3D Vision: Technologies and Applications

SE Selected Topics in Multimedia 2012W

Werner Robitza
MNR a0700504
a0700504@unet.univie.ac.at

ABSTRACT

3D vision is a promising new branch in the entertainment industry that encompasses applications for cinema, television and video games alike. The twenty-first century uprise of three-dimensional technology can only be explained by a mix of technological and social factors. Despite this, vendors are still struggling with reaching a critical mass for the establishment of new products. Some users experience discomfort when using stereoscopic devices due to the difference in perceived and reproduced reality. To address these problems, recording and display technologies need to be compared and evaluated for their applicability and usefulness. We give an overview on today's 3D technology, its benefits and drawbacks, as well as an outlook into the possible future.

1. INTRODUCTION

The idea of capturing and presenting content to viewers in three dimensions is not new. Stereoscopic cameras have existed since the beginning of the twentieth century. The idea of stereoscopic TV was born in the 1920s [23], and stereoscopic cinema made its colorful debut with the 1952 movie "Bwana Devil" [38].

Ever since the sixties, special-interest 3D theater productions have become popular within the IMAX network of cinemas, which now extends to almost 600 3D-capable cinemas worldwide [2]. Although 3D cinema never became a first class citizen of the public entertainment, it began to rise in popularity. A notable event in the history of cinema altogether is marked by the 2009 3D movie *Avatar*, which—with a budget of estimated $230 million—not only was one of the most expensive movies to make [21], but also set records becoming the number one movie to lead the all-time box office charts.

Despite its uprise in popularity, especially in cinemas, 3D television (3DTV) can hardly be found in the homes of the average consumer. In fact, the theoretical availability of consumer 3D technology and equipment seems to surpass the actual demand for 3D in the home. Similarly, mobile 3D devices have not seen wide adoption as of yet. This raises multiple questions, including the search for why the breakthrough is still overdue—or why it may not happen at all during the next years. In search of answers to these questions, one discovers technical and human reasons alike.

This paper aims at explaining both the anatomic and technical background of three-dimensional vision, capturing and reproduction, as well as describing the key issues that vendors currently face at the stage of deploying 3D in the home. In Section 2 we give an overview on the principles of the human visual system, depth perception and health effects. We then describe the technologies used to record, store and reproduce stereographic content in Section 3. An overview of the current developments and an outlook to the near future of 3D is given in Section 4. We conclude with Section 5, summarizing the paper.

2. PRINCIPLES OF 3D VISION

The complex human visual system allows us to perceive the surrounding world in three dimensions: To width and height of observed objects we add the notion of "distance" (or "depth"). As photons travel through their optic components, they create a two-dimensional representation on the retina, once for each eye. The human brain uses this planar picture from both eyes to reconstruct a three-dimensional model of our environment. This is possible because the images are disparate. The ability to perceive 3D does not solely depend on the eyes though: Several other factors allow for a more precise modeling of depth.

2.1 Depth Cues and Perception

To generate depth information from its sensory input, the brain relies on a number of so-called "cues" . In addition to the images captured by the retina, the eye muscles also provide us with feedback on our surroundings. We can therefore differentiate between *visual* cues and *oculomotor* cues. Visual cues rely on retinal images, whereas oculomotor cues are based on muscle movement.

Visual cues can further be divided into two groups, depending on whether they require one eye ("monocular") or both eyes ("binocular"). According to [8, 19], monocular cues include:

- *occlusion* (e.g. objects overlapping each other)
- *relative size/density*
- *height in visual field*
- *aerial perspective* (e.g. mountains whose colors seem washed out when seen from far away)
- *motion perspective*
- *textural gradients*
- *light and shading*
- *linear perspective*

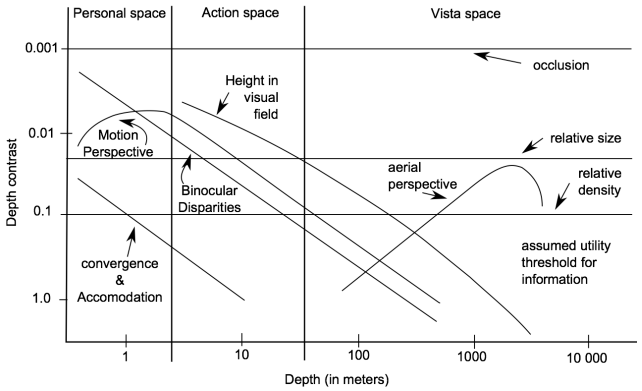

Figure 1: Depth thresholds as a function of distance from observer, with different depth cues. The smaller the threshold, the more potent the cue with regards to estimating the depth. (Figure from [19] based on results from [8])

These cues allow the brain to estimate the depth of an object. Especially the motion perspective is important for a reliable and fast judgment: When we move our head slightly, near objects appear to shift further than far objects. In 3D display technology, we can however neglect most of the monocular cues, since binocular and oculomotoric cues predominate for object distances below 10m (see Figure 1).

Oculomotoric cues include *accommodation*, *convergence* and *pupillary constriction*. These are also known as the "ocular near triad" [27]. The three functions belong to the motoric parts of the human eye and can be modeled as a feedback system, where each component influences the behavior of the other: When the eyes focus on an object of the distance x, they (subconciously) change the optical power of the lens to accommodate for distance x, influenced by the autonomic nervous system. At the same time, the eyes converge to a common point, so that their axis of vision intersects near the distant point. This is necessary because the interocular distance for human eyes is around 64mm [16]. If the optical axes did not converge, the central parts of each retina would not capture the same object.

Once the eyes have accommodated and converged, the brain composes two disparate images of the same source. Only the main object in focus is not disparate. These non-disparate objects lie on a half circle also known as the "Vieth-Müller Circle", or "Horopter" (see Figure 2). The radius of this circle is defined by accommodation power, nodal points of the eye and consequently their convergence angle. The brain can now distinguish between points in front or in the back of the Horopter, which are called "crossed" and "uncrossed", respectively, and estimate their depth depending on the amount of crossing and disparation [30].

Due to the optical features of the eye, the Horopter is surrounded by an area whose size is determined by the eye's depth of field. This area is called "Panum's area of fusion" [29], since actual disparities of points within this range are "fused" into one by the brain (see Figure 2). The disparate points appear as one. In 3D reproduction, points that are displayed outside this fusional area can be seen as double images and cause major disturbances and visual discomfort, as explained later.

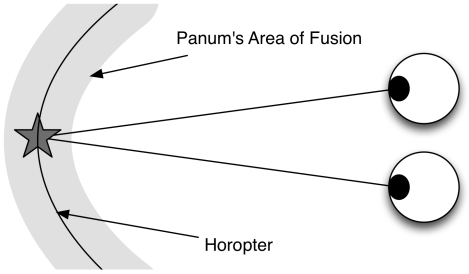

Figure 2: The Horopter (Vieth-Muller Circle) and Panum's Area of Fusion in the field of vision.

2.2 Parallax and 3D Displays

As an attempt to faithfully reproduce a stereographic environment, 3D displays need to offer more than one representation of the same object to the human eye (so-called "views"). Moreover, the representation needs to evoke oculomotoric action in the same way as nature does, so that the brain can correctly interpret the cues. It will then translate images and oculomotoric cues into a model of an experienced three-dimensional world.

In a simple scenario, a 3D display can show one view for the left and right eye each, both on the same plane (e.g. a cinema canvas, a TV screen). To quantify the simulated depth of two corresponding points, their distance in the two views is measured. This distance is called "Parallax". There are three types of parallax, closely related to the locations of points around the Horopter (see Section 2.1). We observe *zero parallax* when both points are shown at the same position. *Positive parallax* is created by a point that lies more to the right in the right-eye view and vice-versa. This parallax is "uncrossed". Similarly, *negative parallax* is achieved through "crossed" points, where the same point lies more to the left in the right-eye view.

The parallax gives us important information about the simulated three-dimensional world. Objects with positive parallax are seen "behind" the screen (the so-called "screen space") whereas negative parallax objects appear "in front" of the screen ("viewer space"). It is also the main point of action when designing three-dimensional content, since not all ranges of parallax can be reproduced, and some might lead to an unpleasant viewing situation [18].

2.3 Visual (Dis)comfort

The parallax determines the experienced depth of an object, created mostly due to monocular cues and convergence of the eyes. However, 3D systems exhibit a principal flaw in the way they represent depth. This causes visual discomfort—for some observers more than for others. This is called the "vergence–accommodation conflict".

Let us assume a plane in a fixed distance x, e.g. a 3DTV screen, 2m away from the viewer. As long as two-dimensional content is seen, the eyes both accommodate (focus) and converge (cross) towards the screen. The main object of attention is isolated in this process, e.g. a news speaker. When the third dimension is added (by providing the eyes with two different views), the autonomic nervous system will simultaneously attempt to accommodate to the distance of the screen where the object is displayed and converge to the point where the object is simulated to be at. As long as

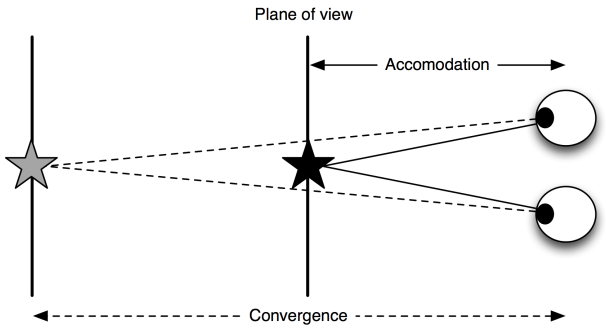

Figure 3: Vergence–accommodation conflict. The eyes converge on the simulated object in the screen space (positive parallax, grey star), while they accommodate at the real plane of view (zero parallax, black star)

the parallax p is zero, both accommodation and convergence points are at the same distance x, e.g. 2m. Once the parallax is shifted in order to simulate a depth of $p \times d$, where d is a factor depending on the environment, the eyes converge to a different angle, namely the one needed to intersect at the object's new perceived depth ($x + p \times d$). The accommodation however stays the same (x). This leads to a mismatch between the two autonomous nerve systems, which are usually in synchronization with each other (Figure 3).

This conflict is often cited as the main cause for visual discomfort, resulting in eyestrain, nausea, headaches or stiff shoulders. The mismatch seems to draw accommodation focus away from the screen, towards the object's perceived depth [33]. Panum's area of fusion (c.f. Section 2.1) also plays a role here: As long as the depth of the object does not exceed the eye's depth of field, there is no need to shift accommodation. In this case, visual discomfort would be kept to a minimum, but content producers have to take care not to exceed the fusional area, whose effective size in turn depends on the viewing situation itself. For a detailed description of this conflict and its consequences, see [10].

The vergence–accommodation conflict is not the only possible source of visual discomfort associated with 3D viewing. In nature, one can observe differences in human visual system from one person to another. One aspect here is the inter-pupil distance (IPD). It can be calculated that for viewers with smaller IPD—assuming the same viewing setup—perceived depth increases when compared to a human with larger IPD, and vice-versa.

Studies show that 5 to 10% of the population may suffer from stereo blindness, thus are not able to fully perceive the world in a stereoscopic manner [18]. There are multiple reasons for this occurrence, ranging from senso-motoric inabilities (i.e. convergence) to other disabilities acquired during childhood. Not only children may suffer from limited perception with regards to 3D content: The ability of the eye to accommodate to a certain distance may decrease with age and lead to difficulties correctly interpreting stereoscopic viewing situations [24]. It is therefore advisable to test one's stereoscopic vision before using 3D applications, e.g. with a *Randot stereo test*, although the efficiency of such tests is disputable depending on the application [20].

It is generally not known to which degree one can expect viewers to experience visual fatigue or even discomfort. The factors leading to these effects comprise different viewing situations and display techniques as well as age, previous experience and daily mood. The transmission quality and severeness of visual artifacts also has an impact on the discomfort, such as cross-talk between the individual views and the composition of the scene [18]. A certain degree of "simulator sickness" may always be expected, yet in psycho-visual experiments carried out with active 3D systems and computer monitors at the University of Vienna, observers did not mention any problems related to the 3D effect, before and after the test [28].

All these health aspects cause overhead when designing or producing content for 3D viewing systems. One has to take into account both typical human factors as well as the possibilities for outliers who perceive stereoscopic content in a different way—or might be prone to experience visual discomfort and fatigue. Recommendations for post-production are still under study [16]. The following section will aim at describing the common production schemes as well as the playback methods used, with regard to these guidelines.

3. CAPTURING AND DISPLAYING STEREOSCOPIC CONTENT

The primary goal of a stereoscopic video system is to faithfully reproduce the sensoric impressions the human typically perceives from a real three-dimensional scene. The 3D system should create a feeling of immersion and a more in-depth experience, which in turn is likely to improve the viewer's subjective value of entertainment services (the so-called "Quality of Experience", QoE). In technical applications, a successful 3D system could for example improve the accuracy of maintenance tasks or the effectiveness of computer-assisted surgery. The number of possible applications is vast, yet it requires specific knowledge to create a successful product.

In practice, careful preparations have to be undertaken to ensure that the filmed objects are properly staged, cameras are accurately aligned, media is reliably stored and transmitted, and the playback system is precisely set up according to the viewing environment and the previous experience of the target audience. It is in the nature of stereoscopic video that not one solution will fit all usage scenarios. For example, the parallax distances used in 3D cinema have different depth effects than those in 3DTV at home, assuming the depth cues seen in Figure 1. This is simply due to the distance between screen and audience. In three-dimensional surgery, precision of display is more important than any "artistic" component.

In the typical production chain for 3DTV and 3D cinema, many aspects can be considered crucial with regards to the final outcome, i.e. the quality of the product. The remainder of this section identifies those.

3.1 Stereographic Recording

In order to be able to present the viewer with at minimum two views, they have to be (a) captured from a real scene, (b) synthesized from existing two-dimensional material, or (c) rendered from digital compositing software (e.g. *Blender*). These methodologies to generate stereographic content all have very different applications.

Figure 4: Semi-professional stereo camera rig with adjustable inter-ocular distance

3.1.1 Stereo and Multi Camera

This native approach at capturing stereoscopic content follows the principles of the human visual system, modeling it with two cameras. The optical axis centers of the cameras are placed next to each other at the average inter-pupil distance. The cameras have to converge to the point that will later be reproduced with zero parallax. In order to achieve this, two approaches can be used: (a) "toed-in", where the cameras themselves are rotated inwards, or (b) "parallel", where the cameras stay fixed with their optical axes crossing at infinity, while only the sensor is shifted inside the camera. This creates a plane of convergence instead of just one point. The latter approach usually results in better image quality, however is harder to set up [16]. The parallel setup allows control over perceived depth by changing the sensor shift h, with the requirement that $h \neq 0$.

The stereo camera setup provides a relatively low-complexity solution for acquiring 3D content. It makes for small shooting rigs and short calibration times, where after setup, 3D footage can be continuously shot. The captured results are natural, however it is very challenging to provide more than two views. For example, when a movie is watched by three or more viewers at home, every viewer might want to see "their own" geometrically accurate view of the scene. This is not possible with stereo camera material, but additional views may be generated in a post-production step. Nonetheless, stereo camera systems are being employed by cinematographers since the production overhead is minimal. Stereo camera rigs for the (semi-)professional market exist, such as the one Figure 4 depicts[1]. For the consumer, handheld stereo cameras are available, mostly with a combined lens system.

An extension to stereo camera, the multi camera approach adds a number of cameras to the filming rig. The same principles apply: All cameras have to converge to one point ("toed in"), or one plane ("parallel") and must be perfectly calibrated. This method results in as many views as there are cameras, which produces high quality and realistic material. Typical purposes are of technical or medical nature. A representative multi camera video set can also be used to test multi-view 3D equipment[2]. However, the required overhead for setup time and calibration as well as the financial means needed make this recording procedure unpractical.

3.1.2 Depth Camera

In situations where only one camera is available, stereographic material can be easily produced with the addition of depth information. Given a two-dimensional image source and a depth map, it is possible to generate an infinite number of disparate images (e.g. one for the left eye, one for the right). By constructing a depth model, the parallax needed to project elements from the two-dimensional source into the screen or viewer space can be calculated for all possible viewing situations.

In practice, depth is captured with an infrared camera, also known as the *Time of Flight* (ToF) camera [14]. It shines infrared light towards the scene and measures the phase correlations of the return signal, which are then translated into depth information. For each pixel in the actual visible-light video, one can now assign a depth value (e.g. between 0 and 255).

Depth cameras usually provide lower resolution than the visible-light cameras (e.g. no High Definition material). However, they are more accurate at finding occlusions and disparations regardless of missing textures or absence of visible light. This makes the depth camera approach very suitable for darker scenes and much more flexible in terms of calibration, since the final image is rendered in a post-production step. A huge disadvantage is the limited visible range of ToF cameras, which makes outdoor shooting impractical.

3.1.3 2D to 3D Conversion

One of the main drawbacks in stereographic video production is the technical and financial overhead. Furthermore, the majority of cinematographic and broadcast material was recorded in 2D only. This calls for a solution to convert existing 2D material into 3D. It would be desirable to be able to simply convert any existing 2D source to 3D, yet this task demands a lot of computational effort.

We can identify three schemes of 2D to 3D conversion: (1) manual conversion, (2) human-assisted conversion and (3) automatic conversion. Naturally, the manual conversion process is time-consuming: 3D artist(s) have to examine each frame of the source video and manually assign depth information to scene objects. Alternatively to stereographing the entire scene, parallax information can be just added to objects of interest, leaving the remainder of the scene in a zero-parallax plane. The latter approach saves time and allows for emphasizing important objects in a 3D scene.

To be able to (semi-)automatically convert two-dimensional material, a typical conversion system consists of two stages [37]: Firstly, it estimates the real depth of an object from a two-dimensional projection thereof. Secondly, it generates the 3D images. In order to achieve the former, one might take into account typical monocular cues like depth blur (as introduced by camera lenses), color gradients and texture occlusion. Automatic systems may estimate the motion parallax by analyzing a sequence of frames from the original video. In this step, an internal model of the 3D scene is created. Human assistance naturally improves the performance, since the possibility of errors is reduced. For example, a 3D artist could supply a rough depth "template" for a scene, which the

[1] Source: `koreaittimes.com`

[2] For example, the Fraunhofer Heinrich Hertz Institute published 16-camera multi-view 3D test sequences under `ftp.hhi.de`, login details to be found under `http://sp.cs.tut.fi/mobile3dtv/stereo-video/`

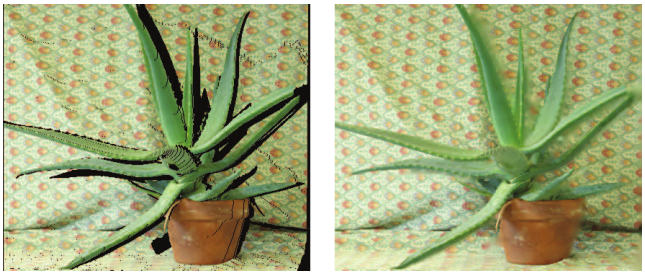

Figure 5: Synthesized view before and after hole filling [32]. One can observe the difficulties for the algorithm to fill textured areas such as along the vertical middle line (behind the plant).

system uses to properly align and estimate the geometry and the depth of scene objects [34].

As mentioned before, the automatic conversion system must synthesize at least two views from the previously generated depth model. This method is called *Depth Image Based Rendering* (DIBR) and consists of subtasks, such as processing the depth model, warping the images and so-called "hole-filling". Hole-filling is a technique often used in DIBR: When an object is placed into the screen or viewer space by shifting its parallax, it exposes an area that was previously occluded, i.e. its background. However, there exists no recorded image information for this area. As an example, in Figure 5, we can observe the occluded tissue behind the plant on the left. On the right, the background was reconstructed with *hierarchical hole filling* (HHF) [32]. Hole-filling algorithms typically fill the disoccluded area by averaging the surround area, thereby "painting" the holes. In Figure 5 it can be observed that this approach does not reliably work for complex textures or strong contrasts. Hole-filling is still under extensive research.

3.2 Storage Formats

Video and audio bitstreams are usually stored in container formats such as defined in the MPEG-2 [1] or MPEG-4 [4] standards. Containers do not necessarily provide facilities to store multiple views of the same source. While in theory, views could be stored as independent bitstreams, this creates redundancy. It occurs because typically, the views depict the same objects, only slightly shifted. Simply adding one view's bitstream could potentially double the required bitrate to transport the media, but broadcast services and storage media are often tailored towards single-view video [11]. Doubling the required bandwidth or even multiplying it by a large number of views would lead to increased costs when compared to traditional broadcasting scenarios.

3.2.1 Multi View Coding

To ameliorate this situation, the ISO/IEC's Motion Picture Experts Group (MPEG) in conjunction with the ITU's Video Coding Experts Group (VCEG) standardized *Multi View Coding* (MVC) as an addition to the H.264/MPEG-4 AVC standard in 2009 [5, 4]. MVC exploits the redundancies between multiple views. It extends the existing AVC standard, thus reuses most of its concepts. This guarantees backwards compatibility with playback devices that are only capable of decoding AVC video.

The main feature of MVC is its inter-view prediction model,

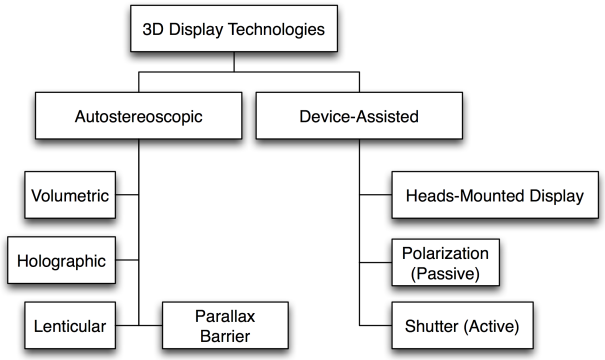

Figure 6: Classification of 3D display technologies

which builds on the inter-picture prediction models from AVC. The encoder may now not only exploit temporal redundancies found within one stream of pictures, but also leverage the similarities between multiple views. In traditional AVC, each picture is in essence composed of macroblocks of pixels. A macroblock may reference one or more other macroblocks from other pictures and only store the difference information to these references, thereby reducing redundancy [36]. In MVC, a macroblock from one view can now reference one or more macroblocks from another view in a similar fashion. As of today, Multi View Coding has emerged as the standard for encoding stereographic video.

3.2.2 Video+Depth Coding

As an alternative to coding both views independently (or with dependencies such as in MVC), one view and the scene's depth information can be transmitted simultaneously. As explained in Section 3.1.2, depth maps allow the receiver to generate an infinite number of views on the fly. This approach makes it possible to stream 3D video in a backwards-compatible way. A receiver without 3D capabilities can simply discard the additional depth channel, whereas a 3D-enabled device can generate the stereographic reproduction on demand. Of course, this method requires a receiver with enough computing power. Moreover, technological advances in 2D to 3D conversion or view synthesis can not easily be propagated to these end devices, whereas already synthesized video transmitted in MVC could leverage the conversion capabilities of a broadcasting provider.

3.3 3D Display Technologies

There exist various techniques to display three-dimensional content, which we can classify depending on the need for external equipment. There is no standardized classification for 3D displays, but loosely following the categories in [23], we can group them into (1) autostereoscopic and (2) device-assisted displays (see Figure 6).

3.3.1 Autostereoscopic Displays

Autostereoscopic displays require no additional gear to create the impression of 3D vision, which is their primary advantage. Without the need for wearing glasses, users can engage in other activities while watching 3D TV or playing 3D games. To a certain degree, viewers can also move around the screen, still keeping the stereoscopic effect. We

can further categorize autostereoscopic displays according to the technology used to build the screen or project the image:

- **Volumetric displays** show the scene spanning all three dimensions, i.e. not by projecting it onto a plane (like for example in TVs). For their complicated technology, volumetric displays are rarely used.
- **Holographic displays** optically project the image with lasers, giving the impression of a 3D scene. Similar to volumetric displays, their use is almost irrelevant to the entertainment industry today.
- **Parallax barrier displays** display both left and right eye views simultaneously in an alternating fashion on vertical stripes of fixed distance. Due to a barrier, each eye only perceives the view designated for it while the other view remains occluded [17]. The technology dates back to 1903 and has since then seen widespread use for its simplicity [31].

  Together with lenticular arrays, parallax barrier designs are mostly used for autostereoscopic TVs. The Nintendo 3DS, a popular mobile gaming console, also features a parallax barrier display. Their main drawback lies in the so-called "sweet spot" that the viewer has to be placed at in order to perceive the 3D effect, which can be very narrow depending on the technology. If the viewer moves out of this spot, they will not be able to see stereoscopic images anymore.
- **Lenticular displays** use cylindrical lenses and project at least two images onto the screen, sliced into alternating columns. An array of lenticular lenses in front of the screen ensures that each eye only sees the matching image slice [17]. This technology allows for more than two views: If the viewer moves their position, another set of images can be exposed. This makes the technology very practical for 3DTV with multiple persons watching at the same time.

Most autostereoscopic devices operate based on very simple principles. This makes them easy to construct and allows the viewer to move around freely while watching, not being distracted by wearing additional devices. Still, the "sweet spot" limits the viewer's free movement, and by splicing the images, autostereoscopic displays cannot offer the same resolution per area as device-assisted displays. They might appear blurry and are therefore potentially less suitable for the consumer market.

3.3.2 Device-assisted Displays

In contrast to the autostereoscopic displays, device-assisted displays require the use of visual aids to perceive stereoscopic content. The necessary additional gear is often very tightly coupled to the display, thus creating "systems" a consumer may buy in a bundle.

- **Heads-mounted displays** consist of two screens placed directly in front of the eyes. They are isolated both from each other as well as from the outside and present each eye with an individual view, thus creating a very immerse viewing situation. While often used for simulations (e.g. flight or train simulations), in medical appliances [9], or when designing and exploring buildings [35], the entertainment industry does not promote them as much as other technologies. This is mainly due to the isolation in which a viewer perceives the content. Interactions with other human beings are practically impossible when wearing heads-mounted displays.
- **Passive stereographic displays** require the viewer to wear a specialized pair of glasses that itself does not need to be connected to a power source or be driven by battery power [22]. A very commonly used and simple technique is called *Anaglyph*. Here, each eye's view is filtered with a different color. The glasses themselves use the same colors, mostly red and cyan, to filter out the corresponding images. Anaglyph systems are cheap and easy to create, yet they don't provide full colors and can therefore be categorized as "toy" systems with regards to today's standards.

  The principle behind most passive displays today is *polarization*. Left and right eye views are projected in polarized light, both orthogonal to each other. The glasses are also polarized in the same order and thus filter out the correct light source for each eye. The main advantage of polarization is that the full frame rate of the video source can be used, since both views are displayed at the same time. This makes for a very smooth presentation. Also, the glasses can be constructed very sturdily and are cheap to create. It is for this reason that polarized glasses are often used in cinemas. The drawback of passive stereoscopic displays is that the polarized glasses themselves minimize the amount of light entering into the human eye, which makes colors less vibrant and dark areas less textured. Also, only half of the original video resolution can be displayed, since the two views are shown at the same time.
- **Active stereographic displays** utilize a pair of glasses that actively creates the stereoscopic effect. As with passive displays, the glasses only work in conjunction with a certain system. In this case, the display shows the left and right views in an alternating fashion, multiple times per second. For example, for video material originally recorded at 50 frames per second, both left and right views will only be shown at 25 fps. The glasses are synchronized with the display and actively block each eye's view for the appropriate amount of time, so that the other eye can see its view [25]. Commonly, liquid crystal fields are used for the shutters; when voltage is applied to the field, they become dark, thus obstructing the view.

  Active 3D systems provide the viewer with the original spatial resolution of the source. Also, the glasses do not filter the light. This results in unaltered video quality in comparison to non-3D systems. Furthermore, almost any television set is 3D capable as long as it can synchronize with shutter glasses. This drastically reduces production costs since screens can be reused. However, the temporal resolution of the source has to be at least 20 to 25 Hz in order to provide a smooth experience, with recommended frame rates of up to 60 Hz [11]. Also, active 3D displays and their glasses—in addition to being battery-driven—are relatively expensive compared to passive systems. As of today, active shutter systems are the prevalent technology for home entertainment.

4. CURRENT DEVELOPMENTS AND FUTURE OUTLOOK

Looking into the recent history of 3D vision and its applications, it is not surprising that predictions about the near future are more akin to fortune telling than specific forecasting. Trends and consumer opinions drastically changed over the last decade, especially in the field of 3D. Much of what the market currently offers is also limited by staggering advances in technology, especially with regards to autostereoscopic displays.

In 2006, the authors of [15] believed that typical TV consumption scenarios would "rule out the use of glasses to experience 3DTV. Active eyewear or other types of glasses would be cumbersome and 'unnatural' for users to wear continuously." Hopes were built on a widespread adoption of autostereoscopic displays—a trend that can not be seen today. Despite the appraisal of those devices by the research community, display manufacturers almost exclusively produce active systems for the 3DTV market. To illustrate the market situation, we surveyed a major Austrian online meta-search engine for consumer electronics.[3] The listing of 3D-capable LCD TV sets reveals that as of December 2012, from 364 available TV models, the majority of 241 (64%) are active stereoscopic, 132 (35%) passive and only 1 autostereoscopic.

In the mobile entertainment field however, autostereoscopic displays seem like the go-to solution: With the MOBILE3DTV project, major efforts were undertaken to understand both technical and social factors behind a mobile 3DTV transmission system, bridging the academic field with the consumer market [3]. In the context of this project, a mobile 3D streaming terminal device was developed in cooperation with industry partners such as Nokia. This lead to a number of usability studies, which identified common problems and pitfalls. It also ensures that future products are developed more efficiently.[4]

Regardless of the technology being employed, the success of 3D also depends on human factors. In the complete production chain, from recording to watching, humans are involved more than with traditional television or cinema. For classical multimedia consumption scenarios, various test procedures exist, which assess the Quality of Experience of a certain service [12, 13]. These subjective tests allow vendors to evaluate their products before bringing them to the market. They could for example measure the influence of certain encoding parameters or transmission schemes as well as different viewing scenarios. For 3D applications however, such recommendations do not exist yet, simply because the third dimension adds an unknown number of factors that influence QoE, ranging from visual discomfort to a feeling of immersion. It is an ongoing field of study to evaluate existing methodologies, adapt them to 3D, or even propose new procedures tailored to capture whatever added benefit 3D vision provides. Further subjective tests are needed to validate newly proposed guidelines [26, 7, 6].

[3] `http://geizhals.at/?cat=tvlcd&xf=33_15%2F38`, last accessed Dec. 7, 2012

[4] A list of the publications originating from the MOBILE3DTV project can be found on `http://sp.cs.tut.fi/mobile3dtv/results/`.

5. CONCLUSIONS

Although 3D technology now exists for more than a century, it still has not found widespread use. Only in the last decade, more and more cinema movies were shown in stereoscopy. Active shutter TVs are now available for prices of normal TV sets. The principles behind the human visual system as explained in Section 2 show that stereoscopic viewing scenarios can lead to visual fatigue due to the vergence–accommodation conflict. This mismatch can never be avoided entirely, but it has to be considered during production.

Several methods exist to capture 3D content, either using single or multiple cameras (possibly with depth analysis). These all have different benefits and drawbacks depending on the intended viewing context. Likewise, 3D display techniques have diverse applications, from the medical field to the consumer's living room. Each location influences the way in which 3D video is perceived, rendering the use of certain display devices inefficient at best or—if they are creating discomfort—unhealthy at worst.

In the future, development in 3D will be further driven by product vendors and the industry, in combination with the academia. In order to create successful products and reach critical mass, tests with real observers have to be undertaken. It is in the interest of any vendor to understand which quality factors are involved, and standardization efforts for 3D testing protocols are underway. Nonetheless, predictions about the success of 3D television or 3D cinema still have to be taken with a grain of salt.

6. REFERENCES

[1] ISO/IEC 13818-1:2007: Information technology – Generic coding of moving pictures and associated audio information: Systems, 2007.

[2] IMAX Annual Report, 2011. `http://www.imax.com/corporate/investors/financial-reports/`, last visited: 2012-10-12.

[3] Mobile 3DTV Content Delivery Optimization over DVB-H System – Final Public Summary. Technical report, MOBILE3DTV, 2011.

[4] ISO/IEC 14496-10:2012 — Information technology – Coding of audio-visual objects, 2012.

[5] ITU-T REC H.264: Advanced video coding for generic audiovisual services, 2012.

[6] W. Chen, J. Fournier, M. Barkowsky, and P. Le Callet. Quality of experience model for 3DTV. pages 82881P–82881P–9, 2012.

[7] W. Chen, J. Fournier, M. Barkowsky, P. Le Callet, et al. New requirements of subjective video quality assessment methodologies for 3DTV. 2010.

[8] J. E. Cutting and P. M. Vishton. *Handbook of perception and cognition*, volume 5, chapter Perceiving layout and knowing distances: The integration, relative potency, and contextual use of different information about depth, pages 69–117. Academic Press, San Diego, CA, 1995.

[9] H. Fuchs, M. Levoy, and S. Pizer. Interactive visualization of 3d medical data. *Computer*, 22(8):46 –51, 1989.

[10] D. Hoffman, A. Girshick, K. Akeley, and M. Banks. Vergence–accommodation conflicts hinder visual

performance and cause visual fatigue. *Journal of Vision*, 8(3), 2008.

[11] N. Hur, H. Lee, G. Lee, S. Lee, A. Gotchev, and S. Park. 3DTV broadcasting and distribution systems. *IEEE Transactions on Broadcasting*, 57(2):395–407, 2011.

[12] ITU-R BT.500-11. Methodology for the subjective assessment of the quality of television pictures, 2002.

[13] ITU-T P.910. Subjective video quality assessment methods for multimedia applications, 1999.

[14] S. Jolly, J. Zubrzycki, O. Grau, V. Vinayagamoorthy, R. Koch, B. Bartczak, J. Fournier, J.-C. Gicquel, R. Tanger, B. Barenburg, M. Murdoch, and J. Kluger. 3D Content Requirements & Initial Acquisition Work. Technical report, 3D4YOU, 2009.

[15] H. Kalva, L. Christodoulou, L. Mayron, O. Marques, and B. Furht. Challenges and opportunities in video coding for 3D TV. In *IEEE International Conference on Multimedia and Expo*, pages 1689–1692. IEEE, 2006.

[16] P. Kauff, M. Müller, F. Zilly, and S. Aijoscha. Requirements on post-production and formats conversion. Technical report, 3D4YOU, 2008.

[17] L. Kong, G. Jin, and X. Zhong. An autostereoscopic projecting system based on parallax barrier and lenticular sheets. In *2011 International Conference on Multimedia Technology (ICMT)*, pages 4816–4819, 2011.

[18] M. Lambooij, M. Fortuin, I. Heynderickx, and W. IJsselsteijn. Visual discomfort and visual fatigue of stereoscopic displays: a review. *Journal of Imaging Science*, 53(3):30201–1–30201–14, 2009.

[19] P. Lebreton, A. Raake, M. Barkowsky, and P. Le Callet. Evaluating depth perception of 3d stereoscopic videos. *IEEE Journal of Selected Topics in Signal Processing*, 99, 2011.

[20] J. Long and C. Siu. Randot stereoacuity does not accurately predict ability to perform two practical tests of depth perception at a near distance. *Optometry & Vision Science*, 82(10):912–915, 2005.

[21] G. Maguire. Populating Pandora's skies for the Academy Award-winning film Avatar (2009) for Industrial Light & Magic. Twentieth Century Fox Film Corporation, 2009.

[22] P. Merkle, K. Müller, and T. Wiegand. 3D video: acquisition, coding, and display. *IEEE Transactions on Consumer Electronics*, 56(2):946–950, 2010.

[23] L. Onural, T. Sikora, J. Ostermann, A. Smolic, M. R. Civanlar, and J. Watson. An Assessment of 3DTV Technologies. In *Proceedings of the NAB Broadcast Engineering Conference*, pages 456–467, 2006.

[24] L. Ostrin and A. Glasser. Accommodation measurements in a prepresbyopic and presbyopic population. *Journal of Cataract & Refractive Surgery*, 30(7):1435–1444, 2004.

[25] D. Park, T. G. Kim, C. Kim, and S. Kwak. A sync processor with noise robustness for 3DTV active shutter glasses. In *2010 International SoC Design Conference (ISOCC)*, pages 147 –149, 2010.

[26] F. Qi, T. Jiang, S. Ma, and D. Zhao. Quality of experience assessment for stereoscopic images. In *IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1712 –1715, may 2012.

[27] S. Reichelt, R. Häussler, G. Fütterer, and N. Leister. Depth cues in human visual perception and their realization in 3D displays. In *Three-Dimensional Imaging, Visualization, and Display*, volume 7690, 2010.

[28] W. Robitza, S. Buchinger, and H. Hlavacs. Impact of reduced quality encoding on object identification in stereoscopic video. In *EuroITV - 9th European Conference on Interactive TV*, Lisbon, Portugal, June 2011.

[29] C. Schor and C. Tyler. Spatio-temporal properties of Panum's fusional area. *Vision Research*, 21(5):683–692, 1981.

[30] K. Schreiber, D. Tweed, and C. Schor. The extended horopter: Quantifying retinal correspondence across changes of 3D eye position. *Journal of Vision*, 6(1), 2006.

[31] I. Sexton. Parallax barrier display systems. In *IEE Colloquium on Stereoscopic Television*, pages 5/1 –5/5, 1992.

[32] M. Solh and G. AlRegib. Hierarchical Hole-Filling For Depth-Based View Synthesis in FTV and 3D Video. *IEEE Journal of Selected Topics in Signal Processing*, 6(5):495–504, 2012.

[33] K. Ukai and P. Howarth. Visual fatigue caused by viewing stereoscopic motion images: Background, theories, and observations. *Displays*, 29(2):106–116, 2008.

[34] Z. Wang, R. Wang, S. Dong, W. Wu, L. Huo, and W. Gao. Depth Template Based 2D-to-3D Video Conversion and Coding System. In *2012 IEEE International Conference on Multimedia and Expo (ICME)*, pages 308–313. IEEE, 2012.

[35] J. Whyte, N. Bouchlaghem, A. Thorpe, and R. McCaffer. From CAD to virtual reality: modelling approaches, data exchange and interactive 3D building design tools. *Automation in Construction*, 10(1):43–55, 2000.

[36] T. Wiegand, G. Sullivan, G. Bjontegaard, and A. Luthra. Overview of the H. 264/AVC video coding standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 13(7):560–576, 2003.

[37] L. Zhang, C. Vazquez, and S. Knorr. 3D-TV Content Creation: Automatic 2D-to-3D Video Conversion. *IEEE Transactions on Broadcasting*, 57(2):372 –383, 2011.

[38] R. Zone. A window on space: Dual-band 3-D cameras of the 1950s. *Film History*, 16(3):216–228, 2004.