3D Throughout the video chain

Bart Barenbrug; Philips 3D Solutions; High Tech Campus 27, 5656 AE Eindhoven, The Netherlands

Abstract

This paper provides an overview of 3D technology throughout the 3D video chain, ranging from content creation to distribution and 3D displays. Central are the formats in which 3D video can be represented and their suitability for different applications such as 3D cinema, 3D TV, 3D mobile, gaming, and advertising. It is shown that both stereo and image+depth representations are important. Which of the two is more suitable depends on the application and stage in the 3D video chain. So both are needed, as well as conversion between them.

Introduction

Stereoscopic, or 3D, imagery is already several centuries old. Stereoscopic drawings were already made in the 16th century, well before photography was even invented. But only now the technology is starting to become available to realize 3D video on a consumer-level scale, in the cinema, at home, and for several other applications which will be mentioned in this paper. With 3D display devices becoming available, the link with content creation, distribution, and formats becomes important. Different applications have different requirements in this respect. This paper provides an overview of different 3D technologies and their impact on a number of (potential) applications to examine which formats are suited.

The paper is structured as follows: first, 3D stereoscopic vision is introduced briefly, along with the different display technologies for providing the stereoscopic depth cue to viewers. Different display technologies lead to different 3D formats. After a discussion of these, we can look at suitability of these for different application areas. Since there are several formats, format conversion will be discussed, before we conclude with some final remarks.

3D vision and display technology

We perceive depth through a number of so-called depth cues. Many of these are present in 2D imagery. For example foreground objects occluding background objects, and far-away objects appearing smaller (for this we need to know their size) and also less saturated in color (an effect that is extreme in the presence of fog).

Motion can also provide depth information: a moving camera will also reveal spatial relationships between objects since far-away objects move differently than objects that are nearby the view point. This is related to the depth cue that we add with our 3D screens: just like a camera that is in a different position at different points in time and thereby reveals depth information, we view the world using two eyes that are spaced apart, and thereby provide a slightly different outlook on the world. The difference between the images from our left and right eyes provides us the binocular or stereoscopic depth cue. Since our eyes are shifted horizontally apart, the differences between left and right take the form of apparent horizontal shifts of objects in the scene, with the amount of shift relating to depth. This shift is called disparity.

An example stereo pair (originating from [1]) is shown in Figures 2 and 4. The flower is closer to the viewer than the green leaves behind it. This is visible, for example, by looking at the background to the left of the flower leaf that touches the lower image border. More of the background can be seen in the left image than in the right image (because you "look around" the leaf more from a position more to the left): the flower seems to have shifted with respect to its background between the two views: a disparity from which our brain can interpret depth.

In order to provide this stereoscopic depth cue to the viewer, a 3D display will have to show a left-eye view of the scene to the viewer's left eye and a right-eye view of the scene to the viewer's right eye. This is often done using displays that send out both views everywhere, requiring the user to wear glasses that block the right eye view from the viewer's left eye and vice versa. The two images can be made separable using color (the anaglyph process, using the red/green or red/blue glasses), polarization (using passive polarized glasses), or time (using active shutter-glasses). Invariably, glasses block half of the light and depending on the application can be cumbersome to wear.

A different class of displays are the so-called auto-stereoscopic displays which do not require the user to wear glasses. These displays separate the different views at the display side. Usually, this is done either using lenses (lenticulars), which focus the light coming from the pixels belonging to the different views into specific directions, or using barriers, which block the pixels containing the other view from the viewer's eyes. In this way regions in the space in front of the display are created where the left view is visible and regions where the right view is visible. If the user is at the boundary of such a region such that his left eye is in a left-view region and his right eye is in a right-view region, depth is perceived. The obvious drawback is that the viewer has only limited positions where proper depth is perceived. Furthermore, for every left/right boundary also a right/left boundary is present where depth is inverted. One solution to this is to track the user and adjust the directions into which the different views are sent. This requires precise tracking and might limit the number of users that can perceive the proper depth simultaneously.

Another way to circumvent these problems is the use of *multi-view* auto-stereoscopic displays, which do not send out only a left and right view, but more than two views. The principle is shown in Figure 1 where for example view 4 is the right-eye view for the pair (3,4) and the left-eye view for the pair (4,5). Using multiple views not only provides the stereoscopic depth cue, but also motion parallax: by moving his head, the viewer can "look around" objects in the scene.

Since the views are spatially multiplexed, the resolution of the underlying 2D display panel has to be distributed over the different views, so there is a compromise between the number of views and the resolution per view. This compromise will lead to different choices for the number of views for different applications. For example in mobile applications, there is usually less resolution avail-

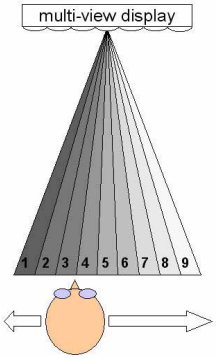

Figure 1. A multi-view auto-stereoscopic display provides a wide viewing area

able than in for example the 3DTV application, so in mobile 5 views are common, whereas 7–9 views are common for applications on larger screens.

Some cross-talk between views is required to have a smooth transition between views when the viewer moves sideways. Having more views helps reduce the need for this cross-talk (while keeping a wide area where proper depth is perceived), so when future display resolutions increase, the number of views can also be increased, allowing for less cross-talk, which in turn allows for better separation between the images projected into the viewer's left and right eyes, and thereby a larger depth range. So similar to how 3D cinema is taking advantage of developments in digital projection, multi-view displays take advantage of the fast increase in display resolution (which like with still cameras is reaching diminishing returns for 2D quality) to add depth and increase the quality of the depth perception.

3D formats

For stereoscopic displays, the stereo format where a left and right image are transmitted is the logical choice. This combination provides transparency for the content makers who can preview the exact images as they will be shown to viewers, and since there is a 1:1 match between the format and the views that are displayed, there is no specific 3D processing required.

However, for multi-view displays, the two views that stereo provides are not enough: more views are needed. Interpolation between or extrapolation between views requires analysis of the stereo material. This will be further discussed in the format conversion section of this paper. A representation that makes it much easier to render different views of the same scene is the image+depth format, where only a single image is enriched with a so-called depth map. A depth map provides a depth value for every pixel in the image that says something about how close or far away from the viewer that pixel is. Figure 3 shows an example of a depth map where a white color in the depth map means "in front" and black "in the back", with the gray-scale values in between providing intermediate depths. Using such a depth map, alternate views can be generated cost effectively, in real-time. See for example [2] for details on how this is done. A result is shown in Figure 5 where a "right" image is shown, which has been re-created from the original left image (Figure 2) and the depth map (Figure 3).

Compared to stereo there are several advantages and disadvantages to the image+depth representation. Image+depth provides a display-independent interface (not depending on the number of views, or how images are interleaved) and thereby decouples content creation and distribution from display. Since depth is explicit, it can be used in compositing, and the depth signal can be manipulated explicitly (see for example [3]). Rendering alternate views from image+depth assures that no vertical disparity is introduced. Vertical disparity can cause discomfort and can be hard to avoid when recording live stereo material (keeping the cameras aligned is a challenge). Depth maps in general compress better than images (no color channels to encode, and smooth objects interiors with little texture), which is advantageous for storage and transmission. Using image+depth, it is easy to vary the amount of depth and placement with respect to the display, tailoring the 3D experience to the capabilities of the display. One important aspect in this area is the size of the display: the relative amount of disparity compared to the display width should be quite different on a small mobile screen than on a huge cinema screen, so the fixed disparity in a stereo pair hinders good display of the same content on such different platforms. Also, by integrating the rendering of the views into the display, the optical and the processing parts of the system can be tuned to one another: processing can help counter-act some optical effects (such as cross-talk), but not others (such as banding), so different choices can be made for the optical system knowing what can be compensated for in rendering. Having the depth explicit in image+depth enables the use of depth-dependent filtering in this area.

Stereo is better suited as an acquisition format, since image+depth cameras often have a limited range and limited quality, and stereo cameras can leverage the ongoing developments in 2D camera technology. Also, image+depth does not handle viewing-direction dependent lighting effects, such as highlights or transparency well, though these effects are rarely prominent in video.

Also, stereo provides some more occlusion information: the image from image+depth lacks the visual information that becomes visible from behind objects when moving the view point sideways, and the second image in stereo may contain such information.

This short-coming is visible when comparing Figure 5 with Figure 4: to the right of the flower leaf that touches the lower image border, some black background can be seen in the original right image, but this information is hidden in the left image, so cannot be recovered when only using the left image and the depth map to generate a new right view. So the part of the image that becomes visible has to be "made up". The combination of the left and newly generated image will be consistent in the sense that there are no contradicting depth cues, and as long as the holes are filled with something reasonable, any artifacts will not detract from the 3D experience.

Also, the amount of occlusion information present in the right frame is exactly enough for stereoscopic displays, but for multi-view displays it is only enough information when interpolating new views between the left and the right views. The amount of depth perceived by a pair of such interpolated views is less than with the original stereo pair, so some extrapolation is also required, and then also stereo lacks the proper occlusion data.

Finally, the image+depth format can be extended with extra information layers to really solve both the occlusion and transparency problems. An example is the use of multiple (if need be, sparse) layers of image+depth (see for example [4]). This makes the image+depth format extendable and scalable (to make a low-cost multi-view renderer, it is an option to let it ignore such extra information).

Applications

There is currently a lot of attention for 3D cinema. With stereo projection being a relatively inexpensive add-on when switching to digital projection, with the technology being there to economically convert 2D movies to 3D, and with the success of movies like Polar Express and Chicken Little in 3D, 3D cinema is on the rise. George Lucas has announced at ShoWest 2005 that he wants to re-release the Star Wars movies in 3D, and Disney has announced that the successor to Chicken Little will be in 3D as well. For the cinema application, stereo is the natural format: the glasses are not much of a problem at the cinema where you solely watch the movie and not do anything else for which the glasses might be distracting or cumbersome. Also, the movie is watched in the dark, so blocking half the light is only a minor issue. Furthermore, with the stereo format, the studios have exact control over what is shown, and there is no need for multi-view rendering at very high resolutions.

However, most money in the movie business is being made at home with the DVD and TV sales (see for example `http://www.factbook.net/wbglobal_rev.htm`). And once 3D cinema becomes more common, people will also want to watch 3D movies at home. However, watching at home is done under different circumstances than at cinema: the environment is much brighter, and TV watching can be much more casual, with the TV running while doing other activities. Glasses are much more of a hindrance under these circumstances, and since multiple viewers are often present, auto-stereoscopic multi-view displays are much more suitable. Of course there will be home cinema enthusiasts who will want to recreate the cinema experience as close as possible, so there certainly will be stereo projection at home, but for a large group, glasses-free watching will be a better choice. For 3D TV transmission, this also means that bandwidth can be saved using the image+depth format, unless the transmission contains both stereo and depth to service both kinds of displays.

Bandwidth is also an important issue in mobile applications. In this area glasses are also not very suited, since watching the display is alternated with watching around a lot. Next to video watching, also gaming is an important application in this domain. 3D graphics rendering of multiple views is a task that a graphics card can do, but it does require a lot more work than just rendering a single view, especially when the scenes become more complex. It is relatively easy to output a depth map from a game without too much overhead since the game's z-buffer contains the right information. Multi-view rendering is not scene-dependent, so its performance is much more predictable and can be done as a next pipeline stage, leaving the graphics card free to render the one image as best as possible. The same reasoning holds for gaming in the home, and since game consoles often use the TV as a display, this matches well with image+depth capable 3DTVs.

For medical applications and scientific visualization, there is less of a chain to take into account, since distribution is in the form of the 3D data files. So 3D rendering and display is more local. Since transparency is often used in these applications, and some more rendering hardware can be afforded, it is often more practical to simply render all views from the 3D data that is present, and display that locally either in stereo or multi-view.

An application where auto-stereoscopic multi-view is a must is the signage advertising market. The 3D effect can be used to attract the attention of people who pass by a 3D display in for example malls, airports, or shop windows. These people will not be wearing 3D glasses, and there are many viewers, so auto-stereoscopic multi-view is a must. In this application a lot of computer graphics material is used, where high quality depth maps can easily be generated, resulting in much better results than real-time estimation of depth from stereo. Here generating the content directly in the image+depth format and use that throughout the chain makes more sense.

Our signage displays therefore use image+depth as interface format, and the multi-view rendering is performed inside the display. We also provide tools for content creation, such as a plug-in for 3D Studio Max®, which allows generation of animation directly in the image+depth format. Furthermore, as evidenced by Figures 2–5, we are working on extraction of depth maps from stereo material to enable access to stereo content for our displays, enabling the use of non computer generated video.

As can be seen from the previous discussion, different applications require different display types and therefore different formats, and also different formats fit better in different parts of the video chain (stereo for camera recordings, image+depth for some editing work where the explicit use of depth can bring advantages, both for different display types at the end of the chain). So both formats are needed, and should be inter-operable: it should be easy to show trailers or clips from movies on a 3DTV or a 3D mobile device, or to re-use the same commercials initially developed for signage displays on those three platforms. Also for games cross-platform portability is a plus.

3D format conversion

Since we have seen that both the stereo and the image+depth formats make sense, the question arises if we can convert between these two representations. Going from image+depth to stereo is equivalent to rendering one extra view from the image+depth representation, so this is relatively easy to do.

The harder problem is converting from stereo to image+depth (which boils down to estimating a depth map given a stereo pair). The depth information that is implicit in the stereo pair has to be made explicit by finding correspondences between left and right images. Optionally any occlusion information has to be encoded in extension layers. This is a task that does not always have a unique solution. For example, depth changes cannot be detected within homogeneous parts of images, since there is simply no detail to guide the correspondence finding.

Many different algorithms have been devised to tackle the problem. A fair many of them are listed and evaluated at [5], though the test images there might not be representative for all applications (for example the aspect of temporal stability of an estimator used for video is not taken into account). Depending on the application, real-time performance is required (e.g. a live broadcast recorded using two cameras but displayed on a multi-view screen), and this can pose extra restrictions on the approach taken. Next to the actual matching algorithms, work has been put into pre-processing, such as aligning left and right images to enable the use of line-based disparity estimators, and also post-processing (such as a left-right check between two estimates to find consistencies, and filling in the holes where inconsistencies were found, see for example [6]).

Since finding correspondences between rather similar images is similar to the problem of motion estimation (which as Philips we have implemented in our television sets already for ten years or so in our Natural Motion feature, which performs temporal up-conversion

Figure 2. *Left image from a stereo sequence*

Figure 4. *Right image from a stereo sequence*

Figure 3. *Depth map generated from left and right images*

Figure 5. *Right view re-generated from left image and depth map*

to increase the frame rate of video), we take a similar approach for depth estimation. This is detailed in [7].

Conclusions

For many applications, auto-stereoscopic multi-view displays are more suited than glasses-based stereo systems. Though stereo is and will remain an important format, especially in video acquisition, there should also be attention for the image+depth format. Even though real-time estimation of depth from stereo is possible, direct generation of depth maps or even off-line processing (optionally with manual interaction) can result in better quality depth maps. This is for example very important in the 3DTV application: creating depth maps at the authoring side and distributing those along with, or instead of a second image of a stereo pair allows for a higher quality 3D experience. This is why there should be more attention for the image+depth format in the areas of authoring and transmission.

References

[1] "Okizarisu" movie from `http://stereomaker.net/ai/3dflower/3d_f001.htm`

[2] R. Berretty et al, Real-time rendering for multiview displays, in Proceedings of the SPIE, 6055, 2006.

[3] Nick Holliman, Mapping Perceived Depth to Regions of Interest in Stereoscopic Images, in Stereoscopic Displays and Applications XV, 2004, available as `http://www.comp.leeds.ac.uk/edemand/publications/hol04a.pdf`

[4] Steven J. Gortler and Li-wei He, Rendering Layered Depth Images, Microsoft Technical Report MSTR-TR-97-09, a.o. available at `http://research.microsoft.com/research/pubs/view.aspx?type=Technical%20Report&id=20`

[5] Middlebury stereo vision page at `http://cat.middlebury.edu/stereo/`

[6] P. Fua, Combining Stereo and Monocular Information to Compute Dense Depth Maps that Preserve Depth Discontinuities, International Joint Conference on Artificial Intelligence, Sydney, Australia, 1991, pp. 1292–1298

[7] Ralph Braspenning and Marc Op de Beeck, Efficient View Synthesis from Uncalibrated Stereo, in Proceedings of the SPIE, 6055, 2006.

Author Biography

Dr. Bart Barenbrug received his master's degree and PhD at the Technical University of Eindhoven working in the field of computer animation. In 1998 he joined Philips Research to work on several topics in the field of computer graphics, in particular combining CG with video processing technology. In that capacity he started to work specifically on processing for 3D displays in 2003. Early 2006, he joined Philips 3D Solutions (www.philips.com/3DSolutions).