3 Internet Video

3.1 Introduction

Today's digital video systems can produce excellent quality visual and auditory experiences at relatively low cost. However, Internet users still encounter many problems that result in an unsatisfactory experience. Although the situation has been steadily improving, buffering delays, incompatible formats, blocky, blurry images, jerky motion, poor synchronization between audio and video are not uncommon and lead to frustration to the point that the user experience of video services involving search is greatly impacted. User's expectations are raised by their familiarity with broadcast television systems, where well defended standards, mature technologies, and abundant bandwidth prevail. In this chapter, we provide background information to shed light on the complexities involved in delivering IP video. We address the practical issues that video search engine systems must resolve in order to deliver their "product" – relevant video information – to users.

3.2 Digital Video

3.2.1 Aspect Ratio

When designing user interfaces for visualizing video search results, the frame aspect ratio (FAR) of the source video and resulting thumbnails must be taken into account. For many years the ratio of width to height for the bulk of video on the Web was 4:3, but with HD cameras dropping in

price, more and more 16:9 format video is appearing. Content sourced from motion picture film may have one of several aspect ratios, but has always had a wider aspect ratio than standard definition television. It is also common to find wide aspect ratio source material digitized within a 4:3 frame in letterbox format with black bars at the top and bottom. When presenting grids of thumbnails for visual browsing, these circumstances present basic layout issues, and make the thumbnails for some content appear smaller than for others, impeding browsing.

Metadata extraction systems must accommodate video with disparate spatial resolutions. For example, a system may detect faces and represent the bounding box results in XML format for content that is 640 x 480 or 320 x 240 but render a user interface with 160 x 120 thumbnails. We can scale the thumbnails or rely on the browser to do so, but we must also scale the bounding box coordinates if we are to plot the detection results overlaid on the thumbnails using Scaleable Vector Graphics (SVG) or Vector Markup Language (VML). So any image region-based metadata must be effectively normalized for query and display to handle source images of various scales and must support different vertical and horizontal scale factors to normalize different frame aspect ratios.

Pixel aspect ratio (PAR) further complicates the matter. Early analog cameras and analog TV systems did indeed have continuous signals along the scan lines that varied in relation to the illumination – similar to the situation with audio microphones. However, in the vertical direction, the picture was sampled as is done in digital systems. There is a discrete fixed number of "lines" per frame – for NTSC we can count on 480 valid lines of picture information. Of course for digital television, we must sample in the other dimension as well, and then quantize the samples. Since the FAR for NTSC is 4:3, we should divide each line into 640 pixels so that each sample covers the same small extent of the picture in the vertical and horizontal directions – a square pixel. So why should we introduce a "rectangular pixel?" It turns out that the channel bandwidth of NTSC specification justifies sampling the signal at a higher rate to preserve image detail. 720 is commonly used and ATSC DTV also specifies a sampling resolution for standard definition video of 704 x 480. So some content may be sampled with square pixels while other content may have pixels that look like shoe boxes standing on end. A feature detector based on spatial relations (e.g. Viola / Jones) trained on square pixel data will perform poorly on rectangular pixel data, so a preprocessing image conversion step is required. Of course it is possible to scale the detector or make it invariant to scale, but this is more complex. Failure to manage the complexity of FAR and PAR correctly not only degrades metadata extraction algorithm performance, it

results in objectionable geometric distortion: circles looking like ovals, and actors looking like they have put on weight.

A similar issue can arise in the temporal dimension. We may encounter video with a wide range of frame rates. Rates of 30, 29.97, 25 and 24 frames per second are common and lower bit-rate applications may use 15 f/s. Security or Webcam video may forsake smooth motion altogether and use 1 f/s to save storage. Media players can render the video at the proper rate, but motion analysis algorithms that assume a given frame rate may not perform well for all content. This effect is not usually much of a problem since the design of these algorithms intrinsically accommodates a wide range of object velocities. Think here of gait detection or vehicle counters – the absolute estimate of object velocity may be affected but the detection rate may not be.

Interlacing is another source of problems for video systems. Interlacing was introduced years ago with the first television broadcast standards to effectively double the spatial resolution given a limited bandwidth channel. The cost, however, is lower temporal resolution (and increased complexity for video processing engineers.) The frame is divided into two fields, one with the odd numbered lines and one with the even. The fields are sent sequentially transmitted. The result is fine for static pictures, but any objects that are in motion result in saw-tooth edges if the video is paused or sampled at the frame resolution. If we are subsampling to create thumbnails, this may not be a problem. The new HDTV standards perpetuate interlacing (1080i vs. 720p). The term "progressive" is used to refer to non-interlaced video, but amusingly the term "progressive JPEG" refers to something similar to interlacing. Video processing algorithms must handle interlaced sources gracefully, by de-interlacing, dropping fields, or by taking into account the slight vertical sampling offset between consecutive fields.

The relation of illumination or intensity to signal amplitude mentioned above is nonlinear and is represented as an exponential referred to as 'gamma'. Analog television systems were designed for CRTs with a nonlinear response and so precompensated the signal. Computer graphics applications and many image processing algorithms assume a linear relation.

3.2.2 Luminance and Chrominance Resolution

The human visual system cannot resolve image features that have differing hue but similar brightness as well as it can resolve features that vary in lu-

minance. Therefore, compression and transmission systems encode chrominance information at lower spatial resolution than luminance with little apparent loss of image quality. The terms 4:2:2, 4:2:0, 4:1:1, etc. refer to the amount of subsampling of the chrominance relative to the luminance for different applications. When the image is rendered for display, it is converted from a luminance–chrominance color space such as Yuv or Y, Cr, Cb to R,G,B using a linear transform. Nonlinear transformations to spaces such as H,S,V yield a better match to the perceived visual qualities of color, but the simpler linear transformation is sufficient for coding gain. Single chip CCD or CMOS sensors designed for low cost consumer applications such as mobile phones or cameras also take these effects into account. Rather than having an equal number of R,G,B sub-pixels, a color filter array such as the Bayer checkerboard [Bayer76] is used to produce an image with relatively higher luminance resolution. This scheme has twice as many green pixels as red or blue. Another point to consider is that the spectral sensitivity of the human eye peaks in the green region of the spectrum, while silicon's sensitivity is highest in the infrared (IR). IR blocking filters are used to select the visible portion, but the sensitivity of the blue is much lower than the red. The resulting signal to noise ratio for the blue component is always lower than the green or red. Color correction processing as well as gamma correction tends to emphasize this noise. Also, color correction parameters are determined for given illumination conditions and, particularly in consumer applications, poor end-to-end color reproduction is common. Noise in the blue component, subsampled chrominance, and poor color reproduction not only degrade image quality, but also degrade performance of video processing algorithms that attempt to take advantage of color information.

3.2.3 Video Compression

Web media is compressed; users almost never encounter original, uncompressed video or audio – the sheer scale of storage and bandwidth required makes this impractical. Even QVGA resolution requires over 55 megabits per second to render in 24 bit RGB at 30 frames per second, while higher resolutions require even more bandwidth. The requirement that video be compressed has several implications for video search engine systems as we shall see.

Lossless video compression is rarely used since the bitrate reduction attainable is quite limited. Lossy compression offers impressive performance, but comes at the price of information loss – the original image or

video sequence cannot be fully recovered from the compressed version. The distortion between the original the reconstructed image is often measured using the peak signal to noise ratio PSNR although this is well known to be a poor match to perceived image quality. It is extremely difficult to quantify image quality; it is highly subjective and content dependent. PSNR is an example of a "full reference" quality metric as defined by ITU-T Recommendation J.144 – "partial reference" and "no reference" techniques are used for applications where full reference data is not available, for example measuring quality at the set–top box at the end of a video delivery service [J.144]. Compression algorithms are evaluated using rate-distortion plots which reflect attempts to approach the information theoretic limits outlined in Shannon's rate distortion theory. Algorithmic improvements have made great strides in pushing the theoretic limits, while Moore's law has allowed for increasingly complex implementations to be standardized and used in practical systems.

Since video is a series of still frames, one would expect that video compression is related to the JPEG image compression used in digital cameras, and, in fact, this is indeed the case. Many consumer cameras capture video as a sequence of JPEG frames to create "Motion JPEG" (M-JPEG) format since the computational complexity of this approach is minimal. At the high end, professional editing systems use M-JPEG or "MPEG-2 I frame-only" as well. Here the systems are designed for high-quality and ease of cutting and splicing sequences together, rather than on high compression ratios.

JPEG works by dividing an image into small blocks and transforming (using the Discrete Cosine Transform) from the pixel domain to the spatial frequency domain. In this domain, pixels whose intensity values are similar to their neighbors can be efficiently represented – in smooth areas of an image, an entire block can be approximated by just its average (or DC) value or just a few DCT coefficients. To get an intuition for the concept of spatial frequency, take a look at a folder of digital photo files and sort them by the file size. The larger files will have a large proportion of the image in sharp focus with a lot of edge information, say from a brick wall or a tree with leaves. The smaller, more compressed, files will be the out of focus shots or contain a small object on a large homogenous background. Now suppose that we point a camera at a brick building and capture a video sequence in vivid detail. The frames are nearly identical – they have a high degree of temporal redundancy. By subtracting the second frame from the first, we end up with a frame that is mostly uniform, perhaps with a small region where someone sitting by a window in the building moved slightly. As we have found, this is the type of image that compresses well, so that our entire sequence can be efficiently represented by encoding the first

frame (intra-frame coding) followed by encoding the difference between this frame and subsequent frames (inter-frame coding). Now of course there are some complications that arise due to temporal noise in the signal, and illumination changes due to passing clouds, etc. But the main problem in this scenario is that slight camera motion will result in a large difference image in any region where the image is not uniform (e.g. the sky will not cause much of a problem.) Video coders compensate for this using block-matching where a block of one frame is compared to several neighboring blocks in another subsequent frame to find a good match. In the case of a shift in the camera, most blocks will have the same shift (or motion vector). So, video compression from MPEG-1 up through MPEG-4 is based on DCT of motion compensated frame difference images.

Video compression standards are designed and optimized for particular applications; there is no one-size-fits-all codec. The ITU developed the H.261 and H.263 for low bitrate, low latency teleconferencing applications. For these applications, the facts that the camera is usually stationary (perhaps mounted on pan-tilt stage next to a monitor) and that conferencing applications typically involve static backgrounds with little motion greatly help improve the quality at low bitrates. It is reasonable here for coders to transmit intra-coded blocks rather than entire frames. MPEG-1 was developed for CD-ROM applications with bitrates in the 1 Mb/s range. MPEG-2 is used in broadcast distribution and in DVDs where higher quality and interlaced video support are requirements. MPEG-4 brings increased flexibility and efficiency, of course with increased complexity, and finally the ITU and MPEG bodies have achieved interoperability with MPEG-4 part 10, ITU H.264/AVC. For contribution feeds or editing applications M-JPEG or similar intra-coded video at very high bitrates is appropriate to ensure quality downstream.

MPEG-2 Systems [Info00] added a wide range of capabilities that were not available with MPEG-1. While “program streams” are used for file based applications (MPEG uses the term DSM – Digital Storage Media) which have negligible error, the notion of a transport stream was introduced to allow for efficient delivery over noisy channels such as may be found in typical broadcast systems such as cable or today’s IPTV over DSL. The transport stream specification also supports multiplexing several (even independent) media streams which enables secondary audio programming or alternative representations of the video at different resolutions and bitrates [Haskell97]. Table 3.1 lists a few common video compression standards and bitrates typically encountered. For actual maximum and minimum bit rates supported, readers should consult the standard documents.

Table 3.1. Applications of video compression systems (bit rates are approximate, and assume standard definition).

Standard	Typical bitrates	Common applications
M-JPEG, JPEG2000	Wide range, up to 60M	Low cost consumer electronics, High end video editing systems
DVCAM	25M	Consumer, semi-pro, news gathering
MPEG-1	1.5M	CD-ROM multimedia
MPEG-2	4–20M	Broadcast TV, DVD
MPEG-4 / H.264	300K–12M	Mobile video, Podcasts, IPTV
H.261, H.263	64K–1M	Video Teleconferencing, Telephony

Within all of these standards, there are "profiles" which are particular parameter settings for various applications. The latter standards have a wide range of flexibility here which allows them to span a wide range of applications while the earlier standards are more constrained. So it is possible for an MPEG-4 decoder not to be able to decode an MPEG-4 bit stream (e.g. if the decoder only supports a baseline profile). Profiles are intended for varying degrees of complexity (i.e. required computational power of encoders / decoders) as well as latency or error resilience. For example, for DVD applications, variable bit rate (VBR) encoding allows bits required to represent high action scenes to be effectively borrowed from more sedate shots. Of course, the player has to read large chunks of data from the disk and store them in a local buffer in order to decode the video. On the other hand, for digital broadcast TV, rapid channel change is desirable so the buffering requirements are kept to a minimum. The quality difference between DTV and DVD leads many viewers to think that DVDs are HD while in fact only Blue Ray and HD-DVDs support higher resolution than standard definition. Some of this confusion arises because DVDs are often letterbox, but primarily it is due to the lack of obvious coding artifacts such as blocking or contouring. Higher bitrates play a role, but even at the same bitrate, real-time encoding for low latency applications results in lower quality. Additionally, the quality of the source is key – some digital television sources are of dubious quality, perhaps with multiple generations of encoding – as well as the fact that mastering DVDs is done offline, allowing for two-pass encoding. DVD mastering is really an art; a bit like making a fine wine as opposed to producing grape juice. So, encoding systems designers have a challenging job to balance latency, complexity, error resilience, and bandwidth to achieve the quality of experience that the viewer ultimately enjoys.

What implications do these video compression systems have for video search engines?

- Video content analysis / indexing algorithms must either support the formats natively, or transcode to a format that is supported. Since many algorithms operate in the pixel domain as opposed to the compressed domain, this "support" may simply imply that the system can decode the video. However, the video quality does have an effect on indexing accuracy – noise or image coding artifacts such as blocks can be significant problems. Also, in some cases, periodic quality fluctuations due to poor bit allocation between intra- and inter-coded frames can produce more subtle artifacts.
- Of course from a systems perspective, high bitrate video may not be practical to archive on-line at scale. Further, each format must be supported by the client media player, and by media servers as well. This problem of incompatible media players and formats is driving a move to Flash formats, which at least offers a degree of independence from the client operating system.
- Finally, as we have seen, these codecs are highly optimized for particular applications, and this typically does not include streaming or fine grained random access.

MPEG frames are organized as "groups of pictures" or GoP which consists of an intra-coded frame (I frame) and several predicted frames (P and B frames). Applications such as media players can't jump into a video stream in the middle of a GoP and start playing – they must refer back to the I frame. So in effect the GoP length determines the precision for media replay requests. For many applications the GoP length is less than a second (15 frames is common) so this has only minor effects on the user experience, but for high coding efficiency applications, "Long GoP" coding is used where there may be several seconds between I frames. H.264/AVC introduces many more complex options in this area such as multiple reference frames for different macroblocks which further exacerbate random access [Rich03].

3.3 Internet Protocol Media Systems

3.3.1 Transport

Video search engines deliver their product to clients over IP connections in several ways:

- Download – This simple delivery system has been available since the beginning of HTTP where MIME types are used by browsers to launch the appropriate media player after the media has been downloaded to a local file.
- Progressive Download – Again, a basic HTTP server delivers the media file, but in this case its play-out is initiated via a media player before the entire file is downloaded.
- HTTP with byte offsets – The byte range feature of HTTP/1.1 is used to support random access to media files. Clients map user play position (time seek) requests to media stream byte offsets and issue requests to the server to fetch required segments of the media file.
- Managed Download – A specially designed client application provides additional features such as DRM management, expiration, reliable download and HTTP or P2P is typically used for transport. There are many types of these applications, from applications that operated in the background without much of a UI, to iTunes which include download management capabilities for Podcasts and purchased media.
- HTTP Streaming – These systems require a dedicated media server that parses the media file to determine the bit rate and delivers the content accordingly. Random access and other features such as fast start, fast forward, etc. may also be supported.
- RTSP / RTP – A media streaming server delivers the content via UDP to avoid the overhead of TCP retransmissions. Some form of error concealment or forward error correction can be used. Some IPTV systems use a "reliable UDP" scheme where selective retransmission based on certain conditions is employed.

3.3.2 Searching VoD vs. Live

Most video search applications inherently provide personalized access to stored media – essentially this is a "video on demand" (VoD) scenario, although the term VoD is commonly used to refer to movie rental on a set-top box delivered via cable TV or IPTV. For VoD, the connection is point to point and unicast IP transmission is appropriate. However, IPTV and Internet TV are channel based where many users are viewing the same content at the same time so multicast IP is employed. As the number of these feeds grows, users will need searching systems to locate channels of interest. In this scenario, EPG/ESG data including descriptions will provide the most readily accessible metadata for search. Live streams can be processed in real time to extract up to the minute metadata for more de-

tailed content-based retrieval. Of course, prepared programming and rebroadcasts of live events can be indexed a priori and used to provide users with more accurate content selection capabilities.

3.3.3 IPTV

IPTV is often heralded as the future of television, promising a revolution on the same scale as the Web. With all this potential, there are many groups co-opting the term IPTV to their own advantage. Does IPTV imply any television content delivered over an IP network? Well, we have been able to see video content streamed over the Internet for years so it makes sense to restrict the term IPTV to a narrower connotation. Of course, as more bandwidth has become available and desktop computers more powerful, we can experience full-screen video delivery and begin to approach broadcast TV quality (although HD delivery to large audiences over unmanaged networks is much more demanding and may be slow to evolve). The term "Internet TV" has been used to describe this type of system, and the term IPTV is generally accepted to mean delivery of a television-like experience over a managed IP network. To avoid confusion for the purposes of standardization, the IPTV Interoperability Forum (IIF) group formed by the Alliance for Telecommunications Industry Solutions (ATIS) [ATIS06] has defined IPTV as:

> the secure and reliable delivery to subscribers of entertainment video and related services. These services may include, for example, Live TV, Video On Demand (VOD) and Interactive TV (iTV). These services are delivered across an access agnostic, packet switched network that employs the IP protocol to transport the audio, video and control signals. In contrast to video over the public Internet, with IPTV deployments, network security and performance are tightly managed to ensure a superior entertainment experience, resulting in a compelling business environment for content providers, advertisers and customers alike.

In the context of video search, IPTV is a significant step towards an evolved state of video programming where the entire end-to-end process is manageable using generic IT methods. While there is clearly a long way to go in terms of interoperability and standardization for exchange of media and metadata, the IP and accessible nature of the new delivery paradigm paves the way toward making this a reality. This offers the potential for engineers competent in networking and data management technologies to bring their experience to bear on the problem of managing video distribution. The potential for metadata loss through conversions through the delivery chain is greatly reduced. Of course, today's IPTV systems use IP for

distribution to consumers, but IP is not necessarily used for contribution of broadcast content. Traditional and reliable methods used for cable delivery such as satellite, pitcher / catcher VoD systems, etc. will persist for the foreseeable future. In addition to ATIS, several other bodies including ETSI (DVB-IPTV), OMA (BCAST) and OpenIPTV are participating in drafting IPTV recommendations for a range of applications.

Although not specified in the ATIS/IIF definition, IPTV deployments are usually delivered via DSL links that do not have enough bandwidth to support the cable model of bringing all channels to the customer premises and tuning at the set-top. With VDSL2, downstream bandwidth is typically 25Mb/s which can accommodate two HD and two SD channels simultaneously. With IPTV over DSL, only a single channel for each receiver is delivered to the customer – effectively the "tuning" takes place at the central office. This is sometimes referred to as a "switched video" service (although the term is used is used in cable TV delivery as well). To support rapid channel changing, IPTV systems keep the GoP short and employ various techniques to speed up channel change. Of course short GoP and channel change bursts consume bandwidth and systems must balance these factors. Given this optimization, and the necessary FEC for DSL, IPTV streams must be transcoded for efficient archival applications where there is less need for error correction.

As we have seen, there are a wide range of video coding systems in use and each is optimized for its intended set of applications. As video content is acquired and ingested into a video search engine, it is very likely that the encoding of the source video is not appropriate for delivery from the search engine. In some cases the bit rate is simply too high to scale well given the number of concurrent users, or the format may be unsuitable for the intended delivery mechanism. Although some services attempt to redirect users to origin servers, the user experience of switching among multiple players (some of which may not be installed) to view the search results is less than seamless. Therefore many systems have opted to transcode video to a common format and host it. Flash Video is often the format of choice here due to its platform independence and wide installed base of players. The term transcoding is loosely used to refer to changing container formats, encoding systems, or bitrates. Transrating refers to changing only the bitrate (typically via re-encoding, not using scalable coding or multirate streaming). In some cases it is not necessary to fully decode the media streams and re-encode them, such as when changing only the container format. Also, the re-encoding process can be made more efficient by only partially decoding the source (perhaps re-using motion estimation results), but in many general purpose transcoding systems, the source is fully decoded and the results fed to a standard encoder. This approach is taken

because the required decoders and encoders are readily available and have been highly optimized to perform efficiently. Also, search engines may transcode to a small set of formats in order to target different markets such as mobile devices (e.g. YouTube's use of Flash Video required large scale transcoding in order to support AppleTV® and iPod Touch® which did not include support for Flash Video).

3.3.4 Rights Management

In addition to incompatible media formats, digital rights management (DRM) systems are not interchangeable, and systems that hope to process a cornucopia of content must navigate these systems as well. Various DRM systems such as Apple's FairPlay and Real's Helix can be applied to MPEG-4 AAC media, but this does not imply interoperability. While it would be in keeping with the sprit of DRM to allow the purchaser of a song (or a license to a song) to enjoy the media and justly compensate the provider, in practice this notion has been restricted so that the user must enjoy the song on a single vendor's device or player. MPEG-21 attempts to standardize the intent, if not the particular implementation, of rights through the definition of a rights expression language (REL). Examples of limited rights to use content include play once, play for a limited time, hold for up to 30 days and then play many times for up to 24 hours after the first play. The hope is that at least these desired use cases can be codified even though a particular media player device may only support a limited number of DRM systems or only a single system. In reality, choosing a DRM system is tantamount to choosing a media player. Purchased music and media (iTunes, Windows Media), video download services, DVDs and broadcast television all have forms of encryption for prevention of unauthorized copy of content (CCS, AACS for DVDs, conditional access for DVB and Cable, broadcast flag for ATSC). Finally, media watermarking and embedding user information in metadata to enable forensic traceability of a copied asset to its source are additional techniques used to preserve the copyright owner's rights.

3.3.5 Redirector Files

Video search engine systems can make use of redirector (or "metafiles") to provide increased functionality when initiating video playback. Instead of the user interface containing links directly to the media files, the links

point to media metafiles which are small text markup files issued by the HTTP server with a particular MIME type that is mapped to the client media player. At this point the browser has done its job and control of the streaming session is passed to the media player which connects to a media server. This arrangement provides several advantages:

- Response time: the small files download instantly and the media player application can launch quickly and begin video playback using progressive download or streaming.
- Failover / Loadbalancing: The redirector files can include alternative URLs for retrieving the media and media players support a failover mechanism where connection to servers indicated by a list of URLs is attempted in sequence. Applications can also generate metafiles dynamically with URLs pointing to lightly loaded streaming servers if the desired media is avalible on multiple media servers.
- Playtime offsets / clipping: the media play time start and duration can be encoded in the metafile. The ability to seek into the media is critical for directing users to relevant segments in long-form content.
- Playlists / Ad insertion: sets of media files matching user quieres can be represented as a play list and interfaces supported by the media player can be used to navigate amoung them. Preroll or interstitial advertizing can be supported using this mechanism – where essentially one or more clips in the playlist are ads. Much to users chagrin, these clips can be marked so that the ability to skip or fastforward are disabled during playback of ads.
- Additional features: Optionally, directives for including media captions (similar to closed captions) are supported. Also, metadata specific to the session can be included, e.g. the title can be set to "Results for your query for the term: NASA." This mechanism can be used to effectivly override any metadata embedded in the media itself.

Table 3.2. Media metafile systems.

Format	Extension	Comments
Real Audio Metafile	.ram	One of the early streaming Web media formats
Windows Media Metafile	.asx, .wmx, .wvx, .wax	Extensions connote video (v), and audio (a) but the format is the same; 'asx' is deprecated
Synchronized Media Information Language	.smil, .smi	Supports many additional features such as layout.

Some common file formats or protocols for achieving this effect are shown in

Table 3.2; also the playlist formats such as M3U and PLS provide a somewhat similar function, but with a limited subset of the capabilities.

Fig. 3.1 shows a Windows Media Format metafile that includes failover (if the media is not available from mserver1, then mserver2 will be contacted). Also the media play position is set to 120 seconds. For Quicktime, a "reference movie" can be created to point to different bitrate versions of the content. A Reference Movie Atom (rmra) can contain multiple Reference movie descriptor atoms (rmda).

```
<ASX version = "3.0">
 <Entry>
  <Ref href="http://mserver1.company.com/media/video1.wmv"/>
  <Ref href="http://mserver2.company.com/media/video1.wmv"/>
  <StartTime Value="120"/>
 </Entry>
</ASX>
```

Fig. 3.1. ASX Metafile with failover and start offset.

Embedded players: While UIs that launch the media player using metafiles can be extremely lightweight (no client side JavaScript is required) and therefore easily supported by a wide range of browser clients, a more integrated user experience is achieved by embedding the media player in the browser. With this approach, the player plug-in loads once and user navigation of search results can change the media and change the play position. For example Fig. 3.2 shows a client side script fragment for loading a media stream and seeking to a given point using the Windows Media Player object model, assuming that the player has been embedded and named "Player". More recently, immersive interfaces that provide a user experience more similar to TV have been created leveraging emerging technologies including AJAX, XAML, and using the graphics capabilities of clients to their full potential to provide full screen interfaces with overlaid navigational elements.

```
Player.URL = "http://server1.company.com/media/video1.wmv";
Player.controls.currentPosition = 120;
Player.controls.Play();
```

Fig. 3.2. Controlling media playback using client side scripting.

3.3.6 Layered Encoding

Some encoding systems include features to efficiently support scalability. Scalability encompasses several varieties including spatial, temporal, and even object scalability. The idea is to encode media once and enable multiple applications where views may be alternatively rendered for services with bandwidths less than the media encoded bitrate. The concept also supports the notion of a base layer and an enhancement layer where the base layer may represent a lower resolution or lower frame rate version of the media and the enhancement later can include more detail. In a best effort network delivery scenario with variable congestion, the base layer can be delivered with a guaranteed quality of service (QoS), while the enhancement layer can use a lower priority so that the overall system user experience will be improved. (Rather than one user – or worse all users – experiencing video dropouts, all users may see a slight degradation in quality).

Some media streaming systems use a less efficient scheme to provide a similar effect. Using what is called "multirate encoding," multiple versions of a video encoded at different bitrates are merged into a single file. Some implementations of this can be very inefficient, in that each stream is self-contained and doesn't share any information from the other representations of the media. Streaming media players can detect connection bandwidth dynamically and switch among the streams as appropriate. While crude, this improves the situation over the case where the user must select from separate files based on their connection bandwidth. Most users don't have a good understanding of their connection bandwidth in the first place and requiring a selection choice is poor system design which can lead to errors if the wrong setting is chosen.

3.3.7 Illustrated Audio

Illustrated audio is a class of content that fills the gap between full motion video and a bare audio stream. There are two main classes of this; the first is frame flipping where a single still image is displayed at a given point in time until the next event where a different frame is displayed. This can be thought of as non-uniform sampling: instead of each frame being displayed for the same amount of time, e.g. 33 ms, a frame may be displayed for 20 seconds followed by a frame displayed for 65 seconds, etc. An example of this is a recording of a lecture containing slides. The second class involves some form of gradual transition between slides and may include synthesized camera operations such as panning and zooming. Some replay sys-

tems for digital photographs employ this technique using automatically selected operations. Readers may also be familiar with the historical documentary style of Ken Burns where old photographs are seemingly brought to life through appropriate narration and synthesized camera operations [Burns07].

The value of this form of content has justified the creation of systems designed for efficiently compressing and representing this unique material. Microsoft's Photo Story application allows manual creation of slide shows from still frames and encodes the result in Windows media format with a special codec. Alternatively, the Windows Media Format allows for synchronous events to be included in the stream which may include links to images or may encode generic events that can be accessed via client JavaScript at media replay time to take action (which may also include fetching an image from a URL and displaying it).

Apple's Enhanced Podcasts are MPEG-4 files with streams containing specific information that allows for the inclusion and synchronized replay of embedded still images (as well as other information such as links) that can be replayed on iPods. These files typically have the extension .m4a or .m4b. The points in the media where the images are inserted naturally form waypoints for navigating in the content, and Apple emphasizes this by referring to these points as chapter markers and exposing this up through the user interface of iTunes® and iPods®. Other formats also support chapter metadata such as ID3v2 which specifies CHAP (Chapter) and CTOC (Table of Contents) and the DVD specification. In Flash video, the "Cue Point" mechanism is used for synchronizing loading of graphics and providing for navigation of the media.

For video search engines, textual chapter metadata can augment the global metadata and can improve relevance ranking and navigation for systems that support navigating within long form content. Additionally, where archiving systems manage wide varieties of content and adapt it to produce content for consumption scenarios where the primary media track is audio (i.e. mobile listening), the ability to automatically insert chapter markings to aid user navigation is extremely valuable.

3.4 Media Captioning

We have already seen how captioning can be exploited for video search, but further, video search engine systems and IP media systems should preserve any captioning that accompanies the ingested source media in order

to reach the broadest possible audience. Again, it is important to point out that captioning is not just for the hearing impaired, but can improve comprehension and enable media consumption in a wider range of environments (e.g. meetings). Most IP media formats support some form of timed text and these were covered in detail in Chapter 2. The National Center for Accessible Media at WGBH pioneered television captioning [Robson97] and has recently formed the Internet Captioning Forum with industry leaders. The Distribution Format Exchange Profile (DFXP) is a subset of the Timed Text Authoring format intended to aid in interoperability of existing legacy formats. While its scope is limited, the specification includes enough generality to support a very wide range of existing captioning presentations (perhaps only exclusive of sign language representations) so it is not trivial by any means [TT06].

3.5 Conclusion

We have presented many of the practical aspects of digital video that content-based video search engine systems must deal with in order to operate seamlessly on a wide variety of content sources. At the basic level, issues of encoding and container file formats, and DRM systems must be taken into account in the system design. Next, presentation issues such as aspect ratio and transcoding for archival storage and delivery for a range of applications must be considered in the design of user interfaces for search. We also introduced methods for creating networked user interfaces for media replay with thin clients such as media players with dynamically generated playlists or browser plug-ins. Beyond the basic input and output media handling and rendering, systems that operate on the video content must also deal with real-world issues such as subsampled, noisy chrominance, non-square pixels and various temporal sampling rates. While a theoretician might correctly dismiss many of these issues as engineering decisions arising from legacy (or worse, commercially motivated proprietary and incompatible) implementations, some are related to basic principles or physical properties. There are limits to the fractional bits per pixel to which video can be compressed and the signal to noise ratio of imaging devices.

References

[ATIS06] Status Report on the work of the ATIS ITPV Interoperability Forum (IIF), ATIS / ITU, Document 34-E, Geneva (2006).

[Apple07] QuickTime File Format Specification, Apple Computer, Inc. (2007).

[Rich03] Richardson, I.: *H.264 and MPEG-4 Video Compression: Video Coding for Next-generation Multimedia*, Wiley, Chichester, West Sussex, England (2003).

[Bayer76] US Patent 3,971,065 Bryce E. Bayer Color imaging array, July 20, 1976.

[Haskell07] Haskell, B. et al.: *Digital Video: An Introduction to MPEG-2,* Chapman & Hall, New York (1997).

[Info00] Information Technology – Generic Coding of Moving Pictures and Associated Audio Information: Systems, ISO/IEC, International Standard 13818-1, 2nd ed., December 1, 2000.

[Burns07] Burns, K.: Museum of Broadcast Television article, http://www.museum.tv/archives/etv/B/htmlB/burnsken/burnsken.htm Encyclopedia of TV, 2nd ed., cited 11 January 2007.

[TT06] Timed Text (TT) Authoring Format 1.0 – Distribution Format Exchange Profile (DFXP) W3C Recommendation, November 16, 2006.

[Robson97] Robson, G., *Inside Captioning,* Cyberdawg Publishing (1997).

http://www.springer.com/978-3-540-79336-6

Introduction to Video Search Engines

Gibbon, D.C.; Liu, Z.
2008, XVI, 276 p. 79 illus., Hardcover
ISBN: 978-3-540-79336-6