Representation and Coding Formats for Stereo and Multiview Video

Anthony Vetro

TR2010-011 April 2010

Abstract

This chapter discusses the various representation and coding formats for stereo and multiview video in the context of next-generation 3D video services. Several application scenarios are discussed including packaged media such as Blu-ray Disc, as well as the delivery over cable, terrestrial and Internet channels. The various types of 3D displays are also described and the data requirements for each examined. A review of different representation formats for stereo and multiview video is given including multiplexed formats, full-channel formats and depth-based formats. The corresponding compression technology for these formats is then described. The chapter concludes with a discussion of future outlooks and research challenges.

Representation and Coding Formats for Stereo and Multiview Video

Anthony Vetro

Mitsubishi Electric Research Laboratories, USA avetro@merl.com

Summary. This chapter discusses the various representation and coding formats for stereo and multiview video in the context of next-generation 3D video services. Several application scenarios are discussed including packaged media such as Blu-ray Disc, as well as the delivery over cable, terrestrial and Internet channels. The various types of 3D displays are also described and the data requirements for each examined. A review of different representation formats for stereo and multiview video is given including multiplexed formats, full-channel formats and depth-based formats. The corresponding compression technology for these formats is then described. The chapter concludes with a discussion of future outlooks and research challenges.

1 Introduction

Delivering higher resolution video and providing an immersive multimedia experience to the home has been a primary target for industry and researchers in recent years. Due to advances in display technology, signal processing and circuit design, the ability to offer a compelling 3D video experience on consumer electronics platforms has become feasible. In addition to advances on the display side, there has also been a notable increase in the production of 3D contents. The number of title releases has been steadily growing each year; a number of major studios have announced all future releases in 3D. There are also substantial investments being made in digital cinema theaters with 3D capability. The push from both production and display side have played a significant role in fuelling a renewed interest in 3D video.

There are a number of challenges to make 3D video a practical and sustainable service that consumers will enjoy. For instance, it will be essential to determine an appropriate data format for delivery under different constraints. Interoperability among various devices will be essential. It will also be critical to ensure that the consumer has a high quality experience and is not turned off by viewing discomfort or fatigue, or the need to wear special glasses.

This chapter discusses the various representation and coding formats for stereo and multiview video that are available or being studied, which could be used to drive next-generation 3D video services. Several application domains and distribution environments are discussed including packaged media such as Blu-ray Disc, as well as the delivery over cable, terrestrial and Internet channels. The various types of 3D displays are also described and the data requirements for each examined. A review of different representation formats for stereo and multiview video is given including multiplexed formats, full channel formats and depth-based formats. The corresponding compression technology for these formats is then described with particular focus on the recently finalized extension of the H.264/MPEG-4 Advanced Video Coding (AVC) standard [1] on multiview video coding. Future outlooks and research challenges are also discussed.

2 Application Domains

A diagram illustrating a system level view of the end-to-end delivery chain for 3D video, including production, distribution and consumption, is shown in Figure 1. Each of these domains will be discussed further in the following subsections.

2.1 Production

The main approaches to creating 3D content include camera capture, computer generated, and conversion from 2D video. Most 3D video that is captured use stereo cameras. While some multi-camera setups exist, large camera arrays coupled with the increased amount of data currently prohibit widespread use of multi-camera capture. Computer generated content is much easier to produce since scene information such as depth is an inherent part of production process, and the depth information enables great flexibility in editing and repurposing of the content. Yet another way to create 3D video is by taking conventional 2D video and adding depth information. The typical process is to identify a series of objects from the 2D image, assigning relative depth to each object, then fill in occluded areas. The creation of a depth map from 2D allows for the creation of multiple views through a rendering process.

An important element of the production domain is a master format. Whether the content is a 3D cinema production or a live event, the master format specifies a common image format along with high level metadata that are required to make sense of the data and prepare the data for distribution. The format is generally independent of any specific delivery channel.

In August 2008, the Society of Motion Picture and Television Engineers (SMPTE) formed a task force to investigate the production requirements to realize 3D video to the home. After a 6-month study, the final report of the task force recommended standardization of a *3D Home Master* which would essentially be an uncompressed and high-definition stereo image format, i.e., 1920×1080 pixel resolution at 60Hz per eye [2]. The mastering format will also

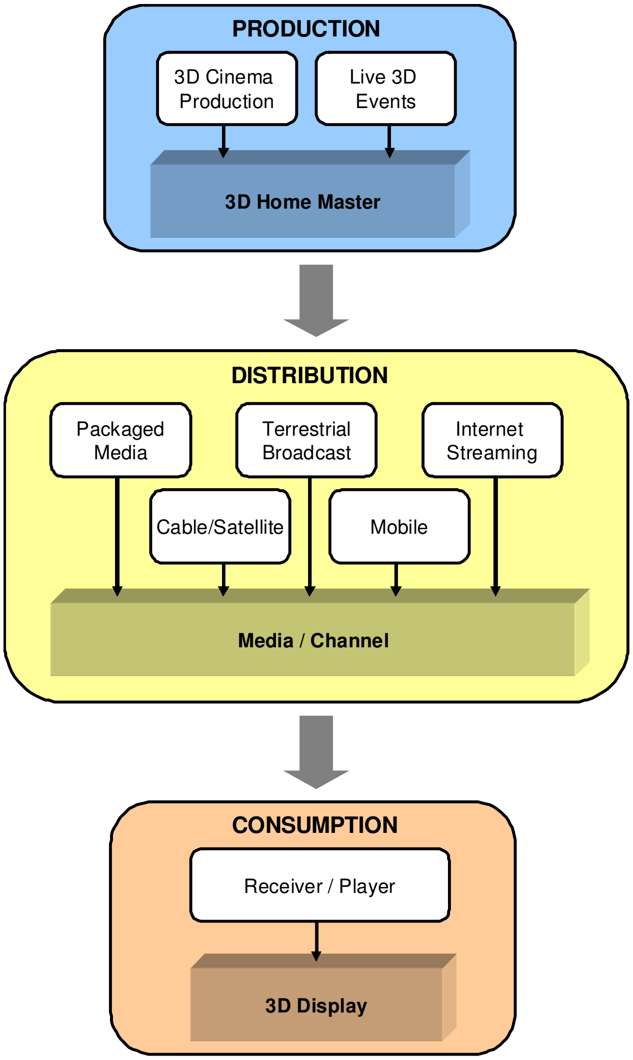


Fig. 1. Illustration of major areas and components in an end-to-end 3D video delivery chain, including production, distribution and consumption.

specify metadata, e.g., signaling of left and right image frames, as well as scene information such as the maximum and minimum depth of a scene. The master format is also expected to include provisions to associate supplementary data such as pixel-level depth maps, occlusion and transparency data. SMPTE expects to finalize the mastering standard for 3D to the home by the end of 2010.

2.2 Distribution

It is expected that 3D content will reach the home through a variety of different channels, including packaged media such as Blu-ray Disc, through cable or terrestrial broadcast, as well as Internet streaming or download. It is an open question whether the delivery formats between each of these distribution channels could be harmonized given the unique constraints associated with each. The key challenges of each distribution channel are discussed in the following.

Blu-ray discs have a recording capacity of 50 GB, which is more than enough for feature-length movies in a high-definition format, which is typically 1920×1080 pixels at 24Hz. There is also sufficient capacity to store additional streams, which allows for significant flexibility in the way that 3D content is stored on Blu-ray discs. One option is to encode a second view independently of the 2D view and store it separately; combing the two views would then offer a stereoscopic viewing experience. To reduce the bit rate, the second view may also be compressed based on the 2D view. A separate 3D stream that subsamples and multiplexes both the left and right views into a single frame could also be encoded and stored; this format would have the advantage of working with existing 2D players, but sacrifices quality. These different formats will be discussed in greater detail in sections 3 and 4. One of the major issues with delivery of 3D content on Blu-ray discs is backward compatibility with existing players. This means that the disc containing both 2D and 3D content should have no problems playing 2D content on legacy devices that do not have explicit support for 3D.

It is expected that packaged media will be one of the first ways for consumers to receive premium 3D content in the home, but other delivery means are very likely to follow. For instance, the delivery of 3D video services through cable seems very promising [3]. Cable operators may offer premium channels with 3D video as part of their line up or offer 3D video through on-demand services. While bandwidth is not a major issue in the cable infrastructure, the set-top boxes to decode and format the content for display is a concern. A 3D format that is compatible with existing set-top boxes would enable faster deployment of new 3D services; a multiplexed format could be useful for this purpose. New boxes could also be introduced into the market with full resolution and 3D capabilities. This tradeoff between deployment speed and 3D capabilities is part of the ongoing discussion among cable operators and within the Society of Cable Telecommunications Engineers (SCTE), which is the standards organization that is responsible for cable services.

Terrestrial broadcast is perhaps the most constrained distribution method. Among the organizations responsible for digital televisions broadcast are the Advanced Television Systems Committee (ATSC) in the US, the Digital Video Broadcast (DVB) Project in Europe and the Association of Radio Industries and Businesses (ARIB) in Japan. Many analog systems around the world have converted or are in the process of converting to all digital broadcast, where the allocated bandwidth for each channel is fixed and somewhat limited. Furthermore, most countries around the world defined their digital broadcast services based on the MPEG-2 standard, which is often a mandatory format in each broadcast channel. Between the limited channel bandwidth, the legacy format issues, costs associated with upgrading broadcast infrastructure and the lack of a clear business model on the part of the broadcasters to introduce 3D services, over-the-air broadcast of 3D video is likely to lag behind other distribution channels.

With increased broadband connectivity in the home, access to 3D content from web servers is likely to be a dominant source of content. There are already media servers on the market that ensure sufficient processing capability and interfaces to support 3D video, including interactive gaming media, and such capabilities could integrated into the television or other home appliances in the near future. Existing online services that offer content as part of a subscription or charge a fee for download or streaming of video are likely to add 3D services to their offerings in the near future. Mobile phone services are also expected to introduce 3D capable devices and offer 3D content as well.

In summary, while there are many obstacles in delivering 3D content to the home, there are also several viable options. With an increased demand for 3D in the home combined with the push for an additional revenue stream by services providers, we should be able to expect the availability of some 3D services in the very near future.

2.3 Consumption

There are a wide variety of different 3D display systems designed for the home user applications, starting from classical two-view stereo systems that require glasses to more sophisticated auto-stereoscopic displays that do not require glasses [4]. Auto-stereoscopic displays emit a large number of views, but the technology ensures that users only see a stereo pair from any particular viewpoint. In the following, some of the more prominent display technologies and their corresponding data format requirements are outlined.

There are two major categories of stereo displays: passive and active. The main difference between these devices is the way in which the light for each eye is emitted and viewed.

With passive devices, the images for left and right eyes are polarized in an orthogonal direction and superimposed on a display screen. To view the 3D scene, a pair of polarized glasses are used that match the orientation of the views being displayed for left and right eyes. In this way, the glasses are used to separate or filter the light with a specific orientation, i.e., only light with an orientation that matches that of the lens will be able to pass through. As a result, images that correspond to the left and right views can be properly viewed by each eye to achieve a 3D viewing effect.

There are a number of passive displays based on LCD technology available today. These displays are implemented using alternating lines for left and right views with opposite polarization. The native display format in this case is referred to as row interleaving.

In contrast to passive displays, active displays operate by rapidly alternating between the left-eye and the right-eye image on the screen. To maintain motion continuity, a frame rate of 120Hz is typically used. Active shutter glasses have lenses which turn from opaque to transparent, e.g., based on liquid crystals, in perfect sync with the image being displayed. To maintain synchronization, an active signal must be sent from the display to the glasses. Infrared emitters are typically used for this purpose.

Currently, active displays are implemented based on DLP or plasma technology, where each left and right frame are displayed at alternating time instants. The native display format for some active displays based on plasma technology is frame sequential, which outputs a full spatial resolution for each eye. The native display format for all DLP-based displays and some plasma displays is checkerboard, which is essentially applies a quincunx sub-sampling to each view. Next-generation LCD panels with higher frame rates may also enter the active display category.

Auto-stereoscopic displays do not require glasses and achieve 3D by essentially emitting view-dependent pixels. Such displays can be implemented, for example, using conventional high-resolution displays and parallax barriers; other technologies include lenticular sheets and holographic screens [4]. Each view-dependent pixel can be thought of as emitting a small number of light rays in a set of discrete viewing directions, typically between eight and a few dozen. Often these directions are distributed in a horizontal plane, such that parallax effects are limited to horizontal motion of the observer. Obviously, these multiview displays have much higher data requirements relative to conventional stereo displays that only require two views.

Since there are a number of displays already on the market that use different formats, the interface from distribution formats to native displays formats is a major issue. There is currently a strong need to standardize the signaling and data format to be transmitted from the various source devices in the home such as TV receivers and players, to the many types of sink devices, i.e., displays. HDMI v1.4 has recently been announced and includes support for a number of uncompressed 3D formats [5]. Efforts are underway to also update other digital interface specifications including those specified by the Consumer Electronics Associations (CEA). There are also new initiatives within CEA to standardize the specification of 3D glasses, as well as the interface between display devices and active glasses [6].

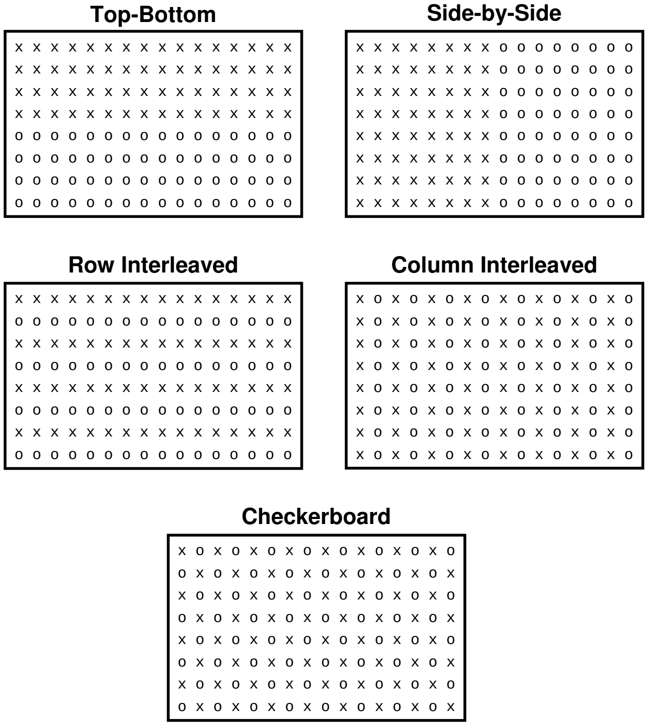
3 Representation Formats

3.1 Full-Resolution Stereo and Multiview

The format that most people first think of for stereo and multiview video are full-resolution formats. In the case of stereo, this representation basically doubles the data rate of conventional single view video. For multiview, there is an N-fold increase in the data rate for N-view video. Efficient compression of such data is a key issue and will be discussed further in Section 4.

3.2 Stereo Interleaving

Stereo interleaving is a class of formats in which the stereo signal is essentially a multiplex of the two views into a single frame or sequence of frames. Some common spatial interleaving formats are shown in Figure 2. Another common name for such representation formats are *frame-compatible* formats.



Top-Bottom

```
x x x x x x x x x x x x x x x x
x x x x x x x x x x x x x x x x
x x x x x x x x x x x x x x x x
x x x x x x x x x x x x x x x x
o o o o o o o o o o o o o o o o
o o o o o o o o o o o o o o o o
o o o o o o o o o o o o o o o o
o o o o o o o o o o o o o o o o
```

Side-by-Side

```
x x x x x x x x o o o o o o o o
x x x x x x x x o o o o o o o o
x x x x x x x x o o o o o o o o
x x x x x x x x o o o o o o o o
x x x x x x x x o o o o o o o o
x x x x x x x x o o o o o o o o
x x x x x x x x o o o o o o o o
x x x x x x x x o o o o o o o o
```

Row Interleaved

```
x x x x x x x x x x x x x x x x
o o o o o o o o o o o o o o o o
x x x x x x x x x x x x x x x x
o o o o o o o o o o o o o o o o
x x x x x x x x x x x x x x x x
o o o o o o o o o o o o o o o o
x x x x x x x x x x x x x x x x
o o o o o o o o o o o o o o o o
```

Column Interleaved

```
x o x o x o x o x o x o x o x o
x o x o x o x o x o x o x o x o
x o x o x o x o x o x o x o x o
x o x o x o x o x o x o x o x o
x o x o x o x o x o x o x o x o
x o x o x o x o x o x o x o x o
x o x o x o x o x o x o x o x o
x o x o x o x o x o x o x o x o
```

Checkerboard

```
x o x o x o x o x o x o x o x o
o x o x o x o x o x o x o x o x
x o x o x o x o x o x o x o x o
o x o x o x o x o x o x o x o x
x o x o x o x o x o x o x o x o
o x o x o x o x o x o x o x o x
x o x o x o x o x o x o x o x o
o x o x o x o x o x o x o x o x
```

Fig. 2. Common spatial interleaving formats, where `x` represents the samples from one view and `o` represents samples from the another view.

With a spatially multiplexed format, the left and right views are sub-sampled and interleaved into a single frame. There are a variety of options for both the sub-sampling and interleaving. For instance, a quincunx sampling may be applied to each view and the two views interleaved with alternating samples in both horizontal and vertical dimensions. Alternatively, the two views may be decimated horizontally or vertically and stored in a side-by side or top-bottom format, respectively.

With a time multiplexed format, the left and right views would be interleaved as alternating frames or fields. These formats are often referred to as frame sequential and field sequential. The frame rate of each view may be reduced so that the amount of data is equivalent to a that of a single view.

A major advantage of such formats is that the stereo video is represented in such a way that is compatible with existing codecs and delivery infrastructure. In this way, the video can be compressed with existing encoders, transmitted through existing channels and decoded by existing receivers and players. This format essentially tunnels the stereo video through existing hardware and delivery channels. Due to these minimal changes, stereo services can can be quickly deployed to capable display, which are already in the market.

The drawback of representing the stereo signal in this way is that spatial or temporal resolution would be lost. However, the impact on the 3D perception may be limited. An additional issue with interleaving formats is distinguishing the left and right views. To perform the de-interleaving, some additional out-of-band signaling is necessary. Since this signalling may not be understood by legacy receivers, it is not possible for such devices to extract, decode and display a 2D version of the 3D program. While this might not be so problematic for packaged media since special 3D players could be upgraded and new discs might be able to store both 2D and 3D formats, it is certainly a major issue for broadcast services where the transmission bandwidth is limited and devices cannot be upgraded.

The signalling for a complete set of interleaving formats has been standardized within the H.264/MPEG-4 AVC standard as Supplementary Enhancement Information (SEI). In general, SEI messages provide useful information to a decoder, but are not a normative part of the decoding process. An earlier edition of the standard already specified a Stereo SEI message that identifies the left view and right view; it also has the capability of indicating whether the encoding of a particular view is self-contained, i.e., frame or field corresponding to the left view are only predicted from other frames or fields in the left view. Inter-view prediction for stereo is possible when the self-contained flag is disabled. This functionality has been combined with additional signaling for the various spatially multiplexed formats described above as a new SEI message referred to as the Frame Packing Arrangement SEI message. This new SEI message has recently been specified as an amendment of the AVC standard [7].

3.3 Depth-based Formats

ISO/IEC 23002-3 (also referred to as MPEG-C Part 3) specifies the representation of auxiliary video and supplemental information. In particular, it enables signaling for depth map streams to support 3D video applications. Specifically, the 2D plus depth format is specified by this standard and is illustrated in Figure 3. The inclusion of depth enables a display-independent

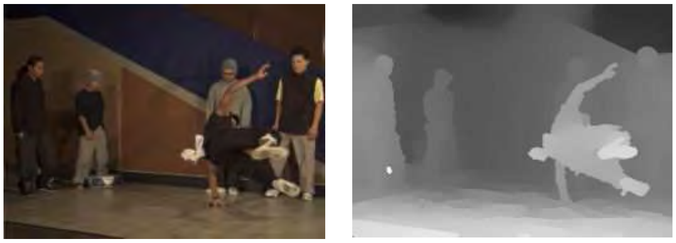


Fig. 3. Illustration of the 2D plus depth format.

solution for 3D that supports generation of an increased number of views as need by any stereoscopic display.

A key advantage of this representation format is that the main 2D video provides backward compatibility with legacy devices. Also, this representation is agnostic of the actual coding format, i.e., the approach works with both MPEG-2 and H.264/MPEG-4 AVC video coding standards. In principle, this format is able to support both stereo and multiview displays, and also allows adjustment of depth perception in stereo displays according to viewing characteristics such as display size and viewing distance.

The main drawback of this format is that it is only capable of rendering a limited depth range and was not specifically designed to handle occlusions. Also, stereo signals are not easily accessible by this format, i.e., receivers would be required to generate the second view to drive a stereo displays, which is not the convention in existing displays. Finally, while automatic depth estimation techniques have been a heavily explored topic in the literature, their accuracy is still not sufficient to support the synthesis and rendering requirements of this format. Therefore, some semi-automatic means are needed to extract depth maps with sufficient accuracy, which could add substantially to production costs and may not be practical for live events.

To overcome the drawbacks of the 2D plus depth format, while still maintaining some of its key merits, MPEG is now in the process of exploring alternative representation formats and considering a new phase of standardization. The targets of this new initiative were discussed in [8] and are also illustrated in Figure 4 [9]. The objectives are:

1. Enable stereo devices to cope with varying display types and sizes, and different viewing preferences. This includes the ability to vary the baseline distance for stereo video so that the depth perception experienced by the viewer is within a comfortable range. Such a feature could help to avoid fatigue and other viewing discomforts.
2. Facilitate support for high-quality auto-stereoscopic displays. Since directly providing all the necessary views for these displays is not practical due to production and transmission constraints, the new format aims to enable the generation of many high-quality views from a limited amount of input data, e.g. stereo and depth.

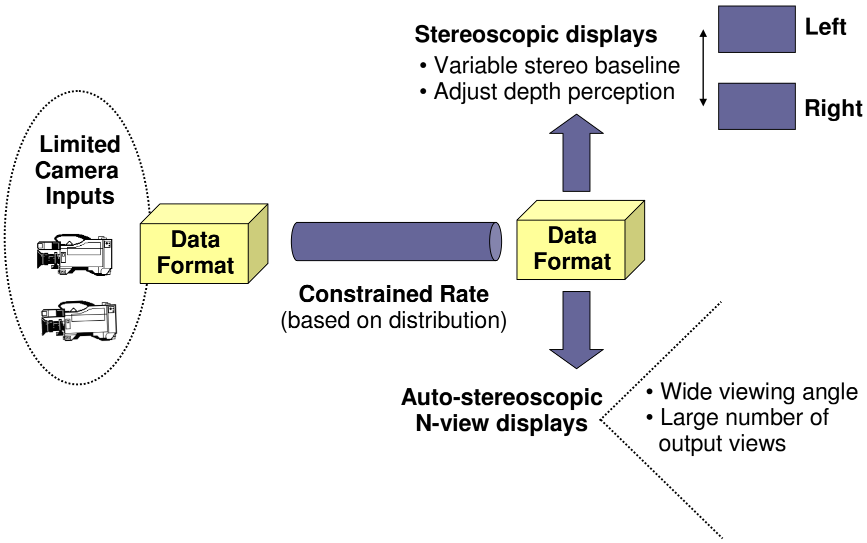


Fig. 4. Target of 3D video framework including limited camera input, fixed rate transmission and capability to support auto-stereoscopic displays and advanced stereoscopic processing.

A key feature of this new 3D video (3DV) data format is to decouple the content creation from the display requirements, while still working within the constraints imposed by production and transmission. Furthermore, compared to the existing coding formats, the 3DV format aims to enhance 3D rendering capabilities beyond 2D plus depth, while not incurring a substantial rate increase. Simultaneously, at an equivalent or improved rendering capability, this new format should substantially reduce the rate requirements relative to sending multiple views directly. These requirements are outlined in [10].

4 Compression Technology

To realize an efficient coded realization of the 3D representation formats discussed in the previous section, compression of the scene data is required. This section describes related compression techniques and standards. In particular, the multiview video coding extension of the H.264/MPEG-4 AVC standard [1] is reviewed. Coding tools that have been proposed for efficient coding of multiview video, but have not been adopted to the standard, are covered as well. Finally, specific techniques for coding of depth map information are described.

4.1 Multiview Video Coding Standard

A straightforward means to represent stereo or multi-view video is independent encoding of each view. This solution has low complexity since depen-

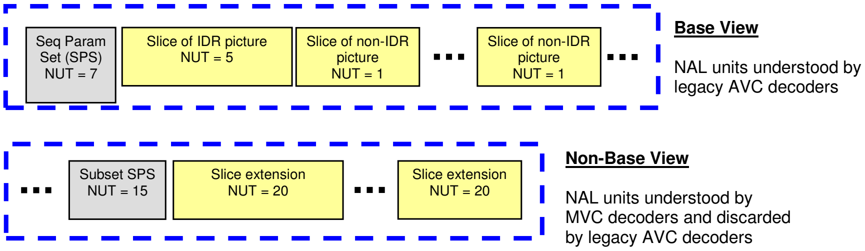

Seq Param Set (SPS) NUT = 7
Slice of IDR picture NUT = 5
Slice of non-IDR picture NUT = 1
Slice of non-IDR picture NUT = 1
Base View
NAL units understood by legacy AVC decoders
Subset SPS NUT = 15
Slice extension NUT = 20
Slice extension NUT = 20
Non-Base View
NAL units understood by MVC decoders and discarded by legacy AVC decoders

Fig. 5. Base structure of an MVC bitstream including NAL units that are associated with a base view and NAL units that are associated with a non-base view.

dencies between views are not exploited, thereby keeping computation and processing delay to a minimum. It also enables backward compatibility with existing 2D solutions since one of the views could be easily accessed and decoded for legacy displays. The main drawback of the simulcast method is that coding efficiency is not maximized since redundancy between views is not considered.

To improve coding efficiency of multiview video, the redundancy over time and across views could be exploited. In this way, pictures are not only predicted from temporal neighbors, but also from spatial neighbors in adjacent views. This capability has been enabled most recently as the Multiview Video Coding (MVC) extension of the H.264/MPEG-4 AVC standard [1]. Several key features of the standard are reviewed below.

Bitstream structure

A key aspect of the MVC design is that it is mandatory for the compressed multiview stream to include a base view bitstream, which is independently coded from other non-base (or enhancement) views. Such a requirement enables a variety of uses cases that need a 2D version of the content to be easily extracted and decoded. For instance, in television broadcast, the base view would be extracted and decoded by legacy receivers, while newer 3D receivers could decode the complete 3D bitstream including non-base views.

As defined in the AVC standard, there exists a Network Abstraction Layer (NAL) and coded data is organized into NAL units. There exist many types of NAL units, some which are designated for video coding data, while others for non-video data such as high level syntax information. MVC extends the NAL unit types used for single video to provide backward compatibility for multiview video.

To achieve this compatibility, the video data associated with a base view is encapsulated in NAL units defined for single view video, and the video data associated with additional views are encapsulated in a new NAL unit type for multiview video. The base view bitstream conforms to existing AVC profiles for single view video, e.g., High profile, and decoders conforming to an

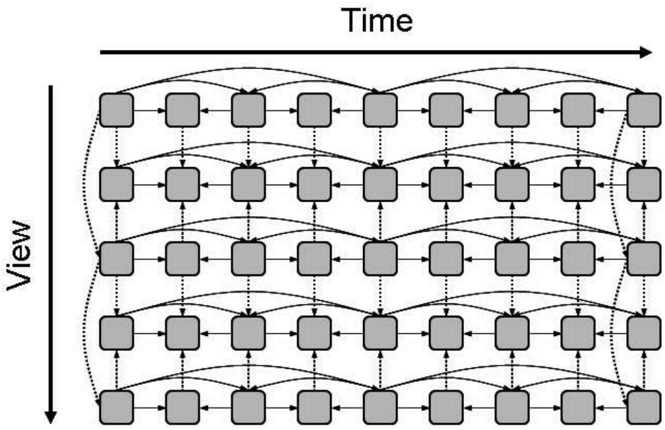


Fig. 6. Illustration of inter-view prediction in MVC.

existing single view profile will ignore and discard NAL units corresponding to the multiview data since it would not recognize those NAL unit types. Decoding the additional views with these new NAL unit types would require a decoder that conforms to one of the MVC profiles and recognizes the new NAL unit types. The basic structure of the MVC bitstream including NAL units associated with a base view and NAL units associated with a non-base view is shown in Figure 5. Further discussion of the high-level syntax is given below, and MVC profiles and levels are discussed further below.

Inter-view prediction

The basic idea of inter-view prediction, which is employed in all related works on efficient multiview video coding, is to exploit both spatial and temporal redundancy for compression. Since all cameras capture the same scene from different viewpoints, inter-view redundancy is present. A sample prediction structure is shown in Fig. 6. Pictures are not only predicted from temporal references, but also from inter-view references. The prediction is adaptive, so the best predictor among temporal and inter-view references is selected on a block basis.

Inter-view prediction is a key feature of the MVC design and is enabled through flexible reference picture management of AVC, where decoded pictures from other views are essentially made available in the reference picture list. Specifically, a reference picture list is maintained for each picture to be decoded in a given view. This list is initialized as usual for single view video and would include any temporal references that may be used to predict the current picture. Additionally, inter-view reference pictures may be appended to the list and thereby available for prediction of the current picture.

According to the MVC specification, inter-view reference pictures must be contained within the same access unit as the current picture, where an access unit contains all the NAL units pertaining to a certain time instance. In other words, it is not possible to predict a picture in one view at a given time from a picture in another view at a different time. This would involve inter-view prediction across different access units, which would incur additional complexity for limited coding benefits.

To keep the management of reference pictures inline with single view video, all the memory management control operation commands that may be signalled through an AVC bitstream apply to a particular view. The same is true for the sliding window that is used to mark pictures as not being used for reference; this process of AVC is also applied independently for each view. Also, just as it is possible to re-order the reference pictures in a reference picture list including temporal references, the same can be done with reference picture lists including inter-view references. An extended set of re-ordering commands have been adopted to the MVC specification for this purpose.

It is important to emphasize that the core block-level decoding modules do not need to be aware of whether a reference picture is a temporal reference or an inter-view reference. This distinction is managed at a higher level of the decoding process.

In terms of syntax, the standard only requires small changes to high-level syntax, e.g., view dependency as discussed below. A major consequence of not requiring changes to lower block-level syntax is that MVC is compatible with existing hardware for decoding single view video with AVC. In other words, supporting MVC as part of an existing AVC decoder should not require substantial design changes.

Since MVC introduces dependencies between views, random access must also be considered in the view dimension. Specifically, in addition to the views to be accessed (target views), any dependent views also need to be accessed and decoded, which typically requires decoding time or delay. For applications in which random access or view switching is important, the prediction structure could be designed to minimize access delay.

To achieve access to a particular picture in a given view, the decoder should first determine an appropriate access point. In AVC, Instantaneous Decoding Refresh (IDR) pictures provide a clean access point since these pictures can be independently decoded and all the coded pictures that follow in decoding order can be decoded without temporal prediction from any picture decoded prior to the IDR picture. In the context of MVC, an IDR picture prohibits the use of temporal prediction for any of the views at that particular instant of time; however, inter-view prediction may be used to reduce the rate overhead. MVC also introduces a new picture type, referred to as an anchor picture. Anchor pictures are similar to IDR pictures in that they do not utilize temporal prediction, but do allow inter-view prediction from views within the same access unit. The difference between anchor pictures and IDR pictures is similar to the difference between the open GOP and closed GOP concept in MPEG-2, respectively, where closed GOPs do not allow pictures from one GOP to be used as a reference for pictures in a different GOP. In contrast, open GOPs effectively allow the I-frame of one GOP to be used as a backward reference for a B-frame that is earlier in display order. In MVC, both IDR and anchor pictures are efficiently coded and provide random access in both time and view dimensions.

High-level Syntax

The decoding process of MVC requires several additions to the high-level syntax, which are primarily signalled through a multiview extension of the Sequence Parameter Set (SPS) defined by AVC. Three important pieces of information are carried in the SPS extension:

- View identification
- View dependency information
- Level index for operation points

The level index is an indicator of resource constraints imposed on a decoder that conforms to a particular level; it is mainly used to bound the complexity of a decoder and discussed further below. In the context of MVC, an operating point corresponds to a specific temporal level and a set of views including those intended for output and the views that they depend on.

The view identification part includes an indication of the total number of views, as well as a listing of view identifiers. The view identifiers are important for associating a particular view to a specific index, while the order of the views identifiers signals the view order index. The view order index is critical to the decoding process as it defines the order in which views are decoded.

The view dependency information is comprised of a set of signals that indicate the number of inter-view reference for each of the two reference picture lists that are used in the prediction process, as well as the views that may be used for predicting a particular view. Separate information is maintained for anchor and non-anchor pictures to provide some flexibility in the prediction, while not over-burdening decoders with dependency information that could change for each unit of time.

For non-anchor pictures, the view dependency only indicates that a given view may be used for inter-view prediction. There is additional signaling in the NAL unit header indicating whether a particular view at a given time is used as an inter-view reference for any other picture in the same access unit. The view dependency in the SPS and this syntax element in the NAL unit header are used to append the reference picture list to include inter-view references as described earlier.

The final portion of the SPS extension is the signalling of level information and associated information about the operating points to which they correspond. An MVC stream with 8 views may include several operating points, e.g., one corresponding to all 8 views, another corresponding to a stereo pair,

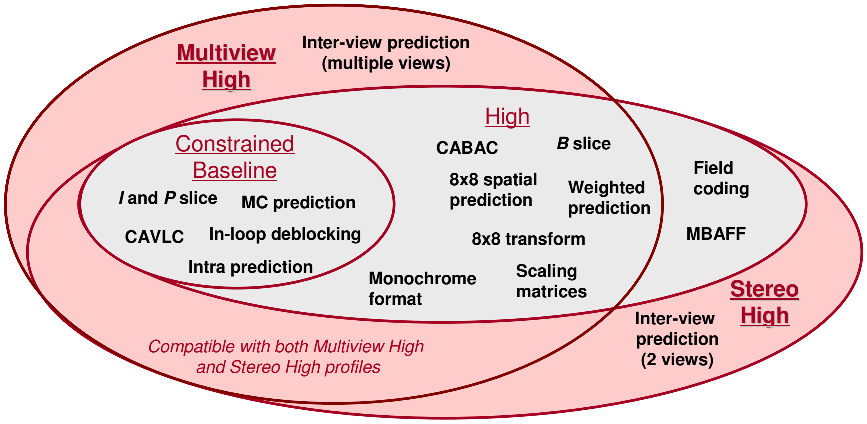


Fig. 7. Illustration of MVC profiles including Multiview High and Stereo High profiles, as well as region that is compatible with both profiles.

and another corresponding to a set of three particular views. According to the MVC standard, multiple level values could be signalled as part of the SPS extension, with each level being associated to a particular operating point (i.e., temporal level and target set of views). The syntax indicates the number of views that are targeted for output as well as the number of views that would be required for decoding particular operating points.

Profiles and levels

Consistent with prior MPEG standards, profiles determine the subset of coding tools that must be supported by decoder. There are two profiles currently defined by MVC with support for more than one view: Multiview High profile and Stereo High profile. Both are based on the High profile of AVC with a few differences.

- The Multiview High profile supports multiple views and does not support interlaced coding tools.
- The Stereo High profile is limited to two views, but does support interlaced coding tools.

An illustration of these profile specifications relative to High profile of AVC is provided in Figure 7. It is possible to have a bitstream that conforms to both the Stereo High profile and Multiview High profile when there are only two views are coded and interlaced coding tools are not used. In this case, a flag signaling their compatibility is set.

Levels impose constraints on decoder resources and complexity. A similar set of limits as imposed on AVC decoders are imposed on MVC decoders including limits on the amount of frame memory required for the decoding of a bitstream, the maximum throughput in terms of macroblocks per second, maximum picture size, overall bit rate, etc.

The general approach to defining level limits in MVC was to enable repurposing the decoding resources of single-view decoders for multi-view decoders. In this way, some level limits are unchanged such as the overall bit rate; in this way, an input bitstream can be processed by a decoder regardless of whether it encodes a single view or multiple views. However, other level limits are increased such as the maximum decoded picture buffer and throughput; a fixed scale factor of 2 was applied to these decoder parameters. Assuming a fixed resolution, this scale factor enables decoding of stereo video using the same level as single view video at that resolution. For instance, the same Level 4.0 index could be used to decode single view video and stereo video at 1920×1080p at 24Hz. To decode a higher number of views, one would need to use either a higher level and/or reduce the spatial or temporal resolution of the multiview video.

Coding Performance

It has been shown that coding multiview video with inter-view prediction does give significantly better results compared to independent coding [11, 12]. For some cases, gains as high as 3 dB, which correspond to 50% savings in bit rate, have been reported. A comprehensive set of results for multiview video coding over a broad range of test material was also presented in [13]. For multiview video with up to 8 views, an average of 20% improvement relative to the total simulcast bit rate has been reported. In other studies [14], an average reduction of 20-30% of the bit rate for the second (dependent) view of typical stereo movie content was reported, with peak reduction up to 43% of the bit rate of the dependent view.

Fig. 8 shows sample RD curves comparing the performance of simulcast coding (without the use of hierarchical B-pictures) with the performance of the MVC reference software that employs hierarchical predictions in both spatial and view dimensions. Of course, the use of hierarchical B-pictures in the simulcast solution will also provide some gains, but they are not shown in this plot.

There are many possible variations on the prediction structure considering both temporal and spatial dependencies. The structure not only affects coding performance, but has notable impact on delay, memory requirements and random access. It has been confirmed that the majority of gains are obtained using inter-view prediction at anchor positions in Fig. 8. Rate penalties of approximately 5-15% could be expected if the spatial predictions at non-anchor positions are removed [15]. The upside is that delay and required memory would also be reduced.

Prior studies on asymmetrical coding of stereo, whereby one of the views is encoded with less quality, suggest that substantial savings in bitrate for

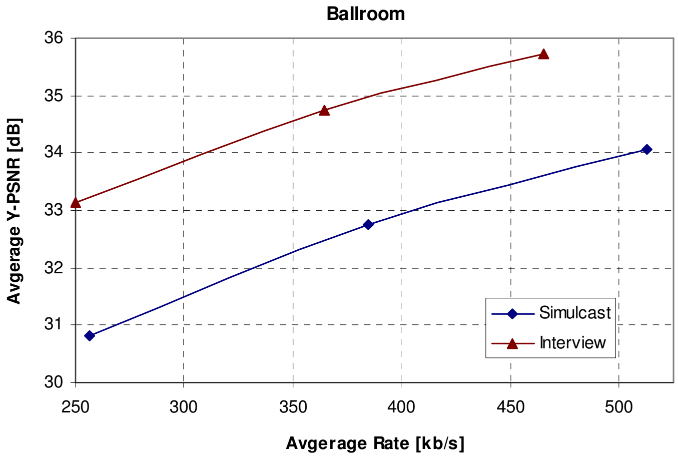


Fig. 8. Illustration of inter-view prediction in MVC.

the second view could be achieved. In this way, one of the views is significantly blurred or more coarsely quantized than the other [16] or coded with a reduced spatial resolution [17, 18], yielding an imperceptible impact on the stereo quality. With mixed resolution coding, it has bene reported that the an additional view could be supported with minimal rate overhead, e.g., on the order of 25-30% additional rate for coding the right view at quarter resolution. Further study is needed to understand how this phenomenon extends to multiview video.

Additional considerations

MVC was designed mainly to support auto-stereoscopic displays that require a large number of views. However, large camera arrays are not common in the current acquisition and production environments. Furthermore, although MVC is more efficient than simulcast, the rate of MVC encoded video is still proportional to the number of views. Of course, this varies with factors such as scene complexity, resolution and camera arrangement, but when considering a high number of views, the achievable rate reduction might not be significant enough to overcome constraints on channel bandwidth.

Despite the challenges associated multiview video, MVC is still an effective format for the delivery of stereo contents. It has been shown that a good level of rate reduction could be achieved relative to simulcast and that backward compatibility with 2D systems could also be provided.

4.2 Block-based Coding Tools for Multiview Video

Several macroblock level coding tools have also being explored during the MVC standardization process. It has been shown that additional coding gains could be achieved beyond the inter-prediction coding supported by MVC. However, these tools were not adopted to the standard since they would require design changes at the macroblock level. It was believed that this implementation concern outweighed the coding gain benefits at the time. The benefits of block-level coding tools may be revisited in the specification of future 3D video formats; the main ideas of the proposed block-level coding tools are reviewed in the remainder of this section.

Illumination Compensation

Illumination compensation is a block-based coding tool for multiview video that compensates for the illumination differences between views [19, 20]. This tool has shown to be very useful when the illumination or color characteristics vary in different views. This is a likely case since even cameras from the same manufacturer could acquire video with very different color properties. While conventional color normalization procedures could be applied prior to encoding, not all applications and content creation settings would allow for this step.

The proposed method determines an offset value that corresponds to the difference in illumination between a current block and its reference. This offset value is calculated as part of the motion estimation process. Rather than compute the typical sum of absolute differences (SAD) between blocks of pixels in the current and reference frame, a mean-removed SAD is computed instead, where there is a mean associated with the current block and a mean associated with a reference block. The difference between these two mean values is the offset that is used for illumination compensation in the decoder. The decoder simply adds this offset value as part of the motion-compensated prediction.

The illumination differences between views have been found to be spatially correlated. Therefore, rather than coding the offset value directly, a prediction from neighboring illumination offset values is used keep rate overhead to a minimum. Coding gains up to 0.6 dB have been reported in comparison to the existing weighted prediction tool of H.264/AVC.

It was also observed by Lai, et al. [21, 22] that there are other types of mismatches present in multiview video. In particular, an adaptive reference filtering scheme was proposed to compensate for focus mismatches between different views.

Motion Skip Mode

An effective method to reduce bit rate in video coding is to infer side information used in the decoding process, e.g., motion vectors for a particular block, based on other available data, e.g., motion vectors from other blocks. This is the basic principle of direct mode prediction in AVC.

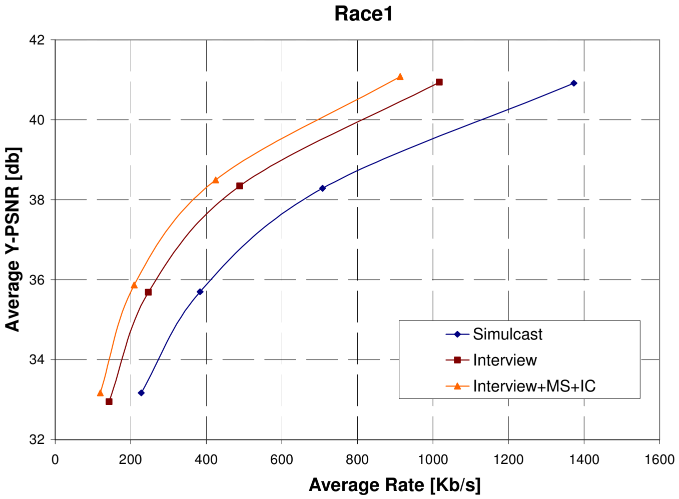


Fig. 9. Illustration of coding efficiency for Race1 sequence demonstrating gains using motion skip (MS) and illumination compensation (IC) coding tools.

In [23, 24], Koo, et al. proposed extensions to the conventional skip and direct coding modes for multiview video coding. Specifically, this method infers side information from inter-view references rather than temporal references. A global disparity vector is determined for each neighboring reference view. The motion vector of a corresponding block in the neighboring view may then be used for prediction of the current block in a different view. This signaling is very minimal and this method has the potential to offer notable reduction in bit rate.

An analysis of the coding gains offered by both illumination compensation and motion skip mode was reported in [13]. A rate reduction of 10% was reported over a wide range of sequences with a maximum reduction of approximately 18%. A sample plot of rate-distortion performance for the Race1 sequence is given in Figure 9. While the gains are considered sizeable, these tools would require block-level changes to the decoding process, which was viewed as undesirable at the time and led to the decision not to adopt such tools into the MVC standard.

View Synthesis Prediction

Another novel macroblock level coding tool that has been explored for improved coding of multiview video is view synthesis prediction. This coding technique predicts a picture in the current view from synthesized references generated from neighboring views.

One approach for view synthesis prediction is to encode depth for each block, which is then used at the decoder to generate the view synthesis data

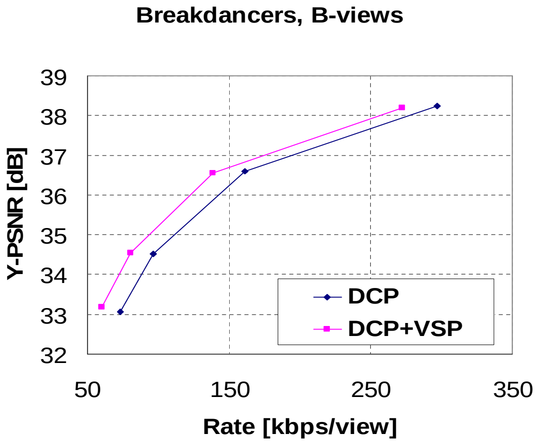


Fig. 10. Comparison of coding performance of disparity-compensated prediction and disparity-compensated prediction with view synthesis prediction for 100 frames of the Breakdancer sequence. Results are computed on B-views only.

used for prediction, as first described by Martinian, et al. [25] and fully elaborated on by Yea and Vetro [26]. Another approach estimates pixel-level disparities at both the encoder and decoder and encodes only disparity correction values [27].

As we know, conventional inter-view prediction maps every pixel in a rectangular block in the predicted view to the corresponding pixel in the reference view, where every pixel in the block is displaced by the same amount using a single disparity vector. In contrast, view synthesis prediction maps the matching pixels according to the scene depth a camera configuration; such a mapping could provide better prediction when the matching area in the reference view is non-rectangular or the correspondence between the views is non-translational. Further details on the view synthesis process, the estimation of depth and the coding of side information are available in the cited papers.

As a sample result to illustrate the coding gains, the results from [26] are discussed. Figure 10 shows a rate-distortion curve that compares the performance of disparity-compensated prediction (DCP) with and without view synthesis prediction (VSP). The results in this plot are averaged over all 100 frames of the B-views in the Breakdancer sequence, i.e., views 1, 3 and 5; these which are the views that utilize two spatially neighboring views from different directions in addition to temporal prediction. While the gains are not substantial at the higher bit rates, we do observe notable gains in the middle to low bit rate range that are between 0.2 and 1.5 dB.

4.3 Depth Compression Techniques

As discussed in previous sections, depth information could be used at the receiver to generate additional novel views or used at the encoder to realize more efficient compression with view synthesis prediction schemes. Regardless of the application, maintaining the fidelity of depth information is important since the quality of the view synthesis result is highly dependent on the accuracy of the geometric information provided by depth. Therefore, it is crucial to strike a good balance between the fidelity of depth data and the associated bandwidth requirement.

As reported in [28], the rate used to code depth video with pixel-level accuracy could be quite high and on the same order as that of the texture video. Experiments were performed to demonstrate how the video synthesis quality varies as a function of bit rate for both the texture and depth videos. It was found that higher bit rates were needed to code the depth data so that the view rendering quality around object boundaries could be maintained.

There have been various approaches considered in the literature to reduce the required rate for coding depth, while maintaining high view synthesis and multiview rendering quality.

One approach is to code a reduced resolution version of the depth using conventional compression techniques. This method could provide substantial rate reductions, but the filtering and reconstruction techniques need to be carefully designed to maximize quality. A set of experiments were performed in [8] using simple averaging and interpolation filters. These results demonstrate effective reduction in rate, but artifacts are introduced in the reconstructed images due to the simple filters. Improved down/up sampling filters were proposed by Oh, et al. in [29]. This work not only achieves very substantial reductions in the bit rate, but also improved rendering quality around the object boundaries.

Another approach is to code the depth based on geometric representation of the data. In [30], Morvan, et al. model depth images using a piece-wise linear function; they referred to this representation as platelets. The image is subdivided using a quadtree decomposition and an appropriate modeling function is selected for each region of the image in order to optimize the overall rate-distortion cost. The benefit of this approach for improved rendering quality relative to JPEG 2000 was clearly shown. In subsequent work, comparisons to AVC intra coding were also made and similar benefits have been shown [31].

A drawback of the platelet-based approach is that it appears difficult to extend this scheme to video. An alternative multi-layered coding scheme for depth was suggested in [32]. In this approach, it was argued that the quality of depth information around object boundaries needs to be maintained with higher fidelity since it as notable impact on subjective visual quality. The proposed scheme guarantees a near-lossless bound on the depth values around the edges by adding an extra enhancement layer. This method effectively improve the visual quality of the synthesized images, and is flexible in the sense that it could incorporate any lossy coder as the base layer thereby making it easily extendible to coding of depth video.

5 Discussion

After many false starts, it now appears that 3D video is forming a positive impression on consumers around the world through high quality digital theater experiences. 3D content production is ramping up and the infrastructure and equipment is being put in place to delivery 3D video to the home. There is still a significant amount of work to be done in the various standards organizations to ensure interoperable 3D services across a wide range of application domains and equipment. One critical issue will be determining suitable 3D data formats among the various options that have been described in this chapter.

Current industry activities are now focused on distribution of stereoscopic video since this is what current display technology, production and delivery infrastructure could practically support. However, there is significant research and standardization initiatives underway that target 3D solutions that would not require glasses. To enable this, a very rich and accurate representation of the scene is needed. The data needs to be coded in an efficient manner and be rendered on devices that are capable of providing an immersive and realistic 3D viewing experience. It is equally important for this to be done in a practical way based on current state of technologies so that the experience could be made available to consumers on a large scale.

References

1. ITU-T Rec. & ISO/IEC 14496-10 AVC: Advanced video coding for generic audiovisual services. (2009)
2. SMPTE: Report of SMPTE Task Force on 3D to the Home. (2009)
3. Broberg, D.K.: Considerations for stereoscopic 3D video delivery on cable. In: IEEE International Conference on Consumer Electronics, Las Vegas, NV (2010)
4. Konrad, J., Halle, M.: 3D displays and signal processing: An answer to 3D ills? IEEE Signal Processing Magazine **24**(6) (2007) 97–111
5. HDMI Licensing, LLC.: HDMI Specification 1.4. (2009)
6. Markwalter, B., Stockfisch, M.: Enabling 3D content in consumer electronics. In: IEEE International Conference on Consumer Electronics, Las Vegas, NV (2010)
7. Video Group: Text of ISO/IEC 14496-10:2009/FDAM1: Constrained baseline profile, stereo high profile and frame packing arrangement SEI message. In: ISO/IEC JTC1/SC29/WG11 Doc. N10707, London, UK (2009)
8. Vetro, A., Yea, S., Smolic, A.: Towards a 3D video format for auto-stereoscopic displays. In: SPIE Conference on Applications of Digital Image Processing XXXI. (2008)

9. Video and Requirements Group: Vision on 3D Video. In: ISO/IEC JTC1/SC29/WG11 Doc. N10357, Lausanne, Switzerland (2009)
10. Video and Requirements Group: Applications and Requirements on 3D Video Coding. In: ISO/IEC JTC1/SC29/WG11 Doc. N10857, London, UK (2009)
11. Merkle, P., Müller, K., Smolic, A., Wiegand, T.: Efficient compression of multiview video exploiting inter-view dependencies based on H.264/AVC. In: IEEE International Conference on Multimedia & Expo, Toronto, Canada (2006)
12. Merkle, P., Smolic, A., Müller, K., Wiegand, T.: Efficient prediction structures for multiview video coding. IEEE Transactions on Circuits and Systems for Video Technology **17**(11) (2007) 1461–1473
13. Tian, D., Pandit, P., Yin, P., Gomila, C.: Study of MVC coding tools. In: ITU-T & ISO/IEC JTC1/SC29/WG11 Doc. JVT-Y044, Shenzhen, China (2007)
14. Chen, T., Kashiwagi, Y., Lim, C.S., Nishi, T.: Coding performance of Stereo High Profile for movie sequences. In: ITU-T & ISO/IEC JTC1/SC29/WG11 Doc. JVT-AE022, London, UK (2009)
15. Droese, M., Clemens, C.: Results of CE1-D on multiview video coding. In: ISO/IEC JTC1/SC29/WG11 Doc. m13247, Montreux, Switzerland (2006)
16. Stelmach, L., Tam, W., Meegan, D., Vincent, A.: Stereo image quality: effects of mixed spatio-temporal resolution. IEEE Transactions on Circuits and Systems for Video Technology **10**(2) (2000) 188–193
17. Fehn, C., Kauff, P., Cho, S., Kwon, H., Hur, N., Kim, J.: Asymmetric coding of stereoscopic video for transmission over T-DMB. In: 3DTV-CON 2007, Kos, Greece (2007)
18. Brust, H., Smolic, A., Müller, K., Tech, G., Wiegand, T.: Mixed resolution coding of stereoscopic video for mobile devices. In: 3DTV-CON 2009, Potsdam, Germany (2009)
19. Lee, Y.L., Hur, J.H., Lee, Y.K., Han, K.H., Cho, S.H., Hur, N.H., Kim, J.W., Kim, J.H., Lai, P.L., Ortega, A., Su, Y., Yin, P., Gomila, C.: CE11: Illumination compensation. In: ITU-T & ISO/IEC JTC1/SC29/WG11 Doc. JVT-U052, Hangzhou, China (2006)
20. Hur, J.H., Cho, S., Lee, Y.L.: Adaptive local illumination change compensation method for H.264/AVC-based multiview video coding. IEEE Transactions on Circuits and Systems for Video Technology **17**(11) (2007) 1496–1505
21. Lai, P., Ortega, A., Pandit, P., Yin, P., Gomila, C.: Adaptive reference filtering for MVC. In: ITU-T & ISO/IEC JTC1/SC29/WG11 Doc. JVT-W065, San Jose, CA (2007)
22. Lai, P., Ortega, A., Pandit, P., Yin, P., Gomila, C.: Focus mismatches in multiview systems and efficient adaptive reference filtering for multiview video coding. In: SPIE Conference on Visual Communications and Image Processing, San Jose, CA (2008)
23. Koo, H.S., Jeon, Y.J., Jeon, B.M.: MVC motion skip mode. In: ITU-T & ISO/IEC JTC1/SC29/WG11 Doc. JVT-W081, San Jose, CA (2007)
24. Koo, H.S., Jeon, Y.J., Jeon, B.M.: Motion information inferring scheme for multi-view video coding. IEICE Transactions on Communications **E91-B**(4) (2008) 1247–1250
25. Martinian, E., Behrens, A., Xin, J., Vetro, A.: View synthesis for multiview video compression. In: Picture Coding Symposium, Beijing, China (2006)
26. Yea, S., Vetro, A.: View synthesis prediction for multiview video coding. Image Communication **24**(1-2) (2009) 89–100

27. Kitahara, M., Kimata, H., Shimizu, S., Kamikura, K., Yashima, Y., Yamamoto, K., Yendo, T., Fujii, T., Tanimoto, M.: Multi-view video coding using view interpolation and reference picture selection. In: IEEE International Conference on Multimedia & Expo, Toronto, Canada (2006) 97–100
28. Merkle, P., Müller, K., Smolic, A., Wiegand, T.: Experiments on coding of multi-view video plus depth. In: ITU-T & ISO/IEC JTC1/SC29/WG11 Doc. JVT-X064, Geneva, Switzerland (2007)
29. Oh, K., Yea, S., Vetro, A., Ho, Y.: Depth reconstruction filter and down/up sampling for depth coding in 3D video. **16**(9) (2009) 747–750
30. Morvan, Y., Farin, D., de With, P.H.N.: Depth-image compression based on an R-D optimized quadtree decomposition for the transmission of multiview images. In: IEEE International Conference on Image Processing, San Antonio, TX (2007)
31. Merkle, P., Morvan, Y., Smolic, A., Farin, D., Mller, K., de With, P., Wiegand, T.: The effects of multiview depth video compression on multiview rendering. Image Communication **24**(1-2) (2009) 73–88
32. Yea, S., Vetro, A.: Multi-layered coding of depth for virtual view synthesis. In: Picture Coding Symposium, Chicago, IL (2009)