

This material is published in the open archive of Mid Sweden University DIVA <http://miun.diva-portal.org> to ensure timely dissemination of scholarly and technical work.

Copyright and all rights therein are retained by authors or by other copyright holders. All persons copying this information are expected to adhere to the terms and constraints invoked by each author's copyright. In most cases, these works may not be reposted without the explicit permission of the copyright holder.

Schwarz, S.; Olsson, R.; Sjöström, M., "Depth Sensing for 3DTV: A Survey," *IEEE MultiMedia*, 20(4), 10-17, 2013.

[DOI: 10.1109/MMUL.2013.53](https://doi.org/10.1109/MMUL.2013.53)

© 2013 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.

# Depth Sensing for 3DTV: A Survey

Sebastian Schwarz, Roger Olsson, and Mårten Sjöström

Mid Sweden University, Sundsvall, Sweden

Three-dimensional television (3DTV) is a hot topic for multimedia researchers, producers and consumers [1]. While traditional, glass-based stereoscopic 3D is steadily making its way from the movie theaters into our living rooms, novel 3D display technologies are emerging. Autostereoscopic multiview displays are set out to remove the drawbacks of stereoscopic representations and present a more immersive 3DTV viewing experience. Instead of just one stereo view pair, such displays provide up to 128 separate views, enabling a realistic “look around” feeling without the necessity of disturbing glasses [2].

Since it is not practical to transmit a large number of views, new video representation formats have been introduced. In particular, the Multiview-Video plus Depth (MVD) format [3] considers a small set of views (e.g., 2 or 3) together with corresponding scene depth information. A typical view with its corresponding depth map is shown in Fig. 1(a). This depth map provides the necessary data to generate arbitrary views using Depth Image Based Rendering (DIBR) [4] as shown in Fig. 1(b). With depth information and projection matrix  $\mathbf{P}$ , containing the projective relationship between original and virtual camera view, we can project any pixel  $\mathbf{b}$  from the original video frame onto position  $\mathbf{b}'$  for a new, virtual video view. Performing this projection for every pixel creates a virtual point of view. In theory, it is possible to create any arbitrary view from a single frame and its depth map. However, in practice, parts of the virtual view are not visible in the original frame. Therefore, virtual views are synthesized from a combination of input frames. Fig. 1(c) shows how DIBR can be used to feed a 5-view multiview display from two inputs.

The quality of the virtual views, and hence the quality of the 3DTV experience, relies heavily on the depth map accuracy. Unfortunately, acquiring depth is not always straight-forward. While for computer-generated imagery (CGI) the depth information is available directly from the modeling software, depth

map acquisition for natural scenes is more tricky.

It is noted that depth maps have a piece-wise smooth value distribution. They consist mainly of large uniform and smooth areas, corresponding to coherent objects, and sharp value transitions, corresponding to jumps in depth between objects. For the application of DIBR two facts are important: First, misaligned depth transitions will lead to rendering artifacts in the virtual view. And second, missing depth values will lead to holes in the virtual view.

In this article, we introduce the interested reader to the field of depth map capture for 3DTV applications. Three main depth sensing concepts are presented: Passive stereo analysis, active stereo analysis, also known as structural lighting, and dedicated depth sensors, i.e. Time-of-Flight cameras. These three concepts and their combinations form the majority of 3DTV depth sensing approaches. We address their individual pros and cons and put them within respect to each other. Thus, this article serves as guideline for aspiring 3DTV content creators, as well as reference for experienced professionals.

## PASSIVE STEREOVISION ANALYSIS

The most common and most established depth sensing concept for 3DTV is passive stereovision analysis [5]. Computer vision algorithms look for corresponding image features in two or more camera views. Considering an array of cameras at different positions, a 3D scene is projected slightly different onto each image plane. The difference between two corresponding points is called *disparity* and gives a measure of depth. This concept is illustrated in Fig. 2(a), where the offsets  $u_1$  and  $u_2$  yield disparity  $\delta = u_1 - u_2$ . With baseline  $B$ , the distance between the cameras, and focal length  $f$  we acquire depth  $z$  with the basic intercept theorem,

$$z(\delta) = \frac{B \cdot f}{\delta}. \quad (1)$$

As shown in Fig. 2(b), close objects have a larger disparity than more distant parts of the scenery. This approach has the advantage that we generate depth maps directly from multiview video. No additional equipment is needed to capture depth. Furthermore, Stereo3D is already an established content format. All major 3D movie releases of the last decade are available in Stereo3D. Depth information gained by passive stereovision analysis makes this content fit for autostereoscopic 3DTV, while still supporting conventional Stereo3D.

Anyhow, there are a few things one should consider before creating MVD content with passive stereovision. First of all, stereovision analysis is highly complex and may be challenging to run in real-time on platforms with limited processing capabilities. Although the first multi-camera real-time applications have already been announced [6], depth estimation is not well suited for live production at this time.

Second, stereovision algorithms can only create depth for points visible in both views. If parts of the scenery are occluded in one view, it is impossible to establish any correspondence. This leads to “depth shadows” around foreground objects. Areas without depth information are shown in Fig. 2(c). The size  $S$  of the shadow in each view is based on the stereo baseline  $B$ , the foreground distance  $D$ , and the distance  $G$  between foreground and background,

$$S = \frac{G}{D} \cdot B. \quad (2)$$

Third, capturing cameras have a finite pixel resolution. If 3D points are projected on the same pixel coordinates, we get a quantization error in depth. This error  $\Delta z$  is related to the camera baseline  $B$ , focal length  $f$  and the capturing pixel width  $\gamma_p$  [5].

$$\Delta z(z) = \frac{z^2}{B \cdot f} \cdot \gamma_p \quad (3)$$

Finally, stereovision analysis relies on detecting common points in both views. This is usually done by a combination of feature and area detectors. Feature detectors look for corners, edges or distinctive lines for robust but sparse depth information. Area detec-

tors consider windows around each pixel to determine similarity between views if no features could be detected. Therefore, it is highly important that all capturing cameras are precisely matched and color corrected. However, stereovision analysis still fails if there is not enough or not distinctive enough information available in the actual scene. Low textured areas, e.g. a white wall, or repetitive structures, e.g. tiling, result in ambiguous correspondences which will yield erroneous depth estimates.

The first two points, complexity and depth shadows, are best addressed in post production, where more relaxed time constraints are imposed. Sophisticated image processing and stereovision analysis algorithms [5] generate high quality depth maps, including filling solutions for “depth shadows”. The depth quantization error can be addressed by adapting your camera setup to the scene requirements, i.e. narrow baseline for close content, wide baseline for distant content. Unfortunately, low or repetitive textures are not so easily handled. However, active stereovision can provide the solution for this problem.

## ACTIVE STEREOVISION ANALYSIS

For active stereovision analysis we replace one “passive” camera with an “active” light source, projecting a predefined structure, e.g. a line grid, on our scene. This projection is usually done in a part of the light spectrum not visible to the human visual system (HVS), i.e. infrared (IR), so the actual content is not disturbed. The geometry of our scene distorts the light structure. We can compare the distortion to the original pattern with an IR camera and get the depth information based on correspondence matching and triangulation similar to passive stereovision analysis.

While the projected light pattern simplifies correspondence matching for low or repetitive texture, we introduce a new constraint, the viability of the projected light pattern. First, the projection has to be powerful enough. If the projector is too weak, the target area too far away, or if there is strong IR background illumination, i.e. sunlight, the pattern will be too weak for detection. Second, the target area has to be within the feasible region for the used light pattern. If the target is too close, the light pattern might have overlaps. If the target is too far, the distance between distinctive points of the light pattern might

be too big for coherent depth estimation. Altogether active stereovision analysis is limited to indoor application within a predefined depth range.

Another important point is to capture the corresponding multiview video. Unlike passive stereovision, active stereovision provides only depth and additional video cameras are required to generate MVD content. One solution is to combine IR projector, IR camera and video camera all in one device. In recent years this approach gained a lot of interest due to the introduction of the Microsoft Kinect sensor, shown in Fig. 3(a), with many fascinating applications for 3DTV [7]. However, since we still have two different viewpoints, the “depth shadow” problem still exists, as the Kinect depth map in Fig. 3(c) clearly shows. This problem can be reduced with dedicated range sensors.

## DEPTH FROM DEDICATED SENSORS

The last depth sensing approach discussed in this article is the use of dedicated range sensors. Such sensors measure the time-of-flight (ToF) of a light beam. There are two different types of ToF sensors: *Pulse runtime sensors*, shown in Fig. 4(a), where a pulsed wave is sent out and a clock measures the time that has passed until the reflected signal is received. Such sensors deliver depth accuracy between 10-20mm for distances of up to a few hundred meters. However, they have a low temporal resolution due to the pulsing scheme, which makes them unsuitable for 3DTV content creation. The other type, *continuous wave sensors*, shown in Fig. 4(b), measures the phase shift between a modulated wave signal and its reflection. The sensor sends out a cosine modulated signal  $s(t)$ . Based on the standard equations for light propagation, we can determine the depth  $z$  of an object based from the phase shift  $\Phi$  of the received reflected signal  $r(t)$  [8]:

$$z(\Phi) = \Phi \cdot \frac{c}{2\omega} \quad (4)$$

For reliable ToF readings we have to make sure that the intensity of the received signal is strong enough. This intensity is called *active brightness* and shown in Fig. 4(d). Areas with lower active brightness are equal to a larger depth error, since the sensor does not get enough information to determine the phase shift[9]. The

active brightness depends on the optical power and travelling distance of the sent signal as well as the sensor exposure time.

Continuous wave ToF sensors are predestined for real-time 3DTV capture. They have a depth accuracy of around ten millimeters and a maximum distance of about ten meters and are well suited for real-time 3DTV capture. Also, they can capture up to 60 depth maps per second, without any time intensive correspondence matching in post production. Unlike passive stereovision analysis, they deliver reliable and accurate depth information in low or repetitively textured areas and suffer less from shadowing, as shown in Fig. 4(c).

Similar to active stereo, current ToF sensors still require additional video cameras to capture MVD content. Yet, recent developments show the possibility of video and depth from the same sensor. In 2012 Samsung presented a combined color video plus ToF chip, capable of capturing 1920x1080 (Full HD) video and 480x270 depth values [10]. However, this chip still shows the drawback of ToF sensors: The limited spatial resolution compared to modern video cameras. Due to the capturing architecture and the need for high active brightness, the size of each capturing pixel element is rather large [8]. Therefore, upscaling algorithms are required to match the multiview video resolution. ToF depth upscaling is a very active research field, with many different approaches. Most solutions share the common idea to utilize texture information from the video cameras for the depth upscaling process. Matching depth and video also guarantees the important correspondence between object borders and transitions in depth. The EU FP7 project SCENE has investigated many of these approaches and summarized solutions for texture-guided real-time ToF upscaling in 25 frames per second (fps) or faster [11].

## COMPARISON & CONCLUSION

In this article, we have reviewed three depth sensing approaches for 3DTV. The characteristics of each approach is summarized in Table I and further discussed below.

First, passive stereovision can create “ready-to-use” MVD content. Active stereovision and ToF depth require additional video sources. Not only do you

need additional cameras, but you also have to make sure that video and depth are matched on the same viewing angle. This could either be done by projection, thus requiring precise camera calibration and creating "shadowing" artifacts, or by optical beam-splitters, thus reducing the available light by one f-stop. Further solutions exist for ToF, e.g. the combined Samsung video plus depth sensor mentioned above and shared camera optics for video and ToF sensors [11].

Passive stereovision is also more flexible in terms of environmental limitations. Active stereovision solutions perform badly in outdoor scenarios with lots of background IR radiation, i.e. sunlight. ToF sensors perform slightly better, but reach their full potential in controlled indoor lighting scenarios, i.e. TV studios.

In terms of real-time capabilities, passive stereovision has some drawbacks. If you aim at live production, e.g. sport events or news, you might run into timing problems with passive stereovision analysis. Active stereo and ToF are the better choice here.

Regarding your actual content, you have to choose carefully for passive stereo, otherwise you run into problems. Occluded parts of your scenery will lead to "shadows" in the depth map and low or repetitive textures will create erroneous depth readings. ToF sensors and active stereovision have no problems with different textures. In case of ToF, integrated video plus depth solutions also eliminate the shadowing problem. The key problem of current ToF sensors, the limited spatial resolution, can be solved with texture guided upscaling.

In terms of temporal resolutions, all approaches are limited by the frame rate of the capturing cameras. Typical frame rates for video are 25fps, current ToF sensors support up to 60fps. Please note that this is not a limitation per se for stereovision. You can always chose cameras with higher frame rates. Anyhow, ToF sensors are already capable of High Frame Rate (HFR) capture.

The final point, depth sensing range, is especially interesting. In passive stereovision you can adapt to your scenery by adjusting the camera baseline and have no theoretical limit. For active stereovision you have to make sure that your light pattern is distinctive enough. Therefore you have a lower and upper limit, e.g. 1.2-4m for the Microsoft Kinect solution. For

ToF capture you have to make sure to get enough reflected light on your sensor, so you depend on the optical power of your light emitter. A typical range for ToF capture is up to 10m [12].

Summarizing, we can conclude that, outside the controlled environment of TV studios, you will have to reside with passive stereovision analysis and all its shortcomings, at least for the time being. Active stereovision analysis can overcome some of those limitations, but only in controlled environments. However, ToF sensors are the better choice for real-time 3DTV content in a controlled environment. They deliver the more accurate depth readings with the least limitations on the capturing scenario.

Looking into the future, it is hard to predict which depth sensing approach will become the number one choice. Most probably, there won't be just one solution. Combinations of passive stereo and ToF appear promising, especially since conventional Stereo3D are already established. New research trends of video + depth with shared optics, or even on one single chip will be interesting to follow.

However, these solutions are still in the future. With this article you now have a good understanding of the requirements and challenges for 3DTV content creation can start producing content tomorrow. We wish you the best of luck!

## ACKNOWLEDGMENT

The authors would like to thank John Doublestein, Craig Scheuermann and Ying-Chih Chen for providing the Andy Rig used in Fig. 1-3.

This work has been supported by grant 2009/0264 of the KK Foundation, Sweden, by grant 00156702 of the EU European Regional Development Fund, Mellersta Norrland, Sweden, and by grant 00155148 of Länsstyrelsen Västernorrland, Sweden.

## REFERENCES

- [1] R. Bajcsy, R. Yang, P. Zanuttigh and C. Zhang, "3D Imaging Techniques and Multimedia Applications," *IEEE Multimedia*, vol. 20, no. 1, pp. 14-16, 2013.
- [2] H. Urey, K. Chellappan, E. Erden and P. Surman, "State of the Art in Stereoscopic and Autostereoscopic Displays," *Proceedings of the*

- IEEE*, vol. 99, no. 4, pp. 540-555, 2011.
- [3] K. Müller, P. Merkle and T. Wiegand, "3-D Video Representation using Depth Maps," *Proceedings of the IEEE*, vol. 99, no. 4, pp. 643-656, 2011.
- [4] C. Fehn, "Depth-Image-Based Rendering (DIBR), Compression, and Transmission for a Flexible Approach on 3DTV," in *SPIE Stereoscopic Displays and Virtual Reality Systems XI*, San Jose, CA, USA, 2004.
- [5] D. Scharstein, R. Szeliski and R. Zabih, "A Taxonomy and Evaluation of Dense Two-frame Stereo Correspondence Algorithms," *International Journal of Computer Vision*, vol. 47, no. 1-3, pp. 7-42, 2002.
- [6] P. T. Kovacs and F. Zilly, "3D capturing using multi-camera rigs, real-time depth estimation and depth-based content creation for multi-view and light-field auto-stereoscopic displays," in *ACM SIGGRAPH 2012 Emerging Technologies*, Los Angeles, CA, USA, 2012.
- [7] Z. Zhang, "Microsoft Kinect Sensor and its Effect," *IEEE Multimedia*, vol. 19, no. 2, pp. 4-10, 2012.
- [8] R. Lange and P. Seitz, "Solid-State Time-of-Flight Range Camera," *IEEE Journal of Quantum Electronics*, vol. 37, no. 5, pp. 390-397, 2001.
- [9] M. Frank, M. Plaue, H. Rapp, U. Köthe and F. Jähne, "Theoretical and Experimental Error Analysis of Continuous-Wave Time-Of-Flight Range Cameras," *SPIE Optical Engineering*, vol. 48, no. 1, p. 013602, 2009.
- [10] J.-S. Kim, B. Kang, J. D. K. Kim, K. Lee, C.-Y. Kim and K. Kinam Kim, "A 1920x1080 3.65  $\mu\text{m}$ -pixel 2D/3D image sensor with split and binning pixel structure in 0.11 $\mu\text{m}$  standard CMOS," in *IEEE International Solid-State Circuits Conference (ISSCC)*, San Francisco, CA, USA, 2012.
- [11] SCENE, "Novel scene representations for richer networked media," 2013. [Online]. Available: <http://www.3d-scene.eu/>.
- [12] S. Foix, G. Alenyà and C. & Torras, "Lock-in Time-of-Flight (ToF) Cameras: A Survey," *IEEE Sensors*, vol. 11, no. 9, pp. 1971-1926, 2011.

**Sebastian Schwarz** is a doctoral student in computer and system science at the Mid Sweden University, Sundsvall, Sweden. His research interests include Time-of-Flight sensors and depth map upsampling. Contact him at [sebastian.schwarz@miun.se](mailto:sebastian.schwarz@miun.se)

**Roger Olsson** is a researcher at the Mid Sweden University, Sundsvall, Sweden. His research interests include plenoptic computational imaging and modeling. Olsson has a PhD in telecommunications from Mid Sweden University. Contact him at [roger.olsson@miun.se](mailto:roger.olsson@miun.se)

**Mårten Sjöström** is a professor in signal processing at the Department of Information and Communication Systems at the Mid Sweden University, Sundsvall, Sweden. His current research interests are within system modeling and identification, as well as 2D and 3D image and video processing. Sjöström has a PhD in Modeling of Nonlinear Systems from EPFL, Switzerland. Contact him at [marten.sjostrom@miun.se](mailto:marten.sjostrom@miun.se)

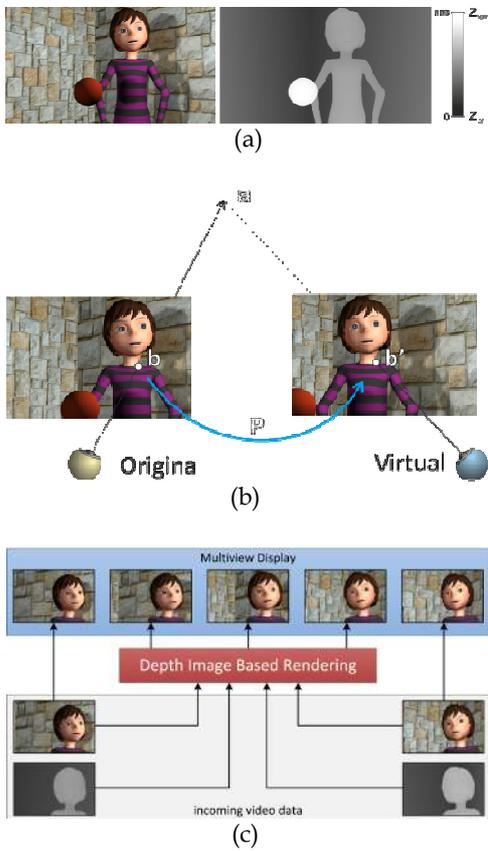


Fig. 1. Example of a video frame and its corresponding 8Bit depth map (a). Point-to-point projection between original and virtual view (b). DIBR multiview generation from two input streams (c).

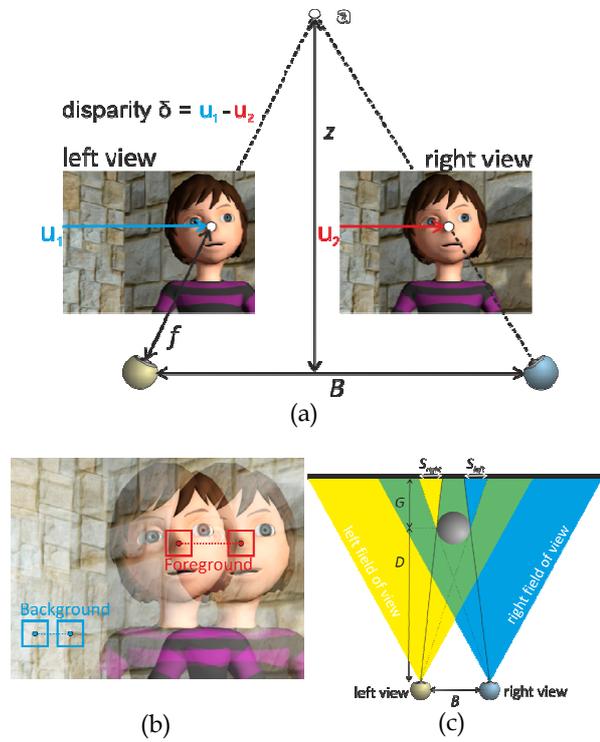


Fig. 2. The relation of depth and disparity between two views (a). Closer objects have more disparity than background objects (b). Depth shadowing due to background occlusion (c).

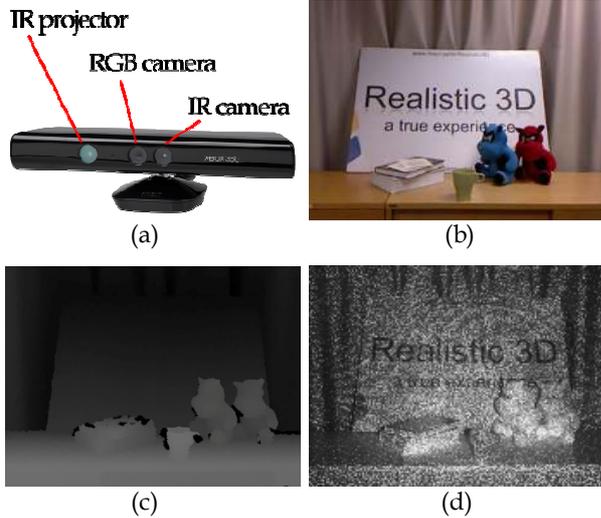


Fig. 3. The Microsoft Kinect structural lighting solution (a): Capturing color video (b) and depth (c), based on the IR pattern shown in (d).

TABLE 1  
DEPTH SENSING APPROACH COMPARISON

	Passive Stereo	Active Stereo	ToF
Video source provided	●	○	○/●
Requires controlled environment	○	●	●
Real-time capability	○/●	●	●
Occlusion/shadowing problem	●	●	○
Low/repetitive texture problem	●	○	○
Upscaling required	○	○/●	●
Frame rate	Typical 25 fps		≤ 60fps
Typical depth sensing range	No limit	1.2 - 4m	≤ 10m

● = yes | ○ = no | ○/● = sometimes

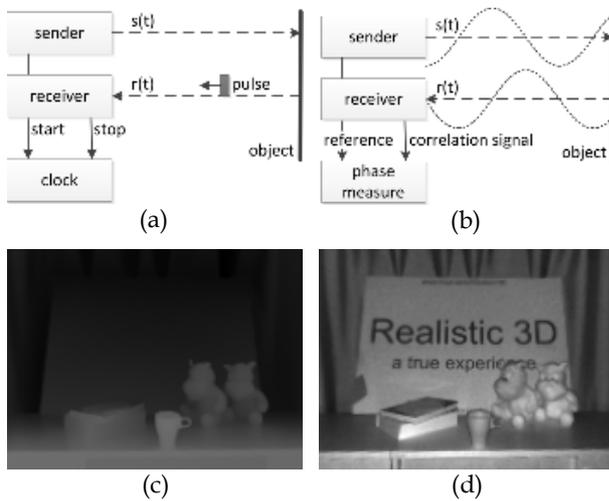


Fig. 4. Pulse runtime (a) and continuous wave (b) ToF principle. Continuous wave ToF depth (c) and active brightness (d) signal.