THE CHALLENGES OF 3D SUPPORT IN THE STB

Kevin Murray, Simon Parnall, Ray Taylor
NDS

Abstract

Distribution of stereoscopic 3DTV has been demonstrated using existing, deployed HD technology providing excellent quality pictures. This shows how easily 3DTV can be made available through broadcast channels as consumer displays reach the market and 3D content becomes readily available. However, a full broadcast service consists of more than just the video—the STB and the features it provides are key parts of the user experience.

This paper looks at several areas of key functionality that the STB provides. It discusses the changes required both in the STB software and the transmissions on which it relies, all whilst utilizing existing HD hardware. Through these discussions we explore how STBs can be updated or extended to support a seamless, high quality 3D aware service.

INTRODUCTION

The set top box (STB) is one part of the distribution path from the broadcaster to the display in the home. It is perfectly possible to utilize this path for stereoscopic 3DTV (S3DTV) without making any alterations to the STB software, by utilizing frame compatible broadcasts – formats that fit a stereoscopic pair into a normal frame. Indeed, broadcasts have been made that do just that. Initially we touch on certain aspects of broadcast formats, and how to minimize stereoscopic specific transmission problems.

In the frame compatible scenario, if the STB is not aware of the format or nature of the video it is carrying, then several functions of the STB will result in very unpleasant visual effects. The main discussion in this paper looks at two key areas of this functionality, and discusses the steps that are needed to make the STB software 3D aware, and the additional signaling or metadata that prevents, or at least minimizes, such unpleasant effects.

The first area we discuss is that of manipulating S3DTV video. Examples of manipulations include supporting such features as picture-in-picture and picture-in-guide. These operations are more complex in S3DTV than in 2D, and in some cases the use of such functionality may be less sensible or visually acceptable. The complexity can vary with format and operation, so we will discuss the key popular formats: side-by-side and top-and-bottom (also known as above-below).

The second key area where updates are needed is in the handling of closed captions. However, the required changes extend to almost all graphical overlays performed by the STB. These updates cover not just the format used to draw the graphics, but also issues in design to reduce eye-strain and in the importance of correct depth placement. Depth placement, and potentially accurate matching of graphics to video, rely on correct signaling and metadata provision in the broadcast. We identify a small set of extensible signaling that assists these key areas.

Although most of this paper is concerned with systems operating with frame-compatible video formats, many of the points apply to non-frame compatible distribution mechanisms and the support required in the STB for such formats. In several cases, the solutions can be identical for frame compatible and non-frame compatible modes. Likewise, many of the examples in this paper

are based on side-by-side formats, but the ideas extend to top-and-bottom formats as well.

STEREOSCOPIC VIDEO FORMATS

There are several formats for stereoscopic video in a frame compatible mode, some of the more common ones (that are also part of the mandatory HDMI 3D formats) are shown below in figures 1 and 2. In the simplest mode where the STB is not aware of the presence of S3DTV, the STB just decodes the received video from this format and outputs it in the same format over the HDMI connection to the display.

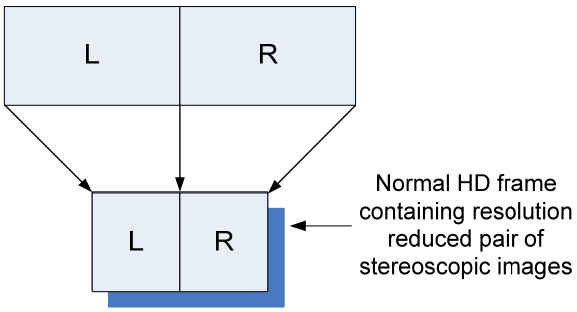

Figure 1: Side-by-Side Format

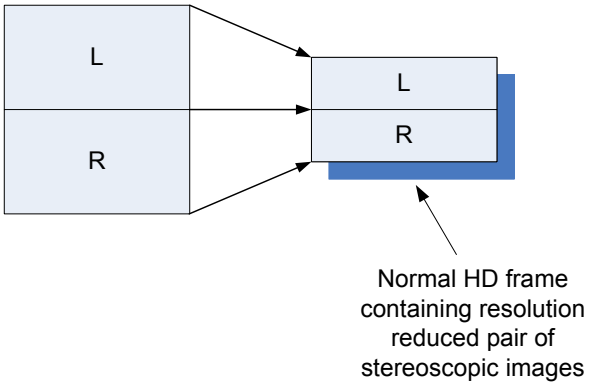

Figure 2: Top-and-Bottom Format

There are two main display technologies available for S3DTV—shutter based and passively polarized—and there is no requirement for a specific format for a specific display technology. These displays use HDMI signaling[1] to identify the input format. The display will then perform the necessary conversion to enable it to display the stereoscopic images. Thus a broadcast in one of the mandatory formats can be supported regardless of the display technology. In the absence of STB supported signaling over HDMI the viewer is required to use a remote control to set the display to the correct input format.

Side-by-side and top-and-bottom represent different trade-offs for resolution reduction. It is worth noting that these different formats result in different effective resolutions on differing display technologies.

A Note on Bitrates

When compressing stereoscopic video it is very important to operate the compression at a level above that where artifacts can occur. If artifacts do occur, there is no guarantee that the left and right eye images will match, and the viewer will experience discomfort.

Format Translations

Just as the displays are able to perform format conversions, so can the STB. Whilst most devices are easily able to scale and resize 2D video, conversion between side-by-side and top-and-bottom formats is probably not within the capability of most deployed devices. In comparison some other conversions, such as from top-and-bottom to line interleaved, are simple to perform.

Further, format conversions can result in a loss of resolution. For instance, conversions between side-by-side and top-and-bottom will normally result in an image where the L and R images are effectively one quarter the resolution of a full HD image. It is better to avoid conversions, or ensure that it is performed with awareness of the native requirements of the display. The chosen broadcast format should also reflect these limitations.

As a simple example, a line-interleaved passively polarized display will normally have a resolution equivalent to that of the top-and-bottom format. Thus if an image is received in top-and-bottom format, converting it unnecessarily to side-by-side would reduce the effective resolution and therefore the quality seen by the viewer.

When considering future STB devices and chipsets, it will be important that they are able to convert whatever formats they receive into the most appropriate output format for the their display. For example, a new device that can support 1080p60 per eye must be able to output at least one mandatory format[1]. Ideally, it should be able to select from more than one to maintain the best quality for the display. This is especially important as the early 3D displays will represent the legacy formats of the future, and it should be possible to drive them with the best quality signal they can accept.

Synchronization

The frame compatible formats have the advantage that they place both the left and right in the same image. This guarantees that the left and right eye images are not swapped in the delivery and display process, and ensures that the left and right images are always perfectly time synchronized.

Approaches for transmitting S3DTV that are not frame compatible can involve twin logical streams that may, or may not, be represented as separate flows within the appropriate transport (e.g. different PIDs within an MPEG-2 transport stream). Whilst both time synchronization and left-right synchronization can clearly be preserved through such twin stream systems, they do represent a point at which errors can occur. The failure of either (or both) synchronization(s) renders the content effectively un-viewable. Alternative approaches such as using a high frame rate stream where alternate frames are for alternate eyes introduces the risk of a left right swap, especially where any processing or frame re-synchronization is performed.

It is tempting to consider encoding left-right images without a reduction in resolution by using either 3840x1080 (for side-by-side) or 1920x2160 (for top-and-bottom). Unfortunately, these sizes of frames fall outside the maximum defined by H.264 for level 4.2 codecs. It seems likely that at least some chipsets will be able to operate with such resolutions, and this may be a desirable avenue of exploration.

Manipulation

There are numerous occasions where video is scaled by an STB, and one such case is picture-in-guide. Figure 3(a) shows how this should occur for side-by-side formats. This approach introduces two main problems: the difficulty of performing the video manipulation, and the potential for an unpleasant visual impact.

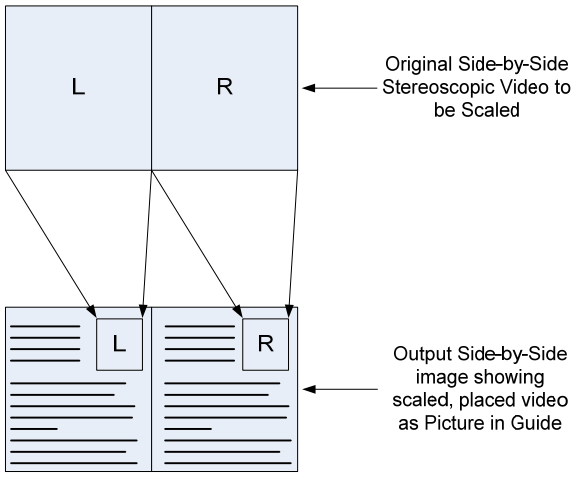

Figure 3(a): Picture in Guide and 3D Scaling

Where the EPG is making use of 3D effects itself, the range of depths used by the EPG may conflict with those of the video, resulting in a strange, and often unpleasant, effect. It may be preferable to simply operate the EPG in a 2D mode, as shown in figure 3

(b) or use only 2D video within the guide as shown in figure 3 (c).

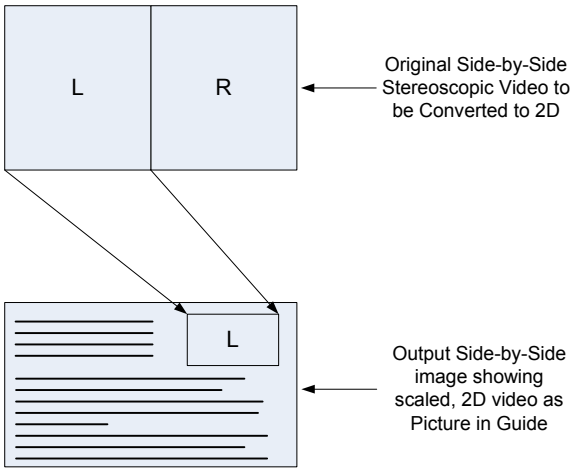

Figure 3(b): Picture in 2D Guide with 3D to 2D conversion

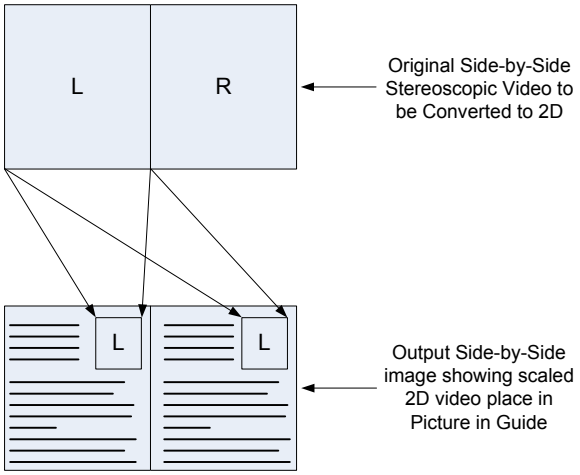

Figure 3(c): 2D from 3D Picture in 3D Guide

Similar visual issues occur with picture-in-picture and where both pictures are in 3D. In the authors' opinion, the conflicts are significantly increased above those with picture-in-guide. Whilst the effect can be partially reduced by providing a border around the inserted picture, to match with the video this border would need to occur in 3D space, and match with the volume shown in the content.

Certain video manipulations, such as those shown in figures 3(a) and 3(c) are significantly different from those normally used in 2D, as they involve two rescaling operations on each video frame. Therefore, these operations can present a significant challenge to existing hardware. In comparison, operations such as those shown in figure 3(b) are common for existing hardware and so easy to implement.

GRAPHICS

One of functions provided by the STB is the provision of graphics or on-screen displays (OSD). This ranges from support for closed captions through channel information banners and interactive applications to electronic program guides (EPGs). There are a range of challenges when handling OSDs and graphics for 3DTV, varying from the relatively simple such as ensuring that they are correctly visible, and that they do not conflict with the video, to including changes in design.

Readability and Format Awareness

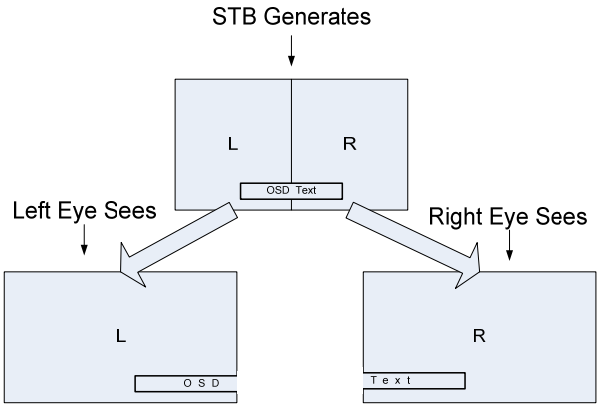

Figure 4: Impact of lack of S3DTV awareness on graphics

Drawing an OSD in a 2D without awareness of the underlying stereoscopic format results in images that are both unreadable and exceptionally disturbing to the viewer. An example of this effect for the case of side-by-side video is shown in figure 4. The top-and-bottom format produces different but still disturbing results with the OSDs only being visible in one eye. By comparison,

frame interleaved formats do not have this readability issue, if handled correctly in the hardware[1].

The readability issue is solved by adapting the graphics stack so that any OSDs are displayed as two images—one image for each eye. This clearly requires accurate signaling of the format so that the STB can alter the details of the graphics to match the underlying video format. When correctly implemented, applications providing graphics do not need to be aware of the 3D nature of the underlying video, as the graphics stack can handle the relevant translations transparently. However, as we shall discuss, there are several reasons why applications may, and in some cases should, choose to be 3D aware.

Depth Conflicts and Awareness

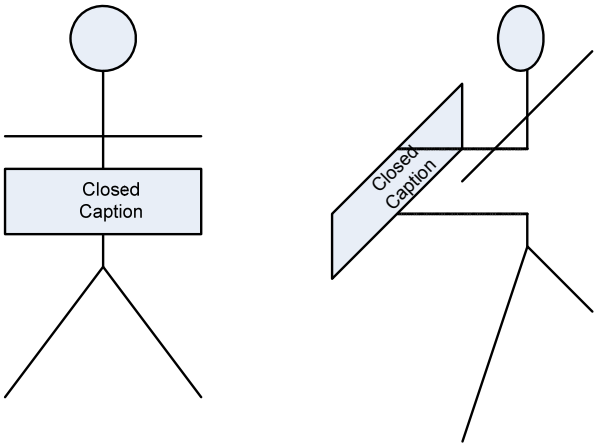

Figure 5: Depth conflicts with Closed Caption Overlays

Once graphics are drawn correctly for the underlying S3DTV video format, the next challenge is to consider where they should be placed. Previously, OSDs only had an x and y location, but with S3DTV they potentially also have a depth, or a z location. Likewise, the video, and the objects in it, occupy a set of depth locations. Careless placement of video can result in a conflict (a visual dissonance) between the objects in the video and the OSD. This happens when graphics are drawn which obscure objects in the video that the viewer knows should be in front of the OSD. An example of this is shown in figure 5.

The perceived depth location of graphics is controlled by the relative positioning of the left and right eye images, and so the STB is able to control the depth at which an object is seen. Figure 6(a) shows the relative offsets of the left and right eye images for a side-by-side format, and figure 6(b) shows the resultant depth that a viewer will see.

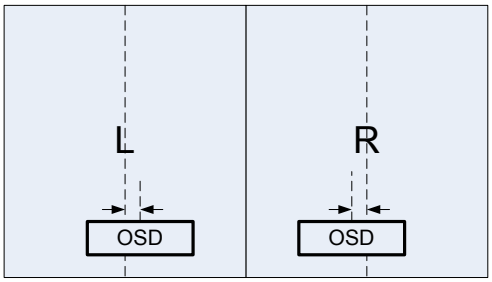

Figure 6(a): Offsetting of OSD in side-by-side S3DTV

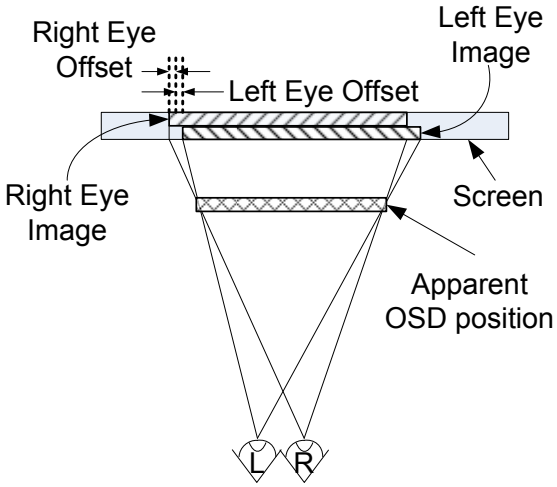

Figure 6(b): Apparent OSD depth position

It is possible to choose a fixed depth position, and always place the OSDs at that location. For example, to minimize the depth conflict, placing the OSD in front of the screen such as shown in figure 6 seems

[1] Some proposed schemes for internal handling treat the frame interleave pair as a single large frame, and so can have identical problems to those of frame compatible formats.

sensible. For short term viewing, this is an adequate general solution, but it does not properly address the entire problem. Firstly, it is possible that occasionally the underlying video will come further forward than the chosen fixed location. Secondly, and more importantly, this can place the OSD significantly forward of objects in the video – unnecessarily increasing the depth budget of the content.

Long durations of high depth budget viewing can increase eyestrain, and consequently approaches to OSD placement which increase the range of depths in use may reduce the appeal of 3DTV. Much content is deliberately created with a careful and conscious choice over the amount of depth that is used. Thus ideally any OSD, be it a closed caption, information banner or an interactive application, should be aware of the depth of the content and seek to fit within, or very close to, the depth range chosen by the content creator.

Design of S3DTV Graphics

The nature of stereoscopic TV is that it provides an illusion of 3D though the brain fusing two images together. In the real world, these two images would change continuously in relation to even the slightest movement of the viewer's head. Without the parallax cue from this continual change, the remaining cues that help the brain fuse the two images together become even more important. In our exploration of the design of 3D graphics[2] we have identified some areas that appear to assist the brain in fusing these images together, or that provide stronger cues, and so result in reduced eyestrain and brain fatigue.

[2] These explorations were performed on polarized, line-interleaved displays, driven by a PC application that rendered a range of graphics at the native format of the display (i.e. no use was made of side-by-side transformations) at 60Hz.

Graphics representing objects with volume present stronger visual cues compared to 2D graphics that are placed at a given depth. Graphics that are 3D objects, or 2D images placed on 3D objects, result in images that are easier to see and appear far more natural.

It is tempting to simply use one image and shift, or offset it, differently for each eye. Whilst this works for a flat planar image, this does not work well for a 3D object, since each eye would normally see a slightly different view. Figure 7 shows, in a somewhat exaggerated fashion, how the left and right eye images differ with a simple box. Thus the design process should ensure that a true 3D model is used and different images are generated with the correct perspective for each eye.

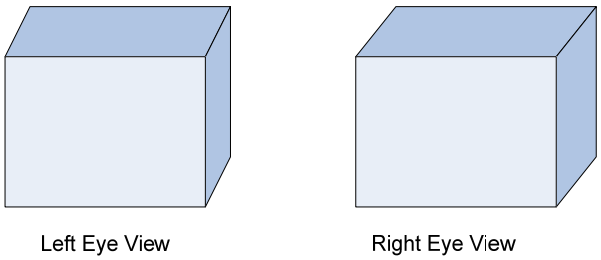

Figure 7: Differences between left and right eye views

When the object is perfectly square onto the screen there are very limited, or no differences between the left and right eye views. Tilting the object in one or more directions introduces differences between each eye's view of the object and provides additional depth cues to help the brain work out the depth placement of the object. Graphics actually look better off-square.

The use of motion on an object provides a much stronger depth and size presence. In part, this is because appropriate motion adds implicit and changing parallax cues through shape changes during movement. So an object that moves into view, changing its depth and rotation as it moves, appears easier on the eye than an equivalent object that simply appears

at a fixed location. Although this effect has a small persistence once motion stops it is not indefinite and it seems beneficial to provide continual animation if only for a small part of the graphic. This provides cues similar to the motion of the viewer's head provides in the real-world, or mimics the continual motion of many objects.

However, there are two aspects of motion that require care. Firstly we found that very fast motion makes it difficult to perceive objects and nearly impossible to determine their depth. Indeed, fast motion provided a worse experience than no motion. Secondly, we found that much of the benefit was lost if the animation process was at too low a frame rate.

The use of a static lighting source helps the reality and natural appearance of objects, especially where lighting related cues such as shadows and color variance are correctly generated and dependent upon true 3D placement. However, it is important that the effects of lighting are updated as objects move and that all objects are subject to the same lighting effects.

The texture of objects also provides depth cues, providing additional reference points on the surface of objects that help in calculating binocular disparity. However, care should be taken to avoid excessive texture, or random noise style textures, as these do not appear to strengthen the 3D presence.

The final aspect is the equivalent of the well-known 2D concept of a safe area. In 3D, we refer to this as a safe volume, and in particular the edges of the display need to be avoided when placing objects in front of the screen. Even objects behind the screen benefit from avoiding the edges. Placing images too close to the viewer is problematic, especially for longer durations (very short periods, especially for disappearing objects is not such a significant problem). Placing graphics far back into the screen is not a significant issue, except for the increased risk of conflict with any underlying video.

Bitmaps and GPU Advances

In the above discussion we have touched upon the need for different representations for each eye to provide some of the correct cues for the stereoscopic illusion, as well as the desire to animate these representations. Earlier, we touched on the need to be able to adapt the placement of graphic objects to differing depths. In many traditional systems, graphics rely heavily (though not entirely) on bitmaps – pre-computed representations of the image to display. Achieving the above goals, especially where multiple formats are to be supported, can easily result in a need for an unacceptably large number of bitmaps.

Newer STB chipsets are becoming available with a powerful graphics processing unit (GPU). These GPUs are able to take complex, abstract representations of a scene, often based on a mesh of triangles with lighting parameters, texture information and camera viewpoint specifications, and then convert this information efficiently into images for display. This approach allows for assets that represent the OSDs to be handled as abstract models and then converted as needed into the appropriate S3DTV format, and placed at the relevant depth.

Moving graphics towards abstract models and exploiting GPUs therefore provides an efficient method to achieve many of the goals above. When implemented correctly, this approach can allow a generic engine to support any abstract model so removing or reducing the need for a new graphics application or set of functions for each graphical asset.

METADATA AND SIGNALING

The sections above have identified a number of STB areas that require updates or changes. Many of these are significantly simpler or sometimes possible only where new or extended signaling or metadata is provided in the broadcast.

Format Signaling

Both AVC/H.264[2] and HDMI[1] provide signaling that is associated with each video frame, and that indicates the format of the stereoscopic data present in the frame. For HDMI this allows a display to perform the correct translations. In the same fashion, the presence, ideally mandatory, of signaling in the video stream allows the STB to adapt its graphics operations to the native, underlying video format.

By providing the signaling in the video, perfect synchronization of the format is enabled. However, this does mean that advance information is not available to the STB, which may be of importance in deciding whether or not to present the item to the viewer. Thus introducing additional signaling is needed in the broadcast. Such information is also useful to assist the box in identifying if it can handle the transmission, and could allow the STB to pre-allocate any increased resources it requires to support S3DTV.

Depth Information

It is possible to provide a single fixed depth value, but as has been discussed above it is desirable for this value to vary to reflect the content. An example of this is shown in figure 8 which represents a bird's eye view of two people walking down a corridor towards the camera. In figure 8(a) the speaker is some distance from the caption, but the caption is placed at the depth position the speaker will reach when the caption disappears. In figure 8(b) the second speaker starts, with a new caption depth placement, which is re-used when the speakers stop walking, as shown in figure 8(c). It is assumed that it is preferable to keep a single caption in at a single depth over the duration of its display, as shown, different captions in the sequence can occur at different locations. However, gentle motion of captions, with a suitable scaling as they approach or recede, is also potentially possible, and ideally the depth information should allow for that eventuality.

The information required for depth could be embedded in the closed caption data stream; however that implies that the values are only available when closed captions are present. An alternative, synchronized stream that might not even be part of the video, allows this information to be used by any OSD, regardless of the presence of caption data. This may be in addition to the depth information contained within a closed caption stream. Various mechanisms for carrying synchronized data exist, and they could easily be extended to carry depth information.

Such a stream of depth information may be generated at the head-end, or during the captioning process. Such depth extraction technology has been demonstrated by Technicolor[3] for off-line caption support and could be implemented within a stereoscopic aware encoder.

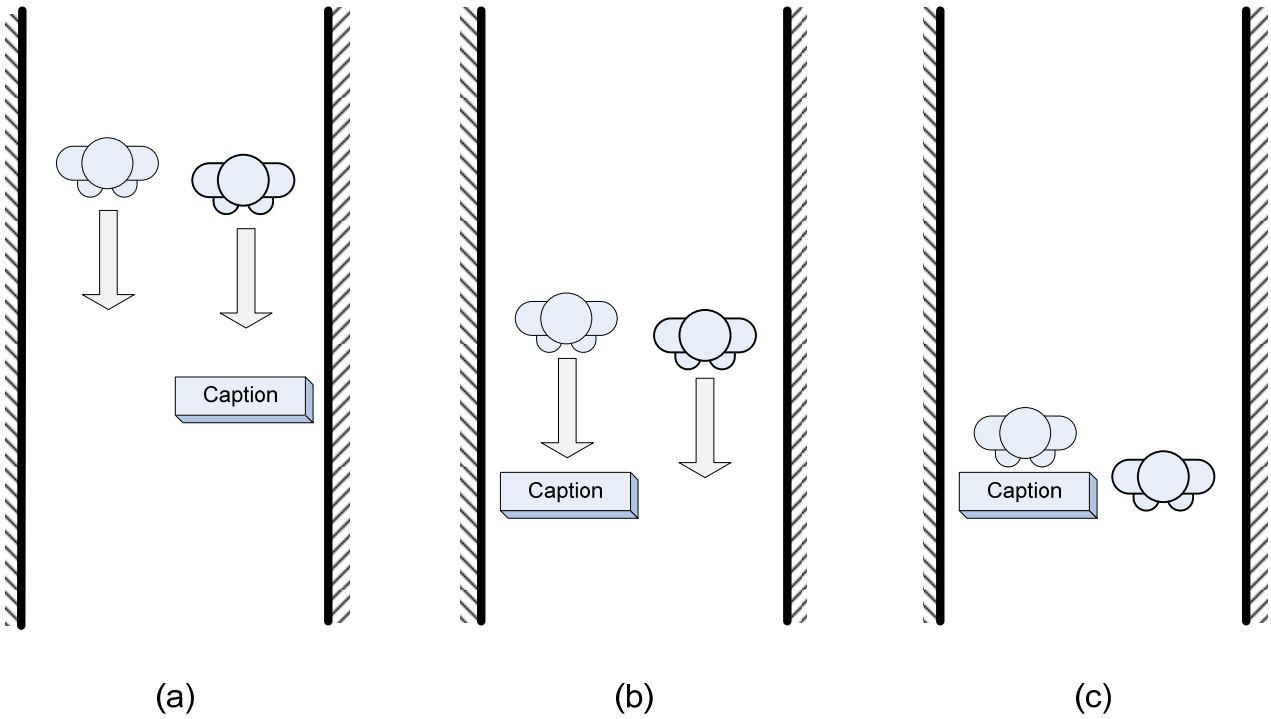

Figure 8: Bird's Eye View Example of Dynamic Depth Caption Placement

Additional Data To Enhance Graphics

When drawing graphics using a GPU, APIs such as openGLES, allow the setting of a wide range of parameters that control how the graphics appear. Some of these parameters correspond to information that is, at least theoretically, available in content creation and production. Examples of such parameters include information typically associated with the camera, such as the focal length of the lens, or that may be known from the setting, such as (e.g. in a studio) the primary lighting directions and sources.

Providing this information to the graphics system allows for a better matching of graphics to the underlying video. This is of most importance where the graphics are closely connected with the video, such as for interactive applications. Developing a system that can carry a wide range of synchronized metadata provides a means for minimizing conflicting cues between the video and the graphics overlays. This, in turn, potentially reduces the unpleasant side-effects from which some people may suffer with longer duration use of mismatched 3D graphics embedded in 3D video.

CLOSING REMARKS

In this paper we have looked at the reasons why the STB needs to be aware of the S3DTV content it is handling, and discussed the basic updates that are required. These start with simple changes to ensure that graphics and OSDs are drawn in a readable fashion, placing them correctly in depth and finally looking at how they can be designed and generated to give a true S3DTV experience. In a similar fashion, we have also looked at issues with processing the video, and explained the limits on performing operations that are simple with 2D but difficult or less desirable to do in S3DTV.

We have also looked at the areas where additional signaling is required, or beneficial. This starts with the S3DTV format in use, and moves through the signaling of depth

information, and discusses some additional metadata that may be useful for graphics.

In closing, it is worth emphasizing how popular S3DTV is likely to be for many viewers. These viewers however will not and should not be aware of the issues outlined in this paper. They will expect everything that they are already familiar with in their television experience to work perfectly in 3D. The areas we have discussed in this paper allow these expectations to be met and help the STB provide its part in the compelling experience of S3DTV.

REFERENCES

1. High-Definition Multimedia Interface Specification Version 1.4a (extract of 3D Signaling Portion)

2. ISO/IEC 14496-10:2009 Amendment 1: Constrained baseline profile, stereo high profile and frame packing arrangement SEI message

3. Technicolor Brings 3D to the Home and Beyond, January 2010