Wenjun Zeng
*University of Missouri-Columbia*

# Converting 2D Video to 3D: An Efficient Path to a 3D Experience

**Xun Cao**
*Tsinghua University*

**Alan C. Bovik**
*University of Texas at Austin*

**Yao Wang**
*Polytechnic Institute of New York University*

**Qionghai Dai**
*Tsinghua University*

Wide-scale deployment of 3D video technologies continues to experience rapid growth in such high-visibility areas as cinema, TV, and mobile devices. Of course, the visualization of 3D videos is actually a 4D experience, because three spatial dimensions are perceived as the video changes over the fourth dimension of time. However, because it's common to describe these videos as simply ''3D,'' we shall do the same, and understand that the time dimension is being ignored. So why is 3D suddenly so popular? For many, watching a 3D video allows for a highly realistic and immersive perception of dynamic scenes, with more deeply engaging experiences, as compared to traditional 2D video. This, coupled with great advances in 3D technologies and the appearance of the most successful movie of all time in vivid 3D (*Avatar*), has apparently put 3D video production on the map for good.

A typical 3D video system includes the steps of content creation, coding and transmission, reconstruction, and display. Three-dimensional processing and display technologies have dramatically advanced—for example, glasses-free (auto-stereoscopic) 3D displays are commercially available on small-screen mobile devices. Large-format 3DTV cable receivers and display systems are now commonly found in the home, which is regarded as the biggest development in home entertainment since high-definition TV. Yet, aside from the wide availability of theatrical 3D productions, there's still surprisingly little 3D content (3D image and video) in view of the tremendous number of 3D displays that have been produced and sold. The lack of good-quality 3D content has become a severe bottleneck to the growth of the 3D industry.

There is, of course, an enormous amount of high-quality 2D video content available, and it's alluring to think that this body of artistic and entertainment data might be converted into 3D formats of similar high quality. Our goal in this article is to introduce and explain emerging theories and algorithms for 2D-to-3D video conversion that seek to convert ordinary 2D videos into 3D.

## Editor's Note

The 3D experience is poised to be the future of entertainment. The path to that future, however, is bumpy. Despite widely available 3D display devices, a lack of 3D content impedes the 3D industry's rapid growth. Here the authors discuss some common 3D content-creation processes and the key elements that affect quality, and present an approach that converts 2D content to 3D content as an efficient path to a 3D multimedia experience.

*—Wenjun Zeng*

## 3D content generation

An obvious approach toward ameliorating the current shortage of good-quality 3D video content is to explore effective 3D content-creation methods. Most contemporary 3D content (for example, commercials and movies) is created using commercial modeling software, such as Autodesk's 3ds Max and Maya. The basic mode of operation is to first build the target object's 3D structure (a graphical mesh representation), then render the meshes using surface texture and lighting models to create realistic appearances (see the middle of Figure 1a). Computer graphics animation is a successful example of this category of 3D content production (for example, *Shrek 3*). The advantage of modeling software is that the synthesized 3D models are accurate and easily manipulated. The drawback is that artificial 3D models are quite time consuming to create, particularly when highly realistic or otherwise complex scenes are required.



1070-986X/11/$26.00 © 2011 IEEE

Published by the IEEE Computer Society

Beyond graphical modeling, 3D capture devices have also greatly progressed in recent years. Stereoscopic cameras and time-of-flight (TOF) cameras, both consumer and professional, are increasingly common and used in a variety of applications. Similar to the binocular vision system found in humans and many other animals, a stereoscopic camera setup with two cameras, located in nearly the same place and pointing in almost the same direction, capture the same visual scene (see the top of Figure 1a). In this way, the video is captured from two viewpoints with slightly different perspectives, yielding slightly different projections onto the cameras' recording apparatus. The left and right views are then integrated in specific visual centers of the brain to form the 3D depth experience, which is called the *cyclopean image*. This 3D experience is quite different from viewing the 2D images separately. The sense of depth, 3D perspective, and 3D shape is profoundly amplified beyond the 3D experience delivered by other visual cues, such as motion, size, surface-luminance gradients, and so on.

Alternatively, TOF 3D cameras directly capture depth information using a single camera. The basic principle of TOF cameras is as follows: the camera emits acoustic, electromagnetic, or other waveforms; these are reflected from object(s) in the scene, some of which are incident upon the sensor's detector. The distance between object and sensor, at each pixel, can be computed from the travel time of the emitted, reflected, and sensed waveform. Once the depth information for each pixel is known, left and right views can be rendered for 3D reproduction using a depth-image-based rendering technique.[1]

While 3D capture, processing, and display techniques have evolved rapidly in recent years; there's a relative dearth of 3D content compared to potential market demand. The ocean of older 2D videos captured during past decades contains a treasure trove of 3D possibilities. Because generating native 3D content is a relatively slow process, 2D-to-3D video conversion offers a tremendous opportunity to alleviate this shortage.

The target accomplishment of 2D-to-3D conversion techniques is to be able to estimate the true 3D information present in the scene (for example, in the form of depths relative to some baseline) from single, monocular video streams. Once the 3D information has been estimated, it can be deployed to create a 3D (stereoscopic) video from the existing 2D video. The successful development of efficient and reliable 2D-to-3D video-conversion algorithms could make it possible to create large quantities of high-quality 3D content at a low cost in terms of time and resources. By analogy, just as many classic movies were once colorized, and as they are now being up-sampled to high definition using super-resolution techniques, we believe that much of the available corpus of attractive archived videos will be converted to 3D formats.

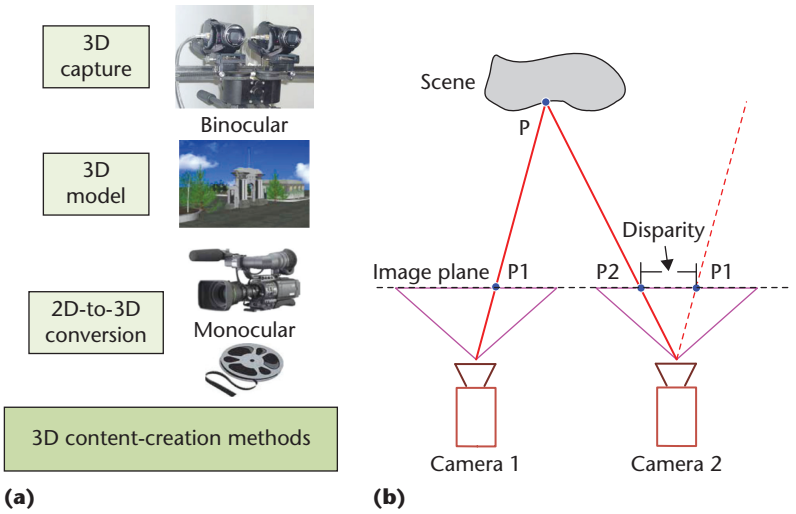


*Figure 1. Generating 3D content. (a) Contemporary methods used for 3D content creation. (b) Stereoscopic camera and the concept of disparity.*

## Perceptual depth cues

The real world is 3D; however, during the process of projection through the lens, one of the three dimensions is lost: depth. To obtain the survival advantage afforded by depth information, humans have evolved two frontally placed eyes separated by an interocular distance of about 65 millimeters (in adults). Spatial points in the scene project to slightly different positions between the 2D projection on left and right eyes; this displacement between the projected positions is called *disparity*, as Figure 1b illustrates. If the left and right views are placed so that the optical axes of the eyes (or cameras) are parallel, then the disparity between the locations of points that were projected from the 3D scene onto the left and right 2D image planes is inversely proportional to depth. Such a camera configuration is called a *parallel optical axis binocular geometry*. This camera arrangement simplifies algorithm design in certain ways but doesn't exploit other advantages

available if the cameras can toe-in, or verge as human eyes are normally able.

A normal human viewer has eyes that are highly mobile and that engage in a number of types of eye movements, including binocular vergence, whereby the optical axes of the two eyes fixate on the same point. Cameras can also be made to verge under computer control provided that points of stereo fixation can be selected in the 3D scene. A vergent stereo imaging geometry is a powerful means for concentrating the visual attention of a vision system on a particular region of 3D interest. However, a more complex geometric relationship exists between points that are projected to a vergent stereo pair of videos. For simplicity we will assume that the acquisition cameras are arranged in parallel-axis geometry in the following discussion. In practice, this is not always the case, as the stereographers will wish to capture the stereo content using a geometry that will create stereo pairs that are comfortable to view in 3D. This is often correctly done using cameras in a vergent geometry that can capture images that are subsequently more consistent with those captured by vergent eyes.

While binocular vision, or stereopsis, is one of the most important modes of human visual-depth sensing, it's not the only source of 3D information nor is it necessarily the most powerful 3D sensing cue. Indeed, people might still perceive depth even with a single eye. Other people, although equipped with two working eyes, have partial or complete stereo deficiency. The number of people who have imperfect (or nonexistent) stereo vision lies between 4 and 10 percent of the overall population, yet these people still experience varying degrees of 3D visual sensation. Depth perception is stimulated not only by binocular depth disparity, but also by a variety of monocular depth cues, such as occlusion (where one visible object covers another), motion parallax (changes in pose with motion), diminution in size or brightness (with distance, shading, and texture gradients), and so on.[2] In the absence of binocular information, 2D-to-3D video conversion would enable the creation of 3D video content by using monocular depth cues such as those just mentioned. In the following, we describe how these monocular depth cues can be incorporated into 2D-to-3D video-conversion systems.

## Fully automatic 2D-to-3D video conversion

*Fully automatic 2D-to-3D video conversion* means that 3D information is estimated from 2D video clips automatically, without a priori knowledge of depths in the scene. This technique aims at scenarios where humans are unable to participate in the depth-computation process, as well as real-time applications where human participation is impossible. For example, a desirable goal is the development of real-time, 2D-to-3D video-conversion modules suitable for incorporation into 3DTV receivers. In this situation, algorithms must be able to extract information from the input 2D video that can be used to directly compute depth dimension. All of the aforementioned monocular depth cues can contribute toward estimating a scene's 3D structure. However, it's unlikely that any monocular depth cue could prove adequate for estimating depths from videos of all kinds of scenes.

One approach to simplifying this issue is to execute a preprocessing step, where we pick a monocular cue category we think will prove most valuable in representing the video's 3D content. The most commonly used monocular depth cues in the fully automatic 2D-to-3D conversion algorithms are occlusion/relative size, texture gradient, perspective/horizon, focus/defocus, and motion parallax. A detailed explanation and illustration for these depth cues can be found at http://media.au.tsinghua.edu.cn/2Dto3D/monocularcues.html.

However, all of these depth cues are imperfect, and given a situation, none are necessarily available or will supply an adequate amount of information to allow for depth computation. Therefore, an overall, fully automatic 2D-to-3D video-conversion system should, ideally, seek to combine multiple monocular depth cues to compute a stable and reliable depth map. For example, when a piece of video containing a sufficient number of frames is acquired, the incoming frames might be categorized into static (no- or low-motion) and dynamic (high-motion) frames using a change-detection algorithm that employs statistical motion information computed from neighboring frames. A third category could be frames where the scene abruptly changes from one situation to another, such as the occurrence of a commercial in ordinary TV programming.

One possible sequence of processing could be to first parse the video at scene changes; then, compute motion information and use it to categorize frames. For static frames, deploy depth cues as occlusion/relative size and focus/defocus to initially estimate the scene depths; whereas for dynamic scenes use motion parallax for the first estimates. In both situations, depth cues, texture gradient, and perspective/horizon can be used to provide supplementary depth information. For complete scenes (between scene changes), consistency over time can be used to improve and affirm the depth computations. Depth post-processing can then be applied to smooth the estimated depth information in both space (within connected objects or textured regions) and over time, to alleviate flicker and other artifacts that might become evident when watching the 3D video. Finally, given the 2D color video and the newly computed depth video, we can create an associated second color video from the associated disparities (presuming a viewing distance and geometry), so that the two color videos can be viewed stereoscopically.

## Semiautomatic 2D-to-3D conversion

Because 2D-to-3D conversion attempts to recover the depth information lost during camera projection, it's a poorly posed problem; primarily because an insufficient reservoir of physical constraints is placed on the possible 3D reconstructed scene by the available projection data. A convenient solution is to involve human participation as a means for injecting prior knowledge about the true 3D scene structure onto the video shot to be converted, thereby enabling both more efficient conversion and more accurate results. In most semiautomatic 2D-to-3D video-conversion systems, specific frames (keyframes) of the video sequence are identified and annotated with 3D (depth or disparity) information via human intervention. The other frames (non-keyframes) are converted to 3D automatically using the information in the keyframes as constraints. This strategy is feasible because the frames in a video taken of a given scene typically have a high degree of correlation, containing objects and surfaces that appear across many video frames, albeit from changing observation points. This high correlation is one of the fundamental attributes of video that allows for high compression: in algorithms such as H.264, an I-frame (similar to a keyframe) is transmitted faithfully using a larger bandwidth, while non-I frames (P-frames, similar to non-keyframes) are compared with others frames, with the differences being compressed and transmitted. Similarly, in semiautomatic 2D-to-3D video conversion, keyframes are given carefully measured human-annotated depth information that indicate partial, or even complete, information about the scene.

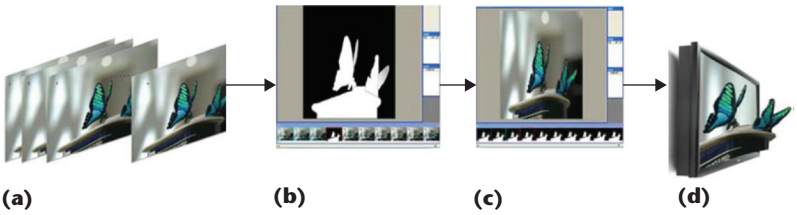


*Figure 2. Workflow of semiautomatic 2D-to-3D conversion. (a) Reading 2D video into frames, (b) keyframe detection and depth annotation, (c) depth propagation, and (d) stereoscopic rendering and display.*

The depth annotations can then be used to initialize and constrain the depths computed from subsequent non-keyframes (see Figure 2). As a design starting point, keyframes might always be assigned at scene changes, as well as at other significant temporal instants in a video—for example, at category changes. Figure 2 shows a typical overall framework for semiautomatic 2D-to-3D video conversion.

Among the four steps of the 2D-to-3D video-conversion framework depicted in Figure 2, the depth-propagation stage is the most important, because it determines the depth quality of most frames of the video. Depth propagation is an inference problem, wherein the known colors (RGB values of all the pixels) from all frames is combined with depth information from just the keyframes, to infer the depth information of the non-keyframes.

A large number of different depth-propagation kernels have been proposed in recent years. Some methods infer the depth information by exploring the relationships between depth and color. This is because pixels of similar color and texture between neighboring frames are likely to come from the same object(s) if their spatial locations are also similar.

The so-called bilateral filter[3] is a powerful tool for approaching such problems. This device combines two attributes, such as color and space, with linear weighting. In our problem, similarity in color space and spatially proximity can be used to define the bilateral filter.

Although a bilateral filter can be used to effectively and smoothly propagate depth information based on color and spatial similarity, problems can occur when there's fast movement in the video. Movement causes luminance to change position relative to the video frame, which can invalidate the assumption that neighboring pixels share similar depths. We therefore advocate semiautomatic 2D-to-3D video-conversion methods that add a third factor into the bilateral depth-propagation framework: motion. Motion estimation is first implemented to establish temporal correlation between the pixels in successive frames. In this way, target pixels being examined in nonkeyframes can be shifted, or motion compensated by displacement relative to the reference keyframes. Bilateral filtering is then performed on the shifted pixels. Such a shifted bilateral filter (SBF) can account for correlations across three attributes: color, space, and time. With an SBF, the motion estimation might proceed by assuming that the shifted pixels occur at the same depth layer; however, in practice there might also be displacements in depth (along the camera's optical axis, forward or backward from the camera). In such cases, the SBF approach can lead to incorrect depth propagation. To solve this problem, a bidirectional propagation strategy can be used to further update the SBF depth-propagation kernel.[4]

## Quality of experience

At this point, it's natural to wonder whether it's possible to estimate the perceptual quality of a converted 3D video, and to likewise compare different conversion algorithms. This question leads us to an important field: image/video quality assessment (QA). The goal of QA is to develop algorithms that can accurately predict the quality of input images and videos as humans will judge them. Therefore, QA contains two important aspects: one is human subjective opinions about converted 3D data, and the other is designed QA models that lead to efficient and accurate algorithms. While the ideal arbiter of converted video quality would be recorded human judgments, such subjective evaluation data requires a lot of participants and a controlled test environment, which is usually infeasible in practice. Therefore, objective QA algorithms that can accurately predict human judgments of converted video quality are highly desirable. The topic of QA on 2D images and videos has been studied for many years.[5] However, understanding the quality of experience (QoE) of human viewers of 3D videos is a much more complicated problem, and promising results are still forthcoming.

First of all, distortions of stereoscopic videos might be perceived differently than distortions of 2D videos, even if the distortions are of the same type (compression, packet loss, and so on). Indeed, these 2D distortions can lead to distortions in the perception of the third dimension. The geometry of stereopsis can lead to other visual artifacts such as depth-plane curvature, shearing, and keystone distortion.[6] Distortions can also arise from imperfect stereoscopic capture, such as poor synchronization of capture of the left and right views, or by faulty left–right image-rendering procedures during 3D content generation.

Even disregarding possible distortions in the 3D video content, a viewer might still feel visual discomfort in seeing the third dimension in a 3D video. This can arise from problems in stereography if the geometry of the captured images (as defined by the distances to objects), the vergence angle between the cameras, and presumed point of 3D fixation don't match what could be viewed by a human observer. In a converted video there's only a single camera, but geometric parameters must be defined to create the second view.

Some 3D displays can also result in visual distortions. For example, commonly used 3D glasses separate left and right images using different polarization modes; however, some amount of the left image might leak into the right eye view and vice versa, which is a form of cross-talk. This might occur because of poor synchronization or separation of the two views. Other 3D display artifacts[7] can reduce the quality or the realism of the 3D viewing experience.

Last, despite decades of research there's much that remains unknown regarding stereo perception. In particular, little is known regarding the perception of 3D distortions, or of the particulars of the way in which geometric inconsistencies affect the QoE. It's particularly difficult to create algorithms that automatically predict the QoE of a stereo 3D video, because there's no ground-truth reference signal available against which to compare the experience. Indeed, the only 3D signal that occurs is the so-called 3D cyclopean image in the brain.

There are only crude methods available to estimate this distal signal, for example, by computer-vision stereo algorithms.

We acquired ten 3D video clips with associated ground-truth depth maps, which we used to evaluate the aforementioned semiautomatic 2D-to-3D video-conversion algorithms. We obtained both subjective evaluations and objective QA algorithm results on these data sets. The subjective test invited 22 viewing subjects of various ages and backgrounds to watch and rate the 3D videos by ranking and score. Here, the percentage of times subjects selected a method as best (first choice) or second best (second choice) are tabulated. In the absence of a standard 3D QA index, and because we believe that 3D QA algorithms will operate using principles different from 2D QA algorithms, we simply measured the mean squared error between depth maps from converted video and ground truth. These results (listed in Table 1) suggest that the efficacy of the algorithms can be distinguished. We hope this data set will help bridge the gap between 2D-to-3D video conversion and 3D QoE research, benefiting both disciplines.

***Table 1. The subjective and objective test on different 2D-to-3D conversion methods.****

| Conversion method | Subjective first choice (%) | Subjective second choice (%) | Mean squared error |
|---|---|---|---|
| Bilateral filter | 4.55 | 0 | 4.51 |
| Shifted bilateral filter | 18.2 | 22.7 | 4.42 |
| Shifted bilateral filter with bi-direct[4] | 22.7 | 45.5 | 3.89 |
| Ground truth | 54.5 | 31.8 | 0 |

* The data sets are available at http://media.au.tsinghua.edu.cn/2Dto3D/Testsequence.html.

## Conclusions

The cost of creating 3D video content remains high, and the amount of quality 3D content being delivered is far from meeting the market demand. Breakthroughs might come with the introduction of efficient and effective 2D-to-3D conversion algorithms. Further studies on the perception of stereoscopic distortions and artifacts are likely to deepen our understanding of the 3D QoE. The database we discussed should prove useful for potential efforts toward creating effective 2D-to-3D conversion algorithms and for assessing the QoE delivered by them. **MM**


## Acknowledgment

The authors would like to thank the National Basic Research Project (no. 2010CB731800) and the Key Project of NSFC (no. 61035002, 60932007, and 61021063).


## References


1. C. Fehn, ''Depth-Image-Based Rendering (DIBR), Compression, and Transmission for a New Approach on 3DTV,'' *Proc. SPIE,* Int'l Soc. for Optics and Photonics (SPIE), vol. 5291, 2004, pp. 93-104.
2. I.P. Howard and B.J. Rogers, *Seeing in Depth: Depth Perception,* Porteous, 2002.
3. C. Tomasi and R. Manduchi, ''Bilateral Filtering for Gray and Color Images,'' *Proc. Int'l Conf. Computer Vision* (ICCV), IEEE CS Press, 1998, pp. 839-846.
4. X. Cao, Z. Li, and Q. Dai, ''Semi-Automatic 2D-to-3D Conversion Using Disparity Propagation,'' *IEEE Trans. Broadcasting,* vol. 57, no. 2, 2011, pp. 491-499.
5. H.R. Sheikh, M.F. Sabir, and A.C. Bovik, ''A Statistical Evaluation of Recent Full Reference Image Quality Assessment Algorithms,'' *IEEE Trans. Image Processing,* vol. 15, no. 11, 2006, pp. 3440-3451.
6. A. Woods, T. Docherty, and R. Koch, ''Image Distortions in Stereoscopic Video Systems,'' *Proc. SPIE,* Int'l Soc. for Optics and Photonics (SPIE), vol. 1915, 1993, pp. 36-48.
7. L. Meesters, W. IJsselsteijn, and P. Seuntiens, ''A Survey of Perceptual Evaluations and Requirements of Three-Dimensional TV,'' *IEEE Trans. Circuits and Systems for Video Technology,* vol. 14, no. 3, 2004, pp. 381-391.



*Contact author Xun Cao at xuncao@gmail.com.*

*Contact editor Wenjun Zeng at zengw@missouri.edu.*