International Conference on Communication, Media, Technology and Design

ICCMTD
09-11 May 2012
Istanbul - Turkey

EVALUATION OF DEPTH CUES IN 3D SUBTITLING

Diekus González, Jordi Carrabina, Pilar Orero
Center for Accessibility and Ambient Intelligence of Catalonia, Universitat Autònoma de Barcelona, Spain

Recent explosion towards 3D content-enabled devices has brought to light many issues regarding user comfort, depth perception and accessibility in media. One important element that touches these areas is subtitling. We define a pipeline to produce extruded texts for evaluating 3D subtitling properties (using traditional 2D depth cues). An evaluation of current 3D tools was carried out, and the chosen platform allows us to create 3D typographies with high varied parameters (parallax, lightning). Produced subtitles were tested in full HD stereoscopic video clips. With the resulting subtitled clips, subjective perception tests were performed in order to evaluate the text's impact on ghosting and readability in different parallax values. The results are a set of rules to use when creating 3D subtitles that will provide: (i) better visual comfort, (ii) ghosting reduction (due to color gradient variations and volume), and (iii) an increment of negative parallax positioning with extrusion.

1. INTRODUCTION

As mammals with binocular vision we are able to extract information from our surroundings. This information is related to positioning and depth of objects in the environment and is used to help us move and interact with it (Sacks, 2010). This concept is called stereopsis and has been studied for decades; early research can be found in "Contributions to the Physiology of Vision" (Whaetstone, 1838). During the last ten years, we have seen stereoscopic 3D's reemergence in films (Mendiburu, 2009, pp. 6-7), though this is not a new format since the first time it emerged was in 1950 –but the technology was not ready to deliver quality 3D. Recently it seems 3D is increasingly capturing audiences, and some films such as *Avatar* (Cameron, 2009) proved to be a blockbuster (IMDb, 2012). The entertainment industry has been marketing and deploying 3D content to consumers with lineups of high definition players, 3D television sets and stereoscopic content available either through physical media or streaming. This transition from monoscopic to stereoscopic display changes the known workflows for production and post production of media, and many techniques and implementations are still not fully developed. This leads to issues in different areas such as: user comfort, quality of experience, usability, accessibility and depth perception, which can cause headaches and dizziness in some users, due to the difference in convergence and accommodation distance (Hoffman, et al., 2008). In this line of research we created a pipeline that allows the creation of volumetric 3D subtitles that combine a series of variables i.e. horizontal disparity and extrusion, and over those subtitles study if traditional 2D depth cues enhance depth perception. The paper is organized with section 2 where 2D visual cues are explained in order to perform tests in the generated subtitles in order to see if the depth perception can be enhanced. Section 3 explains current challenges with subtitles in 3D movies and some questions that arise when tackling the creation of volumetric texts. Section 4 gives an overview of the state of the art of tools available for working with stereoscopic media and 3D subtitling, (software and hardware). Section 5 goes step by step through the process of creation of the 3D subtitles. Section 6 describes the tests made to users and finally sections 7 section 8 expose results and conclusions.

2. VISUAL DEPTH CUES

Images formed in our retina are two dimensional. All the information regarding distance is inferred from the image and by the visual system. This information is gathered and reconstructed in the brain, and permits localization of objects the same way the auditory system can map the source of a sound. Depth can be inferred from three types of cues (Olshausen, 2007): oculomotor, visual binocular and visual monocular.

Oculomotor hints include accommodation which is when the lens of the eye changes size in order to focus an object in the retina; objects far away from us require a low concave shape versus a major concavity required for closer ones. Vergence is the other oculomotor prompt that refers to the movement of the eye when focusing distant objects (that tend to go in parallel lines) and close objects (that tend to bend to position inwards). These cues are related to the physiological processes of the eye.

Binocular cues consist in the horizontal disparity between slightly different images perceived by the left and right eye. Stereopsis is the process by which depth information is extracted from the scene composed in Panum's area. Panum's fusional area is where the two images perceived by both eyes fuse (Puell, 2006). This concept must not be confused with depth perception because we can perceive depth without binocular vision; nonetheless, it is the most advanced state of visual perception.

Monocular cues on the other hand can be obtained using kinetic vision, such as occlusion, size, perspective, parallax and definition of a terrain (Pipes, 2008). Occlusion indicates depth with superposition of objects. Size and perspective alone can also indicate the distance of an object (if the object is familiar and has an established concept in our brain the process is faster). Finally, parallax is one of the most important monoscopic cues because it relates with movement and different points of view. Parallax is the relative position of an object's image in a set of pictures (Mendiburu, 2009, p. 15).

These cues were incorporated in the creation of subtitles in order to try to blend them more naturally into stereoscopic media.

3. 3D SUBTITLE CHALLENGES

Disney was the first studio that announced (Geere, 2009) the inclusion of specially rendered subtitles for its movie *A Christmas Carol* (Zemeckis, 2009) and while more movies are being filmed using stereoscopic cameras –or converted from 2D to 3D-, little changes have been made to the way subtitles are displayed in 2D media –flat text with horizontal pixel offset. The main problem with this implementation is the change of focus between the text and the images when they are located in different depths, or regions. This situation requires a very fast and constant swapping of targets (subtitle/scene) while our eyes converge always to the same distance.

Subtitling stereoscopic media is a challenging task since special attention should be paid to avoiding interference with the content of the videos. Subtitles should be well aligned in the z-axis plane (or the depth plane, referred from now on just as z-plane) to avoid user discomfort. At present there are some 3D subtitle editors in the market, as Poliscript's 3DITOR (Screen, 2011). They allow the positioning of text in a 2D layer over the stereoscopic video and offers a horizontal disparity; nonetheless given the wider array of objects available to gaze upon (Hakkinen et al., 2010), text and other resources over the video must be used in a different way to achieve better results that reduce common problems (headaches, dizziness, tiredness, etc.) with current implementations of subtitling in 3D media.

4. STEREOSCOPIC 3D TOOLS

When working with 3D a distinction must be made between stereoscopic 3D (S3D) and 3D. S3D refers to the process in which slightly different images of the same object are sent to each eye in order to create in the brain depth perception. On the contrary, 3D can be related to the recreation of an open space where a character moves in a videogame (also called 2.5D). With this in mind, software and hardware evaluation was taken on board to select a set of tools that allowed us to produce and play content in stereoscopic form.

Software

a) Creation of 3D content

The main commercial application of S3D right now is in stereoscopic multimedia. Videos and games are rapidly adopting stereoscopic formats to enhance viewer´s immersion and this results in software adapting to produce this new format. 3D modeling software as Autodesk 3DS Max and Blender (through an addon) can produce stereoscopic 3D renders of objects by using horizontally off-set cameras. This allows for the assignment of different cameras to different scenes (points of view) and to export these renderings into an image that can be seen with compatible S3D hardware. The choice was Blender because it is a versatile open source and cross-platform modeling software that has a well documented API and supports a wide
arrange of free available addons that enhance the application capabilities -including the stereoscopic camera, as seen in figure 1.

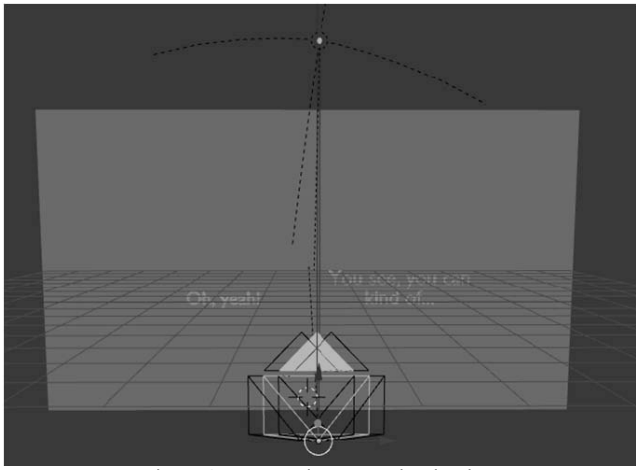

Fig. 1: Stereoscopic camera in Blender

b) Edition of 3D media content

For video editing, professional software is required and the three options in the market are Final Cut Pro (Apple, 2012), Adobe Premier (Adobe, 2012) and Sony Vegas Pro (Sony, 2012). As of today, Final Cut Pro (Apple, 2012) and Adobe Premier require the purchase of add-ins to work with stereoscopic 3D; as an example, Dashwood's Stereo3D Toolbox (Dashwood, 2012) works with these editors. Sony Vegas Pro has native stereoscopic video editing capabilities and can change the horizontal offset of a video or other element superposed in the frame, such as text, plain images or another video, all of which can have different disparities. This was the main reason why Sony Vegas was the chosen program (versions 10 and 11), because it gave the best flexibility out of the box to experiment with stereoscopic videos.

Hardware

Hardware-wise the options for working with stereoscopic content on a consumer level consist of two options. Passive polarized glasses or active shutter glasses (from now on referred to as 'passive' technology and 'active' technology). These technologies will vary in the display format used and the glasses that must be worn to get the stereoscopic effect. In this area NVIDIA has a proprietary active system called 3D Vision that consists of a 120Hz refresh rate display that is synchronized with active shutter glasses via an infrared receiver and radio frequency (NVIDIA, 2011). The main advantage is that a full HD image can be shown to each eye each timeframe. Other drivers like iZ3D and DDD TriDef (along with AMD HD3D) also allow shutter and DLP projectors to be used, and can also handle polarized (passive) technologies.

NVIDIA's 3D Vision infrared technologies were chosen because it is flexible and powerful enough to manage three screens surpassing high definition without complicated configurations, it has independent stereoscopic mode for an independent window, and has a centralized software and driver update system. However, this implies that the hardware platform is limited to active shutter glasses and the Windows operating system, since no 3D Vision drivers are available for other platform. Also, community support, examples, and availability of content exceed the ones of the other available options. For the display, Acer GD245HD 3DVision certified monitor was used. The rest of our configuration consists of an Intel Core i7 machine with 8 cores running at 2.80GHz and 8GB of RAM and a 64 bits processor.

5. 3D SUBTITLE CREATION PROCESS

To create the experimental subtitles a pipeline was defined (see figure 2) that consists of 3D rendering software (blender), image manipulation software (Paint.net), and video editing software (Vegas Pro). This pipeline was established thinking in giving the final generated text maximum detail using shade, hue, luminosity, extrusion, borders and texture. These different parameters would be used to give additional information to the viewer, the same way different colors in subtitles give information about the different characters speaking in TV shows.

Fig. 2: S3D subtitles creation process

The process starts setting a scene in the 3D modeling software with a camera and lighting. The camera must be set to stereoscopic mode

or a second camera must be added with a horizontal offset from the first one. Cameras must be linked if movement is involved. The distance between the lenses needs to be around 65mm to simulate human inter-ocular distance and the texts must be placed in a distance according to the size of the physical display where it will be shown. To achieve this in blender the parameters that need to be set are camera separation and zero parallax. Showing the stereo window and near-far planes is a good guide to know where the positioned text will appear (in z-plane), although this settings will vary through software applications. Text must be added and desired effects are applied. These effects range from diffuse and specular color, to offset, extrusion, depth, size and position. These are the parameters used to modify 2D depth cues. For example, the Catalan TV network TV3 uses occlusion (size and z-position) to mark a different semantic value between cast and titles in the opening credits of the theater recording for the play *Llits* (Danés, 2009).

After the model is generated, separate images (one per camera or lens) need to be generated. In our case, working with an active display, these images will be full HD. The images must be adapted to fit the format used by the content we are subtitling. Standard side by side (half and full) and top-bottom (half and full) are examples of how we need to make the images stack to match the 3D video. In our case side by side half is the format of the videos we are working with, so a conversion is necessary. To convert the images a reduction of size is in place. The resulting image must be the same resolution as the underlying video frame so it can be easily overlaid. Once the subtitles (generated images) are formatted in the same aspect (4:3 or 16:9) as the video they are ready to be incrusted in the video. The images are overlaid in the video edition software and the video is encoded (see figure 3). The process is long and tedious, since images are generated manually and there is no editor or way of seeing the video before it is encoded.

Fig. 3: Extruded subtitles overlaid in video

6. EXPERIMENTS

To test subtitles, four videos were generated with excerpts from the 3D Blu-ray version of *Tron:Legacy* (Kosinski, 2010) in order examine the cues we were testing. Four experiments were designed to measure different variables. The objective of the tests consisted in (i) the evaluation of horizontal disparity values (called the matrix test), (ii) the definition of a range of readability for the subtitles (animated offset test), and (iii) the evaluation of the effect of extrusion in the text (extrusion comparison test) and (iv) the determination of optimal position for text in the z-plane (z-plane comparison test).

Clips were tested with a group of 10 people. The test subjects ranged from 22 to 28 years. The test subjects are university students both male and female without technical knowledge on how stereoscopic 3D works.

The first test (matrix test) consisted in examining parallax values; accordingly, the first video consists of a matrix of nine texts overlaid in the clip with different parallax values (see figure 4). The user was asked to choose from the displayed texts which one matched better with a specific characteristic. These characteristics respond to the questions “which text seems closer to you”, “which text is deeper into the screen”, “which text has better readability”, and finally users were asked to answer “where” in the z-plane did they sense the text they chose as the one with best readability (‘inside of’, ‘outside of’ or ‘over’ the screen).

0	-0.02	-0.01
-0.03	0.03	0.01
-0.04	0.02	0.04

Fig. 4: Text positioned with parallax values of the table on the right

The matrix test was followed with the animated offset experiment. A video that contains a text was shown to the test subjects. This text had its horizontal offset value slowly animated from the most positive offset value that the editing software allows to the most negative one (see figure 5). The user was asked to stop the reproduction of the video when he/she felt comfortable with the readability of the texts. Stop times were logged and matched to the corresponding value of the animated horizontal disparity value. This allowed to corroborate the values obtained in the matrix test and to establish a range of horizontal offset values that can be considered “safe” for positioning of content generated for PC display viewing.

Fig. 5: The text's horizontal disparity is animated

The extrusion comparison test focused on text extrusion in our generated subtitles. It consisted in evaluation of extruded versus non-extruded subtitles, to examine if extrusion enhanced the depth perceived by users. Both rendered texts where made in blender, the only difference was the extrusion applied to the texts and the shading effect that the lightning of the scene produces in the extruded one. Both types of texts (pictured below in figure 6) were used to subtitle a one minute length clip at the same z-depth. These clips were shown to the test subjects and asked to rate how they perceived the subtitles, this is, at the same or different depth. The clips were shown twice to each user, and they were asked to respond to the question which subtitle they felt closer. This test included a question about ghosting, to evaluate if extrusion affected its appearance (figures 7 and 8 show the ghosting present in both types of subtitles).

Fig. 6: Different subtitles used in the clips

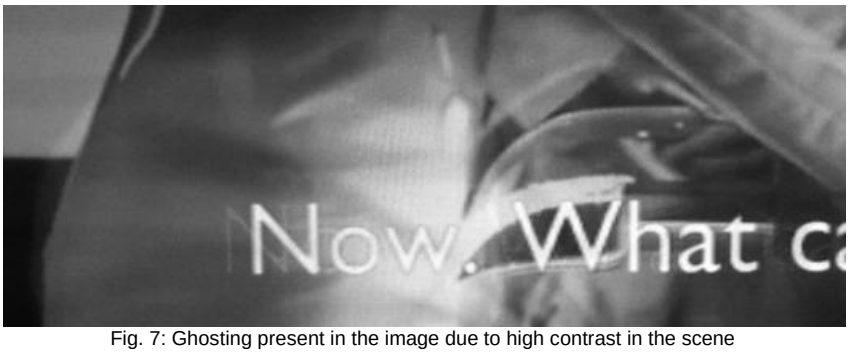

Fig. 7: Ghosting present in the image due to high contrast in the scene

Fig. 8: Reduced ghosting because of decreased contrast in the image

Finally, the z-plane comparison test played the same clip to the users varying the z-position of the extruded subtitles (with negative offset, at screen plane and with positive offset). Again, users were asked to write down their answers to questions on which one was more comfortable for them.

7. RESULTS

The matrix test determined that for users it is easier to perceive which objects are closer to them (in negative parallax) than the ones further away. Of the interviewed subjects, 73% were able to distinguish the most foremost texts (positioned at -0.04 and -0.03) displayed and 46%

could accurately tell which one was the one with the lowest parallax (-0.04). On the other hand, users had difficulties specifying the distance between texts with positive parallax; in this case only the 27% of them distinguished the right objects that were further away (0.03 and 0.04). The majority of responses (55%) indicated opposite disparity values (-0.02 and 0.02). Overall, users had a better sense of the distance between objects with negative parallax than the ones inside the screen (see figure 9). However, this did not imply better readability, since the users were able to distinguish which text was closer to them even if it was outside of Panum's area (which makes it impossible to focus and hard to read).

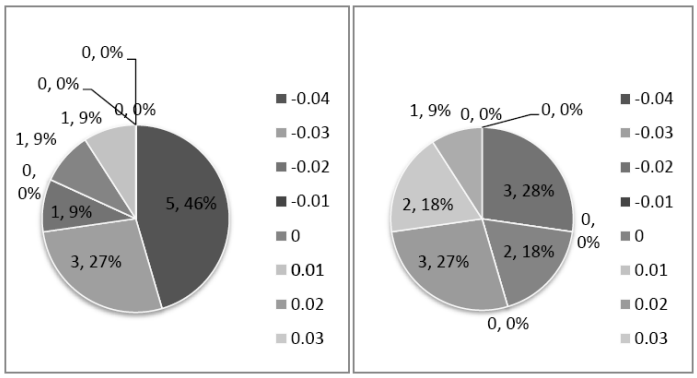

Fig. 9: Left: Most negative parallax values | Right: Most positive parallax values
Evaluation of users

The animated parallax test allowed the determination of a range of values that defined a readability area. By matching the times when the users stopped the video to their corresponding horizontal offset values, we were able first to confirm the results from the matrix experiment regarding which of the nine text objects had better readability and secondly, we notice that our range [-0.03, 0.2] includes the industry's accepted 1% empiric rule (2% for general content creation) (Mendiburu, 2009, p. 74) for small screen content production.

In the extrusion comparison test, it was found that extrusion (as a set of 2D depth cues) in the generated texts did make those subtitles appear to have less parallax (to be more out of the display) that non-extruded ones. Extrusion is a powerful and simple way to force objects nearer without coming too close to the readability limits for them. It also showed that due to the reduction of the contrast produced by the shading of the text models, ghosting was reduced (see figure 8).

Finally, based on the answers we got from oral interviews and written questionnaires with the z-plane comparison test, we determined that all the subjects preferred the extruded texts over the 2D with horizontal offset, since they "blend" with the 3D picture. They expressed that the volume associated with the texts made the subtitles feel like a part of the scene instead of just an overlaid layer.

8. CONCLUSIONS

In this work we have created stereoscopic subtitles that use depth cues to improve current implementations of subtitling in S3D videos. The generated texts provide more visual comfort, decrease ghosting and enhance the perception of negative parallax to the user. This is achieved applying basic monoscopic principles to 3D models of the subtitle.

A set of recommendations when creating subtitles for stereoscopic media include (i) the positioning of the texts close to the screen plane to ensure the best readability, (ii) the utilization of lightning, shades and color gradients to reduce the contrast between the scene and the subtitle and (iii) the creation of an illusion of lower negative parallax to bring objects more outside of the screen while staying at a safe parallax value.

The process to create these subtitles is composed by several software applications that are used to generate extruded texts. Nonetheless, this process can be time consuming due to its manual labor consisting of 3D rendering, image, and video edition. This process can produce subtitles that are enhanced by monoscopic depth cues, proving that colors, shades, size and occlusion can change the perception of depth of an object without changing its horizontal disparity.

During the research it was not difficult to find out that the tools available for working with stereoscopic 3D are very immature (mostly because of the lack of an integrated solution to work with this kind of format) and that there is a lot of investigation to be done until S3D is fully understood. The three major software applications that handle video editing need plug-ins to fully work with S3D and hardware platform choices can limit the viewing technologies available.

The technology is far from perfect and users expressed their concerns for several known issues with it, like darkness of image present in active displays because of the glasses and the uncomfortable design of the glasses, which led several of the users to take them off between tests. Three of them experienced light headaches or dizziness. A comment that was common between them is that 3D depth perception is very tricky, since it depends on the position (viewing angle), lightning in the scene, technology used and personal eye health.

Future work is needed to make this process easier and more accessible to filmmakers and other content producers. A process to generate these subtitles in an independent format also needs to be considered. It is not acceptable to have to encode the video with the subtitles merged into it and unthinkable for distribution and storage of videos. This would imply having to encode different and repeated files for different subtitle languages.

Having evaluated how 2D depth cues affect stereoscopic 3D subtitles, we believe that a more in-depth psychological study is required to document the subtitle's effect in full length movies. Perception and loss of visual information (using eye-trackers (Hakkinen et al., 2010) and questionnaires) must be measured and compared to similar studies in 2D. Also, the automation of the process and a framework that allows the positioning of content over the video, using parameters (cues explained in this paper) to generate the subtitles must be developed –and its uses studied also, these are, the impact of shades and hues in the generated subtitles. The automation would allow interfacing the core process with inputs such as subtitle files (TXT, EVU, SRT, SUB, STL) that can be parsed to generate real time rendering and other inputs like voice (direct application would not be for movies, but the generation of S3D text from voice recognition can be interesting to other live or streaming mediums). Related to this, the adaptation of an existing file format to accommodate new parameters related to depth must be examined, taking advantage of unused conventions in these existing formats.

Probably the most interesting area to explore in the future after the positioning of texts in different depths is the study of semantic and hierarchic meaning of those objects based on their parallax. This, combined with the interaction with other objects during movement,

and in combination with depth maps, might facilitate the way putting text “inside” the video frame is done.

9. ACKNOWLEDGEMENTS

This research is supported by the grant from the Spanish Ministry of Science and Innovation FFI2009-08027, Subtitling for the Deaf and Hard of Hearing and Audio Description: objective tests and future plans, and also by the Catalan Government funds 2009SGR700. It is also funded by the Catalan Government scholarship ECO/2060/2011 (DOGC num. 5951 29.8.2011).

10. REFERENCES

Adobe Systems Incorporated. (2012). *Video Editing Software | Adobe Premier Pro CS5.5*. Retrieved March 20, 2012, from Adobe.com: http://www.adobe.com/products/premiere.html

Apple Inc. (2012). *Final Cut Pro X – Top Features*. Retrieved March 20, 2012, from Apple.com: http://www.apple.com/finalcutpro/top-features/

Dashwood Cinema Solutions. (2012). *Stereo3D Toolbox Plugin Suite*. Retrieved March 20, 2012, from dashwood3D.com: http://www.dashwood3d.com/stereo3dtoolbox.php

Geere, D. (2009). *Disney planning first 3D subtitles*. Retrieved December 11, 2011, from Pocket-lint: http://www.pocket-lint.com/news/29871/disney-plans-first-3d-subtitles

Hakkinen, J. Kawai, T. Takatalo, J. Mitsuya, R. Nyman, G. (2010). *What do people look when they watch stereoscopic movies?*. Electronic Imaging: Stereoscopic Displays & Applications XXI, Proc. SPIE, Vol. 7524

Hoffman, D., Girshick, A., Akeley, K., & Banks, M. (2008). *Vergence–accommodation conflicts hinder visual performance and cause visual fatigue*. *Journal of Vision* .

IMDb. (2012). *All-Time Box Office: World-wide.* Retrieved March 22, 2012, from IMDb: http://www.imdb.com/boxoffice/alltimegross?region=world-wide

Mendiburu, B. (2009). *3D Movie Making: Stereoscopic Digital Cinema from Script to Screen*. Massachusetts: Focal Press.

NVIDIA. (2011). *3D Vision Surround*. Retrieved December 11, 2011, from NVIDIA: http://www.nvidia.co.uk/object/3d-vision-surround-technology-uk.html

Olshausen, B. (2007, May 23). *Psych 129 - Sensory Processes.* Retrieved December 11, 2011, from Berkerley University of California: http://redwood.berkeley.edu/bruno/psc129/lecture-notes/depth.html

Pipes, A. (2008). *Foundations of Art + Design*. London: Laurence King Publishing.

Screen Subtitling Systems. (2011). *Subtitling for Stereographic Media*. Retrieved March 20, 2012, from subtitling.com: http://www.screen.subtitling.com/downloads/Subtitling%20for%20Stereographic%20Media.pdf

Puel, M. (2006). *Óptica fisiológica*. Madrid: Editorial Complutense S.A.

Sacks, O. (2010). *The Mind's Eye*. New York: Alfred A. Knopf

Sony Creative Software Inc. (2012). *Vegas Pro 11 Descripción general*. Retrieved March 20, 2012 from Sonycreativesoftware.com: http://www.sonycreativesoftware.com/vegaspro

Whaetstone, C. (1838). *Contributions to the Physiology of Vision – Part the First*. Retrieved December 13, 2011, from stereoscopy.com - The Library: http://www.stereoscopy.com/library/wheatstone-paper1838.html

11. FILMOGRAPHY

Cameron, J. (Director). (2009). *Avatar* [Motion Picture].

Danés, L. (Director). (2009). *Llits 3D* [Motion Picture].

Kosinski, J. (Director). (2010). *TRON: Legacy* [Motion Picture].

Zemeckis, R. (Director). (2009). *A Christmas Carol* [Motion Picture].