Trend and challenges in 3D digital television: video formats and transmission standards

DRAGORAD MILOVANOVIC, ZORAN BOJKOVIC

University of Belgrade,
Studentski Trg 1, 11000 Belgrade, Republic of Serbia

e-mail: dragoam@gmail.com, http://dragorad.milovanovic.googlepages.com

Abstract: 3D video is new media extension of conventional 2D video into third dimension adding depth sensation and resolving 2D viewing ambiguity. Stereoscopic and free-viewpoint videos are visual representation and coding formats which takes 3D geometry information of acquisition system. Different 3D video acquisition, processing and representation technologies can be used, but all make use of multiple views of the same 3D scene. Advances in multi-camera arrays and display technology enable new applications for 3D television. 3DTV requires transmission of jointly encoded multiple synchronized video signals that show the same 3D scene from different viewpoints. It’s clear that 3DTV, like all other forms of TV before it, needs to be based on well defined and documented technical standards. Technical trend and challenges in 3D video are summarized in this paper with special focus on 3DTV standardization activities in MPEG and DVB.

Key-words: 3D video, multiview video, stereoscopic display, simulcast, StereoSEI, MVC joint coding

1 Introduction

Digital media formats integrate number of media types (video, audio, computer graphics, text, images) into a single file or transport format. The trend is convergence of digital content, terminals and applications [1, 2]. Recently, the convergence of new technologies from computer graphics, computer vision, image processing, multimedia communication enabled also the development of new types of media. An important driving factor in 3D media is the availability of international standards [3, 4].

The next MPEG standardization processes are 3DTV and new types of media:

- **SSV (stereoscopic video)** means a depth impression of the observed 3D scene.
- **FVV (free-viewpoint video)** allows for an interactive selection of arbitrary viewpoint and viewing direction within a certain operating range.

In order to enable using SSV and FVV in 3DTV systems, the complete processing chain (image acquisition, 3D representation, compression, transmission, signal processing, interactive rendering, 3D displays) need to be considered [5-8]. The system optimization takes all parts into account, since there are strong interrelations between all of them. These are quite diverse and various types of 3D scene representations can be employed, which implies a number of different media formats [9, 10].

In the first part of the paper, 3DTV broadcast services, free-view point systems and 3D display technologies are presented. Next, the current status in 3D media standardization (MPEG-C Part 3, AVC StereoSEI, MVC, 3DAV, 3DVC) and 3DTV formats (CSV, MVV, MVD, LDV, DES) are summarized.

2 3DTV systems

From a technical viewpoint, 3DTV systems for the home are already feasible. However, systems mostly involve a redesign of the entire processing chain (new cameras, coding, transmission and display).

- For economical reasons **compatibility** with conventional 2D devices is required. This either involves standard 2D transmission and depth generation at the user side, or the development of a novel 3D video format.
- As 3D display development is a rapidly evolving field, it is required that a wide range of current and future display devices can be used. A **proper choice** of 3D video format is needed, to avoid substantial video processing for format conversions in future displays.
- The need for high-quality **3D content** favors the use of originally recorded content. On the other hand, the viewer preferably should be able to adjust the depth impression to a pleasing amount, which requires **local personalized** image rendering to accommodate for personal preferences.

3DTV broadcast and packetized free-viewpoint video on DVD storage media, require interoperable standard-based systems [11, 12]. This open up large consumer markets for 3D displays, set-top boxes, media, content, and storage media devices, along with the corresponding equipment for production, and transmission [9, 10].
2.1 3DTV broadcast services

The basic psychophysical science of stereo imaging is well established. However, there are many technical implementation gaps in 3DTV development. The key technical issues that have prevented optimized 3DTV systems are now resolved based on MPEG and DVB digital video standards, providing users with more viewing choices than ever before.

- **Transition** from analog to digital television creates both a framework and a necessity for new home media interfaces (HDMI is already 3D capable and has a schedule of 3D improvements for the near future). Moreover, almost all TV programmers and service operators have built DVB/MPEG infrastructures to support digital television and HDTV.
- **Flat panel TVs** are one of the critical enablers of 3DTV. The newest versions of LCDs and plasmas have the high resolution, the high refresh rates, and the support for high frame rate inputs that are needed for good quality 3D experiences. Just as important, flat panel TVs can be improved with optical capabilities (passive and active polarization, or frame-sequential programming) that selectively display left- and right-eye stereo views so efficiently that viewers see 3D smoothly.

An end-to-end 3DTV system based on existing 2D DVB digital TV infrastructure, consists of 3D video representation and compression, transport protocols and systems, and 3D display user is shown in Fig. 1. The main difficulty in the deployment of 3DTV broadcast services is the large bandwidth requirement associated with transport and display of **multiple video streams** (Fig. 2, 3). Research and development of efficient solutions to transport multi-view video signals over broadcast channels, which are compatible with the existing infrastructure and international standards are a challenging task.

2.2 Free-viewpoint systems

Free-viewpoint is the most general and challenging scenario of 3DTV. Many systems have been proposed to realize FVV function. It is common to all systems that a scene is captured with a set of cameras (Fig. 4).

- If the camera density is very high, view generation is simple performed by selecting a camera image or by collecting pixels from camera images. This is referred to as integral photography. If the camera density is not very high, view generation needs some processing.
- If it is moderately high, intermediate views can be generated precisely by interpolation of camera images using camera parameters. This is referred to as ray-space. Although light field and lumigraph belong to this category, precise interpolation was not reported in these methods. If intermediate views are generated, not by precise interpolation, but by warping or projection of camera images, generated views are not precise. This is referred to as image domain.
- If the camera density is low, intermediate views can be generated by detecting objects in the scene. This is referred to as model-based. Generated views are not precise in this method. Detection of objects in natural scenes has been studied intensively in the field of computer vision. However, it is a very difficult problem and has not yet been solved.

If the method of free-view image generation has strong scene dependence, its application is very limited. **Transparency** of the method is low in this case. If the method has no scene dependency, the transparency of the method is high. A typical example of the latter case is TV, which is transparent because it works well for any scene and content.

2.3 3D displays

The human viewer is the final judge of the quality of any 3DTV system and its widespread acceptance and commercial success. Understanding the physiological, perceptual, cognitive, and emotional processes underlying the viewer’s experience and judgment is essential to informing the development of technology that aims to be usable in its interaction, pleasurable to watch, and fun to own.

Some of the human factors limiting the development of a successful stereoscopic 3DTV broadcast service:

- some will be related to the way stereoscopic images are produced, such as keystoning due to a converging camera setup or left-right camera misalignment (e.g., rotational error, vertical offsets, luminance differences).
- transmission-related errors (compression artifacts),
- unwanted side-effect of the specific display technology used to address the left and right eyes separately (cross-talk, loss of resolution).

There are two basic approaches in the underlying 3D display technologies.

- **Aided-viewing** systems relies on special user-worn devices, such as stereo glasses or head-mounted miniature displays, in order to optically channel the left/right-eye views to the appropriate eye (Fig. 3a).
- Another approach integrates the optical elements needed for selective addressing of the two eyes in a remote display device, hence allowing free 3D viewing with the naked eye. In general, **free-viewing** 3D displays are more comfortable to viewers (Fig. 3b,c).

The technologies being pursued for 3D display can be broadly divided into the following categories:
holographic, volumetric, autostereoscopic and stereoscopic displays. For 3D TV applications, it is only interested auto-stereoscopic displays. Most of these displays are based on parallax barrier or lenticular technology, and most of them provide multiple stereoscopic images from a wide range of viewing positions.

Today’s commercial parallax barrier displays use a plate with multiple alternating image stripes per vertical slits as a barrier over an image with alternating strips of left-eye/right-eye images and place parallax barriers on top of LCD or plasma screens (Fig.3b). Parallax barriers generally reduce some of the brightness and sharpness of the image. Some implementations use an LCD screen to display the parallax barriers on top of the viewing screen, which has the advantage that the display can be switched to 2D viewing without any loss in brightness.

Lenticular sheet is a linear array of narrow cylindrical lenses. Each lens (or lenticule) acts as a light multiplexer, projecting a subset of vertical display pixels towards each eye (Fig.3c). The two key parameters of lenticular sheets are the field-of-view (FOV) and the number of lenticules per inch (LPI). Lenticular images found widespread use for advertising, CD covers, and postcards, which has lead to improved manufacturing processes and the availability of large, high-quality, and very inexpensive lenticular sheets. Some modern lenticular displays place diagonally-arranged lenticules on top of high-resolution LCD or plasma monitors.

3 3D media standardization

It’s clear that 3DTV, like all other forms of TV before it, needs to be based on well defined and documented technical standards. In many ways, the MPEG committee has been anticipating the widespread deployment of 3DTV for some time.

• As early as 1996, 3DTV capabilities were included in the MPEG-2 MultiView Profile standard.

• Over the past several years, MPEG has worked to amend MPEG-4 AVC/H.264 to include Multiview Coding (MVC), which is a way of improving compression between multiple views (camera angles) of the same content. Last year, Stereohigh Profile was added to deal specifically with the case in which the multiple views of MVC were the left- and right-eye stereo views.

• Also last year, Frame Packing Arrangement SEI (Supplemental Enhancement Information) messages were added to MPEG-4 AVC. SEI messages inform decoders about any special attributes of the compressed video. The Stereo SEI message tells the decoder that the left- and right-eye stereo views are packed into a single high-resolution video frame either in a top-to-bottom, side-by-side, checkerboard, or other arrangement (Fig.2a). Packing both left- and right-eye stereo views into a single video frame makes it possible to use existing encoders and decoders to distribute 3DTV immediately without having to wait for MVC and Stereo HighProfile hardware to be deployed widely.

• ISO/IEC 23002-3 MPEG-C Part 3 specifies the representation of auxiliary 2D+depth video and supplemental information. In particular, it enables signaling for depth map streams to support 3D video applications.

MPEG-4 BIFS and SNHC extension AFX (Animation Framework eXtension) integrated three 3D video tools: depth image-based rendering (DIBR), point rendering, and view-dependent multitexturing [11].

• The MPEG-4 AFX tool DIBR implements the concept of layered depth images. In this case a 3D object or scene is represented by a number of views with associated depth maps as shown in Fig.2b. The depth maps define a depth value for every single pixel of the 2D images. Together with appropriate scaling and information from camera calibration it is possible to render virtual intermediate views. The quality of the rendered views and the possible range of navigation depend on the number of original views and the setting of the cameras. A special case of this method is stereovision, where two views are generated in accordance with the geometry of the human eyes basis. In this case the depth is often calculated using disparity estimation. Supposing that the capturing cameras are fully calibrated and their 3D geometry is known therefore, corresponding depth values can be recalculated one-to-one from the estimated disparity results. In the case of simple camera configurations (such as a conventional stereo rig or a multi-baseline video system) this disparity estimation can even be used for fully automatic real-time depth reconstruction in 3DTV applications.

• In a point-based representation shape, colour and other properties of 3D objects are defined by a point cloud in 3D. This can be regarded as an alternative to classical 3D mesh models with associated textures. Points are samples of surfaces in 3D. In contrast to 3D meshes such a point cloud representation does not require additional information such as connectivity, texture maps, etc., (note that color and other attributes are assigned to each point individually), which is an advantage over 3D mesh representations. Each point represents a small surface element or surfel, describing the surface in a small neighborhood. A point-based representation as contained in MPEG-4 AFX allows...
for high quality rendering of 3D objects at reasonable computational costs. A point cloud can be regarded as a natural extension of 2D image pixels as they come from cameras into 3D. Since there is no need for connectivity the 3D reconstruction process is also simplified.

- View-dependent multi-texturing is attractive for classical computer graphics applications (games, movie production), but also for reconstructed real world 3D video objects.

**MPEG 3DAV (3D AudioVisual)** is the most recent trend related to standardization for 3D video communications. This activity investigates possible technology for standardization and supports in particular interactivity. Interactivity in the sense of 3DAV is the ability to look around within an audiovisual scene by freely choosing viewpoints and viewing directions. Applications include omnidirectional video, interactive stereo video and free-viewpoint video [11].

- **Omnidirectional (panoramic)** video can be regarded as an extension of the planar 2D image plane to a spherical or cylindrical image plane. Omnidirectional video can be displayed in a suitable player, with the key types of interaction being zoom and rotation to give the effect of looking around. However, in contrast to free-viewpoint video, the user is not able to change the position of the viewpoint interactively. The viewpoint might change, but that implies that the camera has been moved during the period of capturing. With the scene projected onto a dome, or with a head-mounted display, the user can get the impression of being part of the scene.

- **Free-viewpoint** video is the most general and challenging application scenario investigated in 3DAV. Within the context of MPEG 3DAV free-viewpoint video has been divided into model-based and ray-space approaches.

**Ray-space methods**

The Ray-Space solution is based in signal processing, not computer graphics. The Ray-Space can generate free-viewpoint image without 3D models and decomposition process. Therefore, it is suitable for full real-time applications from capturing to display. The scene is captured densely by a high number of cameras (usually 100 cameras) and it is transformed to the Ray-Space domain. The new views are generated selecting the proper data in the Ray-Space and inverting the transform. It is not model based and independent from the geometry, so the image quality is really high and independent from the scene complexity.

**3D model reconstruction**

Model reconstruction methods use a relatively low number of cameras (from 2 to 30). The moving foreground is segmented in real-time by background subtraction and the reconstruction of 3D information is mostly done using shape-from-silhouette methods (hierarchical voxel modeling, marching-cubes, surface smoothing, mesh reduction) which calculate an approximation of the object (the visual hull). There are several possible representations for visual hulls: image based representations (reconstruct a depth image in the desired view and shade it directly via video textures), polyhedral representations (reconstruct a mesh representation of the scene by intersecting the visual cones of the object, which is rendered by viewpoint dependent texture mapping), point based representations (reconstruct an irregular point cloud which gets rendered by forward projection and image reconstruction), and volumetric representations (reconstruct an occupancy voxel set which can be rendered by determining voxel visibility and rendering each voxel using textures).

4 **3DTV formats and coding**

In this section, a brief overview of existing 3D video formats, including both stereo and multiview formats, is provided. The merits of each format will be discussed along with the drawbacks and limitations.

- **CSV (conventional stereo video)** consists of a pair of sequences, showing the same scene for the right and the left eye view. It doubles the amount of data to be stored or transmitted. A derivative of CSV is the MRS (mixed resolution format), which consists of a pair of sequences, showing the same scene for the right and the left eye view. In contrast to CSV one of the two sequences is sub-sampled. Instead of transmitting a 2 view color video (CSV) a single view and an associated depth map can be used. The second view can be rendered using DIBR (depth-image based rendering).

- **MVV (multiview video)** is more general representation, where a number of input cameras capture a scene from different viewpoints. A specific case under investigation is to use linear N view camera settings, which play a special role for 3D Video applications. Having N views form slightly different viewpoints allows for a 3D impression within a limited range by presenting 2 adjacent of the N views as a stereo pair to the user.

- **MVD (multiview video plus depth)** is much more advanced format. It is based on the assumption, that the number of views for transmission can be reduced, if additional information about the recorded scene is available. For MVD, this information is per pixel depth maps, which are provided for each remaining view. **LDV (layered depth video)** is a derivative and alternative to MVD. It uses one color video with associated depth map and a background layer with associated depth map. The background layer includes image content which is covered by foreground objects in the main layer.
A strong requirement for 3D video formats is the compatibility to existing stereo displays, which are mainly based on conventional stereo up to know. To combine the good quality of simple stereo video with the extended functionality of depth data availability, the DES (Depth-enhanced Stereo format) was created. It contains two views as a stereo pair, together with depth data and possibly occlusion color and depth data for both views.

Currently, a number of different coding methods exist, which are part of international standards (Fig.4c). All methods derive from 2D video coding and contain certain structures and coding features, which can be used to code the 3D video representation formats, described above.

- **In H.264/AVC simulcast** one coder is applied to N video sequences in a generic way resulting in N encoded bit- or transport-streams. In simulcast, each view is encoded independent of the other. This solution has low complexity since dependencies between views are not exploited, thereby keeping computation and processing delay to a minimum. It is also a backward compatible solution since one of the views could be decoded for legacy 2D displays. With simulcast, each view is assumed to be encoded with full spatial resolution, and the main drawback is that coding efficiency is not maximized since redundancy between views is not considered.

- **MPEG-4 AVC Stereo SEI** interlaced and coded a CSV sequence is in field coding mode. Dedicated fields tell the decoder that the output has to be deinterlaced into two individual view sequences.

- **MVC (Multiview Video Coding)** is an extension of the Advanced Video Coding (AVC) standard that provides efficient coding of MVV multiview video. The encoder receives N temporally synchronized video streams and generates one bit stream. The decoder receives the bit stream, decodes and outputs the N video signals.

- **MPEG-C Part 3** (ISO/IEC 23002-3 Auxiliary Video Data Representations) was specifically developed as container format allowing to transmit MVD (multiview video plus depth) in separate coding streams. Includes meta data for image-based rendering of virtual views from V+D. The video and the depth sequence are encoded independently, resulting in two bit-streams. ISO/IEC 23002-3 consists of an array of N-bit values which are associated with the individual pixels of a regular video stream. These data can be compressed like conventional luminance signals using already existing (and even future) MPEG video codecs. The format allows for optional subsampling of the auxiliary data in both the spatial and temporal domain. This can be beneficial depending on the particular application and its requirements and allowing for very low bitrates for the auxiliary data. The specification is very flexible in the sense that it defines a new 8-bit code word aux_video_type that specifies the type of the associated data, e.g., currently a value of 0x10 signals a depth map, a value of 0x11 signals a parallax map. New values for additional data representations can be easily added to fulfill future demands.

- **3DVC (3D Video Coding)** is a standard that targets serving a variety of 3D displays (Fig.4c). Such displays here in focus present N views (e.g. N=9) simultaneously to the user, so-called multi-viewed displays. For efficiency reasons only a lower number K of views (K=1,...,3) shall be transmitted. For those K views additional depth data shall be provided. At the receiver side the N views to be displayed are generated from the K transmitted views with depth by depth image based DIBR rendering. This application scenario imposes specific constraints such as narrow angle acquisition (~K out of N views, augmented with K depth sequences). This representation related to stereo-video generalizes the possibilities of MPEG-C, Part 3 and MVC, i.e. the two separate views can be coded together or can be reduced to single view + depth with the second view to be synthesized at the receiver. 3DVC is an ongoing MPEG activity, and a standard is expected in 2011.

There are various approaches to encoding multiview video, which provide trade-off between random access, ease of rate adaptation, and compression efficiency. 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. The Multiview Video Coding (MVC) standard has been specified as Amendment H of H.264/AVC. It has been shown that MVC gives significantly better results compared to simple AVC-based simulcast.

A key aspect of the MVC design is that it does not require any changes to lower-level syntax, so it is quite compatible with single-layer AVC hardware. Also, it is mandatory for the compressed multiview stream to include a base layer stream that could be easily extracted and used for backward compatibility with legacy 2D displays; this base layer stream is identified by the NAL unit type in H.264/AVC. Furthermore, inter-view prediction (Fig.5) is enabled through flexible reference picture management, where decoded pictures from other views are made available in the decoded reference picture buffer. It is important to emphasize that the core decoding modules do not need to be aware of whether reference picture is a time reference or multiview reference from another view. In terms of syntax, the standard only requires small changes to high-level syntax, e.g., view dependency needs to be known for decoding [12].
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.

It is important to note that in the near-term MVC is still useful for delivery of stereo contents. While some rate reduction could certainly be achieved relative to simulcast, the more important factor might be the backward compatibility that it provides to existing 2D services. In contrast to some of the stereo interleaving solutions, the full resolution could also be maintained. MVC also has benefits relative to the 2D+Depth format for stereo in terms of rendering quality for generic scenes.

5 Concluding remarks
Television maintains position as the most important and useful multimedia communication system. However, digital TV enables only a limited view of a real 3D scene. The view is determined not by the user, but by a camera setup in the 3D scene. Although many important technologies have been developed, this function of TV has never changed!? 3DTV are developing a new function such that users view a 3D scene freely from any viewpoint, making an important change in the history of television.

Development of 3D television is challenging task. The special requirements are many different 3D display technologies and varying viewing conditions (distance, display size). User preferences have to be considered (~10% users can see stereo, also users might not want to see all programmes in stereo).

The current status is maturing of 3D technology due to world wide development of whole processing chain from capturing to display. The first 3D video systems are in use, mainly using stereoscopic displays and MVC coding. Restrictions are fixed view number and spatial positions, as well as glasses are needed for stereo display. International consortia (DVB Study mission on 3DTV, DVD Forum, ITU-R SG6, SMPTE WG 3D Home master) works towards next generation 3D, e.g. for home entertainment and mobile applications. New 3D video standardization activities (MPEG 3DVC, 3D Media cluster) will provide generic format for any multi-view display and better 3DTV use acceptance.

Challenges in MVC joint coding based on MPEG4-AVC/H.264 temporal and inter-view dependencies in multiple camera views are reduced computational complexity, fast joint motion and disparity estimation by using coherence condition between motion and disparity in JMVM codec, as well as new prediction structure with better coding efficiency and random access capability.

MPEG develops new 3DVC coding standard including MVD, LDV and DES video format. The main limitations of conventional stereo video format are 2 fixed views of stereo video (head motion parallax viewing is not supported), viewer expects occlusions/disocclusions when watching 3D scene, unnatural impression when moving head (impression that whole scene moves). Improved stereo video + depth format is also limited with disocclusion artifacts increase with distance of virtual view from available original view, does not support wide range multi-view 3D, displays very limited free viewpoint navigation, as well as problem of depth estimation.

Further motivation for 3DVC development are decouple production from coding format, as well as AVC only optimized for 2D color video, but not for depth information. The main 3DVC constraints are considering capturing technology (2-3 recorded views), break linear dependency of coding bitrate from number of target views, providing scene geometry data in general format (pixel-wise depth data), consider statistical properties of depth (and supplementary) data, considering new quality evaluation methods for intermediate views, providing high-quality view synthesis for continuous viewing range, as well as decrease depth and coding errors.

References:
Special issue on 3DTV, capture, transmission and display of 3D video, EURASIP Journal on Advances in signal processing, April 2008.


Fig. 1. 3DTV MPEG/DVB standard-based broadcast system: video coding, transmission and display.

Fig. 2. 3D video formats: a) stereo multiplexed format (line-by-line, side-by-side-half, above-below, checkerboard), b) 2D+z format (2D image + depth map (logarithmical by depth-per-pixel quantization between $z_{near}$ and $z_{far}$ clipping plane)).

Fig. 3. a) Plano-stereoscopic device (active shutter glasses, passive polarized glasses), b) auto-stereoscopic parallax barrier, c) auto-stereoscopic lenticular linear array of narrow cylindrical lenses.
Fig. 4. a) Multi-view video ($N=16$ camera setup), b) free-viewpoint video ($N=5$ camera setup), c) 3D rendering capability versus bit rate for different coding formats [11].

Fig. 5. Typical MPEG JMVM MVC hierarchical temporal/interview prediction structure: a) $N=2$ stereo camera setup, b) $N=5$ simulcast (only temporal prediction), c) $N=5$ star camera setup, d) $N=15$ 2D array camera setup.