describe briefly two of them that have wider adoption and deployment, while leaving the details and other technologies for the demonstration provided in Chapter 4.

1.2.1 3GPP LTE

Long Term Evolution (LTE) was introduced in 3GPP (3rd Generation Partnership Project) Release 8 as the next major step for UMTS (Universal Mobile Telecommunications System). It provides an enhanced user experience for broadband wireless networks.

LTE supports a scalable bandwidth from 1.25 to 20 MHz, as well as both FDD (Frequency Division Duplex) and TDD (Time Division Duplex). It supports a downlink peak rate of 100 Mbps and uplink with peak rate of 50 Mbps in 20 MHz channel. Its spectrum efficiency has been greatly improved so as to reach four times the HSDPA (High Speed Downlink Packet Access) for downlink, and three times for uplink. LTE also has a low latency of less than 100 msec for control-plane, and less than 5 msec for user-plane.

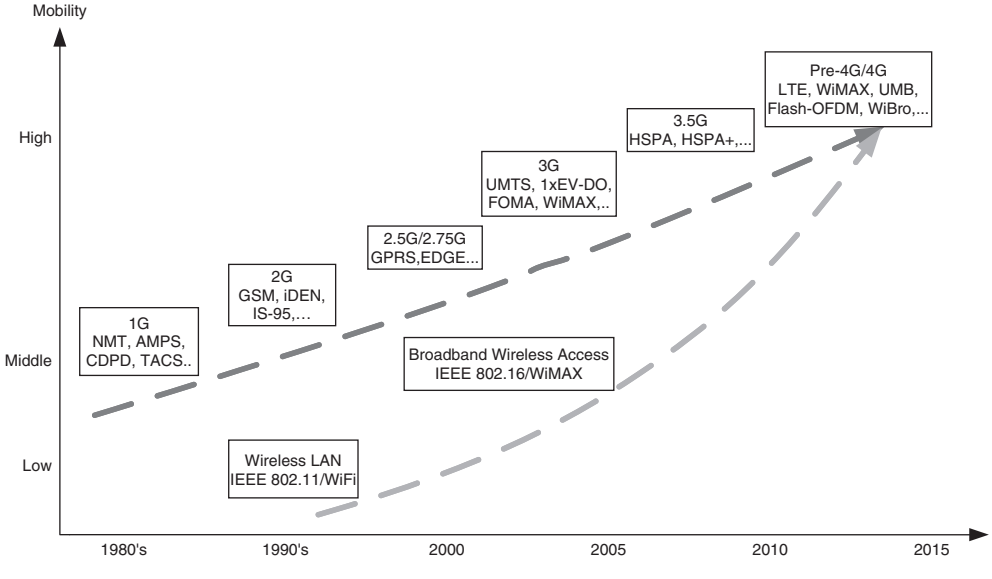

Figure 1.1 Evolving of technology to 4G wireless

It also supports a seamless connection to existing networks, such as GSM, CDMA and HSPA. For multimedia services, LTE provides IP-based traffic as well as the end-to-end Quality of Service (QoS).

LTE Evolved UMTS Terrestrial Radio Access (E-UTRA) has the following key air-interface technology:

- *Downlink based on OFDMA*. The downlink transmission scheme for E-UTRA FDD and TDD modes is based on conventional OFDM, where the available spectrum is divided into multiple sub-carriers, and each sub-carrier is modulated independently by a low rate data stream. As compared to OFDM, OFDMA allows for multiple users to access the available bandwidth and assigns specific time-frequency resources to each user, thus the data channels are shared by multiple users according to the scheduling. The complexity of OFDMA is therefore increased in terms of resource scheduling, however efficiency and latency is achieved.
- *Uplink based on SC-FDMA*. Single-Carrier-Frequency Domain Multiple Access (SC-FDMA) is selected for uplink because the OFDMA signal would result in worse uplink coverage owing to its weaker peak-to-average power ratio (PAPR). SC-FDMA signal processing is similar to OFDMA signal processing, so the parameterization of downlink and uplink can be harmonized. DFT-spread-OFDM has been selected for E-UTRA, where a size-M DFT is applied to a block of M modulation symbols, and then DFT transforms the modulation symbols into the frequency domain; the result is then mapped onto the available sub-carriers. Clearly, DFT processing is the fundamental difference between SC-FDMA and OFDMA signal generation, as each sub-carrier of an OFDMA signal only carries information related to a specific modulation system while SC-FDMA contains all of the transmitted modulation symbols owing to the spread process by the DTF transform.

LTE uses a MIMO (multiple input multiple output) system in order to achieve high throughput and spectral efficiency. It uses 2×2 (i.e., two transmit antennae at the base station and two receive antennae at the terminal side), 3×2 or 4×2 MIMO configurations for downlink. In addition, LTE supports MBMS (multimedia broadcast multicast services) either in single cell or multi-cell mode.

So far the adoption of LTE has been quite successful, as most carriers supporting GSM or HSPA networks (for example, AT&T, T-Mobile, Vodafone etc) have been upgrading their systems to LTE and a few others (for example, Verizon, China Telecom/Unicom, KDDI, and NTT DOCOMO, etc) that use different standards are also upgrading to LTE. More information about LTE, such as its architecture system, protocol stack, etc can be found in Chapter 4.

1.2.2 Mobile WiMAX

Mobile WiMAX is a part of the Worldwide Interoperability for Microwave Access (WiMAX) technology, and is a broadband wireless solution that enables the convergence of mobile and fixed broadband networks through a common wide area broadband radio access technology and flexible network architecture.

In Mobile WiMAX, a scalable channel bandwidth from 1.25 to 20 MHz is supported by scalable OFDMA. WiMAX supports a peak downlink data rate of up to 63 M bps and

a peak uplink data rate of up to 28 Mbps in the 10 MHz channel with MIMO antenna techniques and flexible sub channelization schemes. On the other hand, the end-of-end QoS is supported by mapping the service flows to the DiffServ code points of MPLS flow labels. Security is protected by EAP-based authentication, AES-CCM-based authentication encryption and CMAC and HCMAC based control message protection schemes. In addition, the optimized handover schemes are supported in Mobile WiMAX by latencies of less than 50 milliseconds.

In the Physical layer, the OFDMA air interface is adopted for downlink and uplink, along with TDD. FDD has the potential to be included in the future in order to address specific market opportunities where local spectrum regulatory requirements either prohibit TDD or are more suitable for FDD deployment. In order to enhance coverage and capacity, a few advanced features, such as AMC (adaptive modulation and coding), HARQ (hybrid automatic repeat request), and CQICH (fast channel feedback), are supported.

In Mobile WiMAX, QoS is guaranteed by fast air link, asymmetric downlink and uplink capability, fine resource granularity and flexible resource allocation mechanism. Various data scheduling services are supported in order to handle bursty data traffic, time-varying channel coditions, frequency-time resource allocation in both uplink and downlink on per-frame basis, different QoS requirements for uplink and downlink, etc.

The advanced features of Mobile WiMAX may be summarized as follows:

- A full range of smart antenna technologies including Beamforming, Space-Time Code, and Spatial Multiplexing is supported. It uses multiple-antennae to transmit weighted signals using Beamforming in order to reduce the probability of outage and improve system capacity and coverage; On the other hand, spatial diversity and fade margin reduction are supported by STC, and SM helps to increase the peak data rate and throughput.
- Flexible sub-channel reuse is supported by sub-channel segmentation and permutation zone. Thus resource reuse is possible even for a small fraction of the entire channel bandwidth.
- The Multicast and Broadcast Service (MBS) is supported.

So far more than 350 trials and deployments of the WiMAX networks have been announced by the WiMAX Forum, and many WiMAX final user terminals have been produced by Nokia, Motorola, Samsung and others. On the other hand, WiMAX has been deployed in quite a number of developing countries such as India, Pakistan, Malaysia, Middle East and Africa countries, etc.

1.3 Video Over Wireless

As explained in the title of this book, our focus is on the video processing and communications techniques for the next generation of wireless communication system (4G and beyond). As the system infrastructure has evolved so as to provide better QoS support and higher data bandwidth, more exciting video applications and more innovations in video processing are expected. In this section, we describe the big picture of video over wireless from video compression and error resilience to video delivery over wireless channels. More information can be found in Chapters 3, 5 and 7.

1.3.1 Video Compression Basics

Clearly the purpose of video compression is to save the bandwidth of the communication channel. As video is a specific form of data, the history of video compression can be traced back to Shannon's seminal work [1] on data compression, where the quantitative measure of information called *self-information* is defined. The amount of self-information is associated with the probability of an event to occur; that is, an unlikely event (with lower probability) contains more information than a high probability event. When a set of independent outcome of events is considered, a quantity called *entropy* is used to count the average self-information associated with the random experiments, such as:

$$H = \sum p(A_i) i(A_i) = -\sum p(A_i) \log p(A_i)$$

where $p(A_i)$ is the probability of the event A_i, and $i(A_i)$ the associated self-information.

Entropy concept constitutes the basis of data compression, where we consider a source containing a set of symbols (or its alphabet), and model the source output as a discrete random variable. Shannon's first theorem (or source coding theorem) claims that no matter how one assigns a binary codeword to represent each symbol, in order to make the source code uniquely decodable, the average number of bits per source symbol used in the source coding process is lower bounded by the entropy of the source. In other words, entropy represents a fundamental limit on the average number of bits per source symbol. This boundary is very important in evaluating the efficiency of a lossless coding algorithm.

Modeling is an important stage in data compression. Generally, it is impossible to know the entropy of a physical source. The estimation of the entropy of a physical source is highly dependent on the assumptions about the structure of the source sequence. These assumptions are called the model for the sequence. Having a good model for the data can be useful in estimating the entropy of the source and thus achieving more efficient compression algorithms. You can either construct a physical model based on the understanding of the physics of data generation, or build a probability model based on empirical observation of the statistics of the data, or build another model based on a set of assumptions, for example, aMarkov model, which assumes that the knowledge of the past k symbols is equivalent to the knowledge of the entire past history of the process (for the kth order Markov models).

After the theoretical lower bound on the information source coding bitrate is introduced, we can consider the means for data compression. Huffman coding [2] and arithmetic coding [3] are two of the most popular lossless data compression approaches. The Huffman code is a source code whose average word length approaches the fundamental limit set by the entropy of a source. It is optimum in the sense that no other uniquely decodable set of code words has a smaller average code word length for a given discrete memory less source. It is based on two basic observations of an optimum code: 1) symbols that occur more frequently will have shorter code words than symbols that occur less frequently; 2) the two symbols that occur least frequently will have the same length [4]. Therefore, Huffman codes are variable length code words (VLC), which are built as follows: the symbols constitute initially a set of leaf nodes of a binary tree; a new node is created as the father of the two nodes with the smallest probability, and it is assigned the sum of its offspring's probabilities; this new node adding procedure is repeated until the root of the tree is reached. The Huffman code for any symbol can be obtained by traversing the tree from the root node to the leaf corresponding to the symbol, adding a *0* to the code

word every time the traversal passes a left branch and a *1* every time the traversal passes a right branch.

Arithmetic coding is different from Huffman coding in that there are no VLC code words associated with each symbol. Instead, the arithmetic code generates a floating-point output based on the input sequence, which is described as follows: first, based on the order of the N symbols listed in the alphabet, the initial interval [0, 1] is divided into N ordered sub-intervals with their lengths proportional to the probabilities of the symbols. Then, the first input symbol is read and its associated sub-interval is further divided into N smaller sub-intervals. In the similar manner, the next input symbol is read and its associated smaller sub-interval is further divided, and this procedure is repeated until the last input symbol is read and its associated sub-interval is determined. Finally, this sub-interval is represented by a binary fraction. Compared to Huffman coding, arithmetic coding suffers from a higher complexity and sensitivity to transmission errors. However, it is especially useful when dealing with sources with small alphabets, such as binary sources, and alphabets with highly skewed probabilities. Arithmetic coding procedure does not need to build the entire code book (which could be huge) as in Huffman coding, and for most cases, using arithmetic coding can get rates closer to the entropy than using Huffman coding. More detail of these lossless data coding methods can be found in section 3.2.3. We consider video as a sequence of images. A great deal of effort has been expended in seeking the best model for video compression, whose goal is to reduce the spatial and temporal correlations in video data and to reduce the perceptual redundancies.

Motion compensation (MC) is the paradigm used most often in order to employ the temporal correlation efficiently. In most video sequences, there is little change in the content of the image from one frame to the next. MC coding takes advantage of this redundancy by using the previous frame to generate a prediction for the current frame, and thus only the motion information and the residue (difference between the current frame and the prediction) are coded. In this paradigm, motion estimation [5] is the most important and time-consuming task, which directly affects coding efficiency. Different motion models have been proposed, such as block matching [6], region matching [7] and mesh-based motion estimation [8, 9]. Block matching is the most popular motion estimation technique for video coding. The main reason for its popularity is its simplicity and the fact that several VLSI implementations of block matching algorithms are available. The motion vector (MV) is searched by testing different possible matching blocks in a given search window, and the resulting MV yields the lowest prediction error criterion. Different matching criteria, such as mean square error (MSE) and mean absolution difference (MAD) can be employed. In most cases, a full search requires intolerable computational complexity, thus fast motion estimation approaches arise. The existing popular fast block matching algorithms can be classified into four categories: (1) Heuristic search approaches, whose complexities are reduced by cutting the number of candidate MVs tested in the search area. In those methods, the choices of the positions are driven by some heuristic criterion in order to find the absolute minimum of cost function; (2) Reduced sample matching approaches [10, 11], whose complexities are reduced by cutting the number of points on which the cost function is calculated; (3) Techniques using spatio-temporal correlations [10, 12, 13], where the MVs are selected using the vectors that have already been calculated in the current and in the previous frames; (4) Hierarchical or multi-resolution techniques [12], where the MVs are searched in the low-resolution image and then refined in the normal resolution one.

Transform coding is the most popular technique employed in order to reduce the spatial correlation. In this approach, the input image data in the spatial domain are transformed into another domain, so that the associated energy is concentrated on a few decorrelated coefficients instead of being spread over the whole spatial image. Normally, the efficiency of a transform depends on how much energy compaction is provided by it. In a statistical sense, Karhunen-Loeve Transform (KLT) [21] is the optimal transform for the complete decorrelation of the data and in terms of energy compaction. The main drawback of the KLT is that its base functions are data-dependent, which means these functions need to be transmitted as overhead for the decoding of the image. The overhead can be so significant that it diminishes the advantages of using this optimum transform. Discrete Fourier Transform (DFT) has also been studied, however it generates large number of coefficients in the complex domain, and some of them are partly redundant owing to the symmetry property. The discrete cosine transform (DCT) [14] is the most commonly used transform for image and video coding. It is substantially better than the DFT on energy compaction for most correlated sources [15]. The popularity of DCT is based on its properties. First of all, the basis functions of the DCT are data independency, which means that none of them needs to be transmitted to the decoder. Second, for Markov sources with high correlation coefficient, the compaction ability of the DCT is very close to that of the KLT. Because many sources can be modeled as Markov sources with a high correlation coefficient, this superior compaction ability makes the DCT very attractive. Third, the availability of VLSI implementations of the DCT makes it very attractive for hardware-based real time implementation.

For lossy compression, Quantization is the most commonly used method for reducing the data rate, which represents a large set of values with a much smaller set. In many cases, scalar quantizers are used to quantize the transformed coefficients or DFD data in order to obtain an approximate representation of the image. When the number of reconstruction levels is given, the index of the reconstructed level is sent by using a fixed length code word. The Max-Lloyd quantizer [16, 17] is a well-known optimal quantizer, which results in the minimum mean squared quantization error. When the output of the quantization is entropy coded, a complicated general solution [18] is proposed. Fortunately, at high rates, the design of optimum quantization becomes simple because the optimum entropy-coded quantizer is a uniform quantizer [19]. In addition, it has been shown that the results also hold for low rates [18]. Instead of being quantized independently, pixels can be grouped into blocks or vectors for quantization, which is called vector quantization (VQ). The main advantage of VQ over scalar quantization stems from the fact that VQ can utilize the correlation between pixels.

After the data correlation has been reduced in both the spatial and temporal and the de-correlated data are quantized, the quantized samples are encoded differentially so as to further reduce the correlations as there may be some correlation from sample to sample. Thus we can predict each sample based on its past and encode and transmit only the differences between the prediction and the sample value. The basic differential encoding system is known as the differential pulse code modulation or DPCM system [20], which is an integration of quantizing and differential coding methods, and a variant of the PCM (Pulse code modulation) system.

It is very important to realize that the current typical image and video coding approaches are a hybrid coding that combines various coding approaches within the same framework.

For example, in JPEG, the block-based DCT transform, DPCM coding of the DC coefficient, quantization, zig-zag scan, run-length coding and Huffman coding are combined in the image compression procedure. In video coding, the hybrid motion-compensated DCT coding scheme is the most popular scheme adopted by most of the video coding standards. In this hybrid scheme, the video sequence is first motion compensated predicted, and the resulted residue is transformed by DCT. The resulted DCT coefficients are then quantized, and the quantized data are entropy coded. More information about the DCT-based video compression can be found in Chapter 3.

1.3.2 Video Coding Standards

The work to standardize video coding began in the 1980s and several standards have been set up by two organizations, ITU-T and ISO/IEC, including H.26x and the MPEG-x series. So far, MPEG-2/H.262 and MPEG-4 AVC/H.264 have been recognized as the most successful video coding standards. Currently, MPEG and VCEG are looking into the requirements and the feasibility of developing the next generation of video coding standards with significant improvement in coding efficiency over AVC/H.264. We will now review briefly the major standards, more detailed information can be found in Chapter 5.

H.261 [21] was designed in 1990 for low target bit rate applications that are suitable for the transmission of video over ISDN at a range from 64 kb/s to 1920 kb/s with low delay. H.261 adopted a hybrid DCT/DPCM coding scheme where motion compensation is performed on a macroblock basis. In H.261, a standard coded video syntax and decoding procedure is specified, but most choices in the encoding methods, such as allocation of bits to different parts of the picture are left open and can be changed by the encoder at will.

MPEG-1 is a multimedia standard with specifications for the coding, processing and transmission of audio, video and data streams in a series of synchronized multiplexed packets. It was targeted primarily at multimedia CD-ROM applications, and thus provided features including frame based random access of video, fast forward/fast reverse searches through compressed bit streams, reverse playback of video, and edit ability of the compressed bit stream.

Two years later, MPEG-2 was designed so as to provide the coding and transmission of high quality, multi-channel and multimedia signals over terrestrial broadcast, satellite distribution and broadband networks. The concept of 'profiles' and 'levels' was first introduced in MPEG-2 in order to stipulate conformance between equipment that does not support full implementation. As a general rule, each Profile defines a set of algorithms, and a Level specifies the range of the parameters supported by the implementation (i.e., image size, frame rate and bit rates). MPEG-2 supports the coding and processing of interlaced video sequences, as well as scalable coding. The intention of scalable coding is to provide interoperability between different services and to support receivers with different display capabilities flexibly. Three scalable coding schemes, SNR (quality) scalability, spatial scalability and temporal scalability, are defined in MPEG-2.

In 2000, the H.263 [22] video standard was designed so as to target low bit rate video coding applications, such as visual telephony. The target networks are GSTN, ISDN and wireless networks, whose maximum bit rate is below 64 kbit/s. H.263 considers network-related matters, such as error control and graceful degradation, and specific requirements for video telephony application such as visual quality, and low coding delay,

to be its main responsibility. In H.263, one or more macroblock rows are organized into a group of blocks (GOP) to enable quick resynchronization in the case of transmission error. In encoding, a 3D run-length VLC table with triplet (LAST, RUN, LEVEL) is used to code the AC coefficients, where LAST indicates if the current code corresponds to the last coefficient in the coded block, RUN represents the distance between two non zero coefficients, and LEVEL is the non zero value to be encoded. H.263 adopts half-pixel motion compensation, and provides advanced coding options including unrestricted motion vectors that are allowed to point outside the picture, overlapped block motion compensation, syntax-based arithmetic coding and a PB-frame mode that combines a bidirectionally predicted picture with a normal forward predicted picture.

After that, MPEG-4 Visual was designed for the coding and flexible representation of audio-visual data in order to meet the challenge of future multimedia applications. In particular, it addressed the need for universal accessibility and robustness in an error-prone environment, high interactive functionality, coding of nature and synthetic data, as well as improved compression efficiency. MPEG-4 was targeted at a bit rate between 5–64 kbits/s for mobile and PSTN video applications, and up to 4 Mbit/s for TV/film applications. MPEG-4 is the first standard that supports object-based video representation and thus provides content-based interactivity and scalability. MPEG-4 also supports Sprite coding technology, which allows for the efficient transmission of the background scene where the changes within the background are caused mainly by camera motion. More information about MPEG-4 and object-based video coding can be found in section 6.4.

H.264/MPEG-4 AVC is the latest video standard developed jointly by ITU and ISO. It is targeted at a very wide range of applications, including video telephony, storage, broadcast and streaming. Motion prediction ability is greatly improved in H.264/AVC by the introduction of directional spatial prediction for intra coding, various block-size motion compensation, quarter sample accurate motion compensation and weighted prediction, etc. A 4×4 integer transform was adopted in H.264/AVC so as to replace the popular 8×8 DCT transform, which will not cause inverse-mismatch; smaller size transform seldom causes ringing artifacts and requires less computation. The details of H.264/AVC are provided in Chapter 5.

1.3.3 Error Resilience

Although the video compression algorithms mentioned above can achieve a very high coding efficiency, the resulted compressed video streams are very vulnerable to errors in error-prone communications networks owing to the coding scheme that used. For example, the desynchronization errors caused by VLC coding and propagation errors caused by predictive coding make error handling very difficult. In lossy wireless networks, error resilient techniques [23] can significantly increase the system's robustness by using one of the following methods: encoder error resilience tools, decoder error concealment, as well as techniques that require cooperation bewteen encoder, decoder and the network. Figure 1.2 illustrates the simplified architecture of a video communication system, where the input video is compressed on the transmitter side and the generated bit stream is channel encoded so as to make it more robust against error-prone channel transmission. On the receiver side, the inverse operations are performed in order to obtain the reconstructed video for displaying.

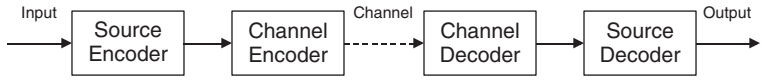

Figure 1.2 Simplified architecture of a video communication system

It is natural to increase the robustness of a compressed video bit stream by optimizing the source coding. The most popular approaches are increasing the amount of synchronization data and using special coding schemes such as RVLC (Reversible Variable Length Codes) [24]. Synchronization markers are bit sequences which have been designed especially so that they can be easily distinguished from other code words or variations of these code words with small perturbations. By inserting a synchronization marker periodically inside the compressed bit stream, any bit error will only affect those data between any two markers, which effectively prevents abusive error flooding. In this way, the decoder can resume proper decoding upon the detection of a resynchronization marker. Forward error correcting (FEC) is another approach, in which the FEC encoder adds redundant information to the bit stream which enables the receiver to detect or even correct transmission errors. This procedure is also known as channel coding. In a case where the data contains various portions of different importance, unequal error protection (UEP) becomes very useful, that is, the encoder uses stronger FEC codes on those important portions, while saving bits from having to protect unimportant portions. More detail on algorithms in this category can be found in sections 7.2 and 7.3.

Decoder error concealment refers to minimizing the negative impact of transmission error on the decoded image. The decoder recovers or estimates lost information by using the available decoded data or the existing knowledge of target applications. The current error concealment approaches can be divided into three types: spatial error concealment, temporal error concealment and adaptive error concealment. As the name indicates, spatial error concealment uses neighboring pixels to recover the corrupted area by using interpolations. It is very useful for high motion images in which frequency and temporal concealment do not yield good results, and it works particularly well for homogenous areas. Temporal error concealment makes use of the motion vectors and the data from previous time instants in order to recreate a missing block. As shown in Figure 1.3, the motion vectors are obtained by interpolation of the motion vectors of the macroblocks

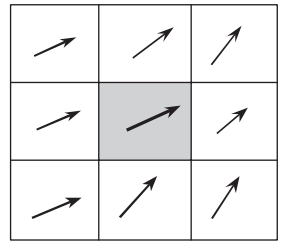

Figure 1.3 Motion vector interpolations in time domain

next to the one lost. This approach works well under the hypothesis that adjacent macroblocks move in the same direction to that of the lost macroblock. For those scenarios presenting high and irregular motion levels and scene changes, spatial concealment gives better results than temporal concealment. So as to take advantage of both the spatial and temporal approaches, adaptive error concealment selects a method according to the characteristics of the missing block, its neighbors and the overall frame, in order to perform the concealment. In this way, some parts of the image might be concealed using spatial concealment, while others might use temporal or other concealment methods, or even a combination of these approaches.

In the system, if there exist mechanisms to provide the encoder with some knowledge of the characteristics of the transmission channel – for example, if a feedback channel is set up from the decoder to the encoder so that the decoder can inform the encoder about which part of the transmitted information is corrupted by errors – then the encoder can adjust its operation correspondingly in order to suppress or eliminate the effect of such error. This type of approach is called network adaptive encoding. Of course, the encoder can also decide if retransmission of the whole video frame or some extra data would be helpful to the decoder so as to recover the packet from error. This is called ARQ (automatic repeat request). ARQ is typically not suitable in real-time communications owing to the intolerable round-trip delay. In summary, the network adaptive encoding approach optimizes the source (and channel) encoder by considering transmission factors, such as error control, packetization, packet scheduling and retransmission, routing and error concealment. It can significantly improve the system's performance (see section 9.4.1 for an example).

In H.264/AVC, many error resilience features have been adopted:

- Slice structured coding, in which slices provide resynchronization points (slice header) within a video frame and the encoder can determine the location of these points on any macroblock boundary.
- Arbitrary slice ordering, in which the decoding order of the slices may not follow the constraint that the address of the first macroblock within a slice is increasing monotonically within the NAL unit stream for a picture.
- Slice data partition, in which the slice data is partitioned into three parts: header, intra and inter-texture data.
- Redundant slices, which are provided as a redundant data in case the primary coded picture is corrupted.
- Flexible macroblock ordering, in which the macroblocks in a frame can be divided into a number of slices in a flexible way without considering the raster scan order limitation.
- Flexible reference frames, by which the reference frames can be changed on the macroblock basis;
- H.264/AVC defines IDR pictures and intra-pictures, so that the intra-picture does not have to provide a random access point function, while the IDR picture plays such a role.

1.3.4 Network Integration

RTP/UDP/IP is the typical protocol stack for video transmission. UDP and TCP are transport protocols supporting functions such as multiplexing, error control and congestion

control. Since TCP retransmission introduces delays that are not acceptable for many video applications, UDP (User Data Protocol) is usually employed although it does not guarantee packet delivery. Therefore, the receiver has to rely on the upper layers in order to detect packet loss. RTP (Real-time transport protocol) is the protocol designed to provide end-to-end transport functionality, however, it does not guarantee QoS or reliable delivery. The RTP header contains important information such as timestamp, sequence number, payload type identification and source identification. RTCP (Real Time Control Protocol) is a companion protocol to RTP, which provides QoS feedback through the use of a Sender Report and Receiver Report at the source and destination, respectively. Thus, a video transmission procedure can be described as the following. At the sender side, the video data are compressed and packed into RTP packets and the feedback control information is transferred to the RTCP generator. The resulting RTP and RTCP packets go down to the UDP/IP layer for transport over the Internet. On the receiver side, the received IP packet are first unpacked by the IP and then the UDP layer and are then dispatched by the filter and dispatcher to the RTP and RTCP packets. The RTP packet is unpacked by the RTP analyzer and tested for loss detection. When packet loss is detected, the message will be sent to the error concealment module for further processing. On the other hand, the RTCP packets are unpacked and the message containing feedback information will be processed. It is important to observe that the feedback information exchange mechanism and feedback control algorithms at the end system provide QoS guarantees. For example, when network congestion occurs, the receiver can catch it by detecting symptoms such as packet loss or packet delay. The receiver then sends feedback RTCP packets to the source in order to inform it about the congestion status. Thus, the sender will decrease its transmission rate once it receives the RTCP packet. This way, the source can always keep up with network bandwidth variation and the network is therefore utilized efficiently.

For H.264/AVC, the elementary data unit for encapsulation by transport protocols (for example RTP) is called the network abstraction layer (NAL) unit. The NAL formats the video content data and provides header information in a manner appropriate for conveyance by particular transport layers. Each NAL unit consists of a one-byte header and the payload byte string, and the header indicates the type of the NAL unit and the relative importance of the NAL unit for the decoding process. During packetization, the RTP packet can contain either one NAL unit or several NAL units in the same picture.

As introduced in [25], in a 3GPP multimedia service the compressed IP/UDP/RTP packet (with RoHC) is encapsulated into a single PDCP packet, and then segmented into smaller pieces of RLC-PDU units. The RLC layer can operate in both the unacknowledged mode and the acknowledged mode, in which the former does not guarantee the data delivery while the later uses ARQ for error correction. In the physical layer, the FEC is added to the RLC-PDU depending on the coding schemes in use.

On the other hand, there are many other protocol stacks for video delivery, for example, the PSS (packet-switching streaming service) supports both the IP/UDP/RTP and IP/TCP stack, and the MBMS (multimedia broadband/multicast service) supports both the IP/UDP/RTP and IP/UDP/LCT/ALC/FLUTE stacks. In PSS, an RTCP extension is standardized so as to support packet retransmissions for RTP applicable to unicast and multicast groups.

In [26], the protocol stacks and the end-to-end architecture of a video broadcast system over WIMAX are introduced. As shown in Figure 1.4, the MBS (multicast/broadcast service) server at the controller side handles coding/transcoding, RTP packetization, mapping video channel ID to multicast connection ID (CID), shaping and multiplexing, encryption, FEC coding, constructing MBS-MAC-PDU for transmission over WiMAX PHY/MAC, burst scheduling and allocating OFDMA data region for each MBS-MAC-PDU. The packets then go through UDP/IP/L2/L1 layers for header encapsulation and are transmitted in the wireless PHY medium. When a packet reaches the Base station, the header de-capsulation is conducted in layers L1/L2/IP/UDP, the obtained MBS-MAC-PDU is then encapsulated with headers by WiMAX MAC/PHY layers, the PHY channel coding is conducted to each MBS-MAC-PUD and they are then mapped to the corresponding OFDMA data region that is determined by the MBS server for transmission. At the MSS (Mobile Subscriber Station) side, the MBS client handles error correction, decryption and constructing RTP video packet and video decoding. It is important to emphasize that the MBS client determines multicast CID according to the selected video channel ID so that only those MBS-MAC-PDUS associated with the selected multicast CID will be decoded.

1.3.5 Cross-Layer Design for Wireless Video Delivery

It is natural to consider whether the interactions between the different network protocol layers can be optimized jointly in end-to-end system design in order to achieve better performance; this is called cross-layer design. Content-aware networking and network-adaptive media processing are two widely-used approaches in wireless video delivery, and they are considered to be two sub-sets of the cross-layer design. In the former approach,

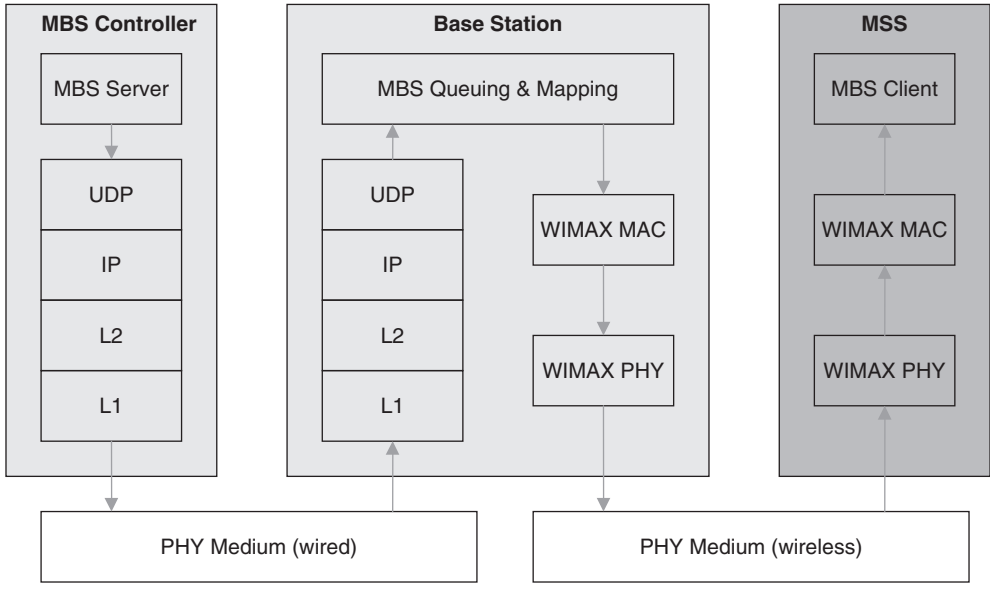

Figure 1.4 Video Broadcasting over MIMAX proposed in [25]

the resource management and protection strategies in the lower layers (i.e. PHY, MAC, network and transport layers) are optimized by considering the specific characteristics of the multimedia applications. The latter approach conducts the media compression and streaming algorithms after taking into account the mechanisms provided by the lower layers for error control and resource allocation. The principal difference between them occurs where the central controller locates: in the content-aware networking scenario, the controller locates at the lower layer, thus the application layer passes its control information, for example rate-distortion table, and requirements to the lower layer, so that after optimization the controller notifies the best strategy or parameters to the application layer for compression and transmission; and in the meantime, the controller also determines the parameters used in the PHY/MAC/IP and transport layers. On the other hand, in the network-adaptive media processing scenario, the controller locates at the application layer, which determines the optimal parameters for all layers given the information provided by the lower layers.

In Chapter 8, a quality-driven cross-layer optimization framework is introduced, which make the quality, in other words the user's experience of the system, the most important factor and the highest priority of design concern. Quality degradation is general is caused by factors such as limited bandwidth, excessive delay, power constraints and computational complexity limitation. Quality is therefore the backbone of the system and connects all other factors. In the optimization framework, the goal of design is to find an optimal balance within an N-dimensional space with given constraints, in which the dimensions include distortion, delay, power, complexity, etc. This chapter introduces an integrated methodology for solving this problem, which constructs a unified cost-to-go function in order to optimize the parameters in the system. It is an approximated dynamic programming (DP) approach, which handles global optimization over time with non-linearity and random distributions very well. The method constructs an optimal cost-to-go function based on the extracted features by using a significance measure model derived from the non-additive measure theory. The unique feature of the significance measure is that the non linear interactions among state variables in the cost-to-go function can be measured quantitatively by solving a generalized non linear Choquet integral.

1.4 Challenges and Opportunities for 4G Wireless Video

As shown in Figure 1.5, the world has become a content producer. People create and upload their own pieces of art onto the network while enjoying other people's masterpieces. It may be expected that in 4G the communication networks will continue to expand so as to include all kinds of channels with various throughputs, quality of services and protocols, and heterogeneous terminals with a wide range of capabilities, accessibilities and user preference. Thus the gap between the richness of multimedia content and the variation of techniques for content access and delivery will increase dramatically. Against such a background of expectations of universal multimedia access (UMA) will become a challenge for the 4G wireless network. The major concept of UMA is universal or seamless access to multimedia content by automatic selection or adaptation of content following user interaction. In Chapter 6, this topic is discussed in detail and the related content analysis techniques and standards are introduced.

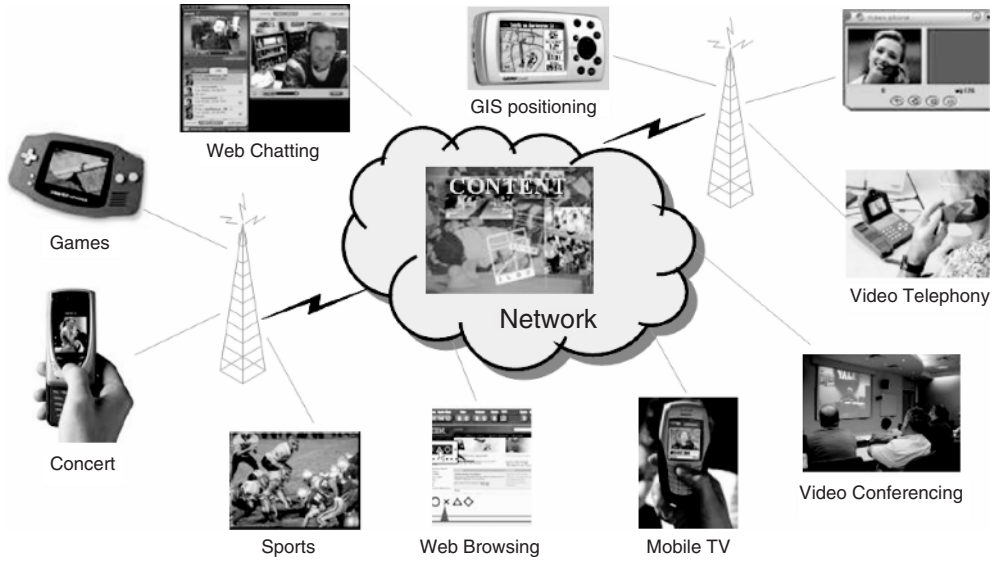

Figure 1.5 Content has become the center of the network communications

Content-based interactivity is highly connected to UMA but imposes higher expectations and requirements on content understanding supports. In the 4G high-speed communication systems, the mobile TV user may customize the content by manipulating the text, image, audio, video and graphics, and may even access information that is not available in the current system (for example, the player's information in a motorcycling racing game, or even the dynamic speed of the motorcycle that he/she is currently riding). Consequently, the 4G wireless system may face the challenge that watching TV or movie has been changed from a passive activity to a new user experience with much higher interactivity. In Chapters 6 and 11, this topic is discussed in detail and the related content-based video representation and communications techniques are demonstrated.

Mobile TV has been a network killer application for a long time, the supported resolution of TV clips by mobile devices has increased dramatically from QCIF (176×144) to VGA (640×480) in the past four or five years; nVidia showed off their APX2500 solution that can support even 720 p (1280×720 resolution) video in April 2008. It is expected that high-definition (HD) TV programmes will soon be delivered to and played in mobile devices in 4G networks, although there are still many challenges to be faced. In Chapter 10, the topics of digital entertainment techniques and mobile TV are discussed in detail.

On the other hand, video on TV will not be flat for much longer; that is, the next step forward is destined to be 3D video or 3D TV services over 4G wireless networks with various representation formats. As shown in Figure 1.6, the expected road map for reality video over wireless was predicted by the Japanese wireless industry in 2005, and it is interesting that the expected deployment of stereo/multi-view/hologram is around the same time period as that of 4G. Currently, stereoscopic and multi-view 3D videos are more developed than other 3D video representation formats, as their coding approaches

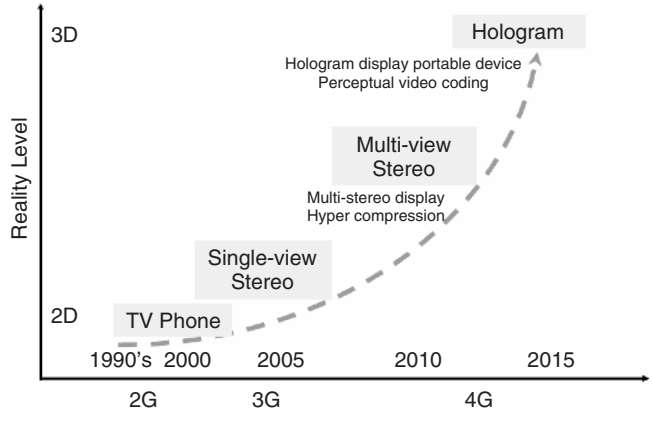

Figure 1.6 Estimated reality video over wireless development roadmap

are standardized in MPEG 'video-plus-depth' and JVT MVC standards, respectively. It is also claimed that coded 3D video only takes a 1.2 bit rate as compared to monoscopic video (i.e., the traditional 2D video). Clearly, higher reality requirements will bring in a larger volume of data to be delivered over the network, and more service and usage scenarios to challenge the 4G wireless infrastructures and protocols. In Chapter 5, some related techniques on the topic of multi-view video are discussed.

Recently, P2P (peer-to-peer) live video streaming has become very popular as it needs much less deployment cost and is able to take advantage of the distributed storage and computing resources in the network. The basic idea of P2P is to treat every node in the network as a peer and the peer can help to pass the video packets to others. However, in the 4G system, the large scale P2P wireless video streaming capability of supporting many viewers would still be very challenging, considering the difficulties of effective incentive mechanisms for peers willing to collaborate, peers' media relay capability, peers' system topology discovery capability, QoS control, etc. In Chapter 12, this topic is discussed in detail and the related video streaming techniques are introduced.

Finally, as so many fancy applications and services are expected to be deployed in 4G networks, the cross-layer design mechanism for content delivery with satisfactory quality of user experience becomes critical. In Chapters 8 and 9, quality-driven cross-layer design and optimization methodology and its applications for various content-based video communication scenarios are addressed in order to resolve this issue.

References

1. P. J. L. van Beek, and A. M. Tekalp, Object-based video coding using forward tracking 2-D mesh layers, in *Proc. SPIE Visual Comm. Image Proc*., Vol. 3024, pp. 699–710, San Jose, California, Feb. 1997.
2. J. Chalidabhongse and C. C. Kuo, Fast motion vector estimation using multiresolution-spatio-temporal correlations, *IEEE Trans. Circuits Syst. Technol.*, Vol. 7, pp. 477–88, June 1997.
3. A. Chimienti, C. Ferraris, and D. Pau, A complexity-bounded motion estimation algorithm, *IEEE Trans. Image Processing*, Vol. 11, No. 4, pp. 387–92, April 2002.
4. C. C. Cutler, "*Differential quantization for television signals*", U.S. Patent, 2,605,361. July 29, 1952.

5. M. Dudon, O. Avaro, and C. Roux, Triangular active mesh for motion estimation, *Signal Processing: Image Communication*, Vol. 10, pp. 21–41, 1997.
6. N. Farvardin and J. W. Modestino, Optimum quantizer performance for a class of non-Gaussian memoryless sources, *IEEE Trans. Information Theory*, IT-30, pp. 485–97, May 1984.
7. B. Furht, J. Greenberg, R. Westwater, *Motion Estimation Algorithms for Video Compression*, Kluwer Academic Publishers, 1997.
8. H. Gish and J. N. Pierce, Asymptotically efficient quantization, *IEEE Trans. Information Theory*, Vol. 14, pp. 676–83, Sept. 1968.
9. ITU-T Recommendation H.261, *Video codec for audiovisual services at px64kbps*, 1993.
10. ITU-T Recommendation H.263, *Video coding for low bitrate communication*, 1998.
11. D. A. Huffman, A method for the construction of minimum redundancy codes, *Proceedings of the IRE*, Vol. 40, pp. 1098–1101, 1951.
12. A. K. Jain, Image data compression: a review, *Proceedings of the IEEE*, Vol. 69, pp. 349–89, March 1981.
13. N. S. Jayant and P. Noll, *Digital Coding of Waveforms*, Englewood Cliffs, NJ: Prentice Hall, 1984.
14. K. Lengwehasatit and A. Ortega, A novel computationally scalable algorithm for motion estimation, in *Proc. VCIP'98*, 1998.
15. S. P. Lloyd, Least squares quantization in PCM, *IEEE Trans. Information Theory*, Vol. 28, pp. 129–37, March 1982.
16. J. Max, Quantizing for minimum distortion, *IRE Trans. Information Theory*, Vol. 6, pp. 7–12, March 1960.
17. Y. Nakaya and H. Harashima, Motion compensation based on spatial transformations, *IEEE Transactions on Circuits and Systems*, Vol. 4, No. 3, pp. 339–56, June 1994.
18. B. Natarajan, V. Bhaskaran, and K. Konstantinides, Low-complexity block-based motion estimation via one-bit transforms, *IEEE Trans. Circuits Syst. Video Technol.*, Vol. 7, pp. 702–6, Aug. 1997.
19. K. R. Rao and P. Yip, *Discrete Cosine Transform: algorithms, advantages, applications*, Academic Press, Boston, 1990.
20. J. Rissanen, G. G. Langdon, Arithmetic coding, *IBM Journal on Research and Development*, Vol. 23, No. 2, pp. 149–62, March 1979.
21. K. Sayood, *Introduction to Data Compression*, Morgan Kaufmann Publishers, Inc. 1996.
22. C. E. Shannon, A mathematical theory of communication, *Bell System Technical Journal*, Vol. 27, pp. 379–423, 1948.
23. T. Stockhammer, M. M. Hannuksela, H.264/AVC Video for Wireless Transmission, *IEEE Wireless Communications*, Vol. 12, Issue 4, pp. 6–13, August 2005.
24. Y. Takishima, M. Wada, H. Murakami, Reversible variable length codes, *IEEE Trans. Communications*, Vol. 43, Nos. 2/3/4, pp. 158–62, Feb./March/April 1995.
25. J. Wang, M. Venkatachalam, and Y. Fang, System architecture and cross-layer optimization of video broadcast over WiMAX, *IEEE Journal of Selected Areas In Communications*, Vol. 25, No. 4, pp. 712–21, May 2007.
26. Y. Wang, S. Wenger, J. Wen, and A. K. Katsaggelos, Review of error resilient techniques for video communications, *IEEE Signal Processing Magazine*, Vol. 17, No. 4, pp. 61–82, July 2000.