International Journal of Informatics and Communication Technology (IJ-ICT)
Vol.7, No.3, December 2018, pp. 141~145
ISSN: 2252-8776, DOI: 10.11591/ijict.v7i3.pp141-145

Addressing Latency Issues in 2D to 3D Conversion: Deploying Available Synthetic Database

N. L. Manasa
GNITS, JNTUH, Hyderabad, India

Article Info

Article history:

Received Jun 20, 2018
Revised Aug 6, 2018
Accepted Aug 20, 2018

Keywords:

3D graphical rendering
3D stereoscopic videos
Depth perception
Gaussian Mixture model

ABSTRACT

Conventional 2D to 3D rendering techniques involve a sequential process of grouping of the input images based on edge information and predictive algorithms to assign depth values to pixels with same hue. The iterative calculations and volume of data under scrutiny to assign 'real-time' values raise latency issues and cost considerations. For commercial consumption, where speed and accuracy define the viability of a product, there is a need to reorient the approach used in the present methodologies. In predictive methodologies one of the core interests is achieving the initial approximation as close to the 'real' value as possible. In this work, 'synthetic' database is used to provide the first approximation through comparison techniques and fed to the predictive tool. It is believed that this work will provide a basis for developing an efficient 2D to 3D conversion methodology.

Corresponding Author:

N. L. Manasa,
GNITS, JNTUH, Hyderabad, India
Email: Nadipally.m@gmail.com; Nadipally.m@gnits.ac.in.

1. INTRODUCTION

In the last decade, the need and consequent demand for 3D (3 Dimensional) images and videos has grown owing to its utility in different fields, such as medical and engineering diagnostics and restorative interventions [1]. In addition, the entertainment industry uses it to add the 'real feel' factor for improved consumption. As such, the need for faster conversion techniques of conventional 2D (2 Dimensional) to 3D imagery has gained wider attention in different industries. Although different approaches have been tried to modulate planar images to 3D images for better visual experience to considerable extent, there still exists need for better quality that is both, faster and commercially viable.

Amongst the principal concerns involved in the conversion of planar, 2D images to 3D visuals is to recreate the 'depth' and 'color' (hue) parameter from the original images. Some of the notable and widely used techniques in the pursuit of generating the 'depth aspect from conventional 2D images include stereoscopic vision (to allow for triangulation), active (dynamic) methods, and 3D graphical (algorithmic, software-based) rendering techniques. While the first two techniques mentioned (stereoscopic imagery and active sensors require specific, and contemporarily high-cost equipment to help assess and render the 'depth parameter, the third (3D graphical rendering) technique involves time-consuming iterative sequence. The stereoscopic methods require multiple cameras to capture 2D images and thereby assess the depth required to create the visually compliant 3D version. In the active image capturing methods, photographers and videographers use different sensors such as time-of-flight and structured light capabilities to estimate the depth factor of 2D images. Such records are not commercially feasible[2] owing to skills required to operate specially devised equipment and the economic feasibility apart from access to such equipment. The third method of graphical rendering techniques used conventionally is usually a skilled and experiential manual process requiring tedious and time-consuming engagement, thereby making it unfeasible for commercial, large-scale consumption. Of late, there is a growing accent on developing automated processes that can

convert 2D images to 3D visuals by integrating the missing third 'depth' parameter in the planar images obtained in conventional photo/video.

In order to improve the quality of 3D rendering, designers are using different predictive techniques and developing algorithms that work faster and more accurately and approach 'real-time' experience sought by consumers in different applications. The human eye estimation of depth is achieved through combined effect and superposition of binocular and monocular depth perception [3]. Whereas, in the binocular visual, humans are able to decode the third dimension through convergence of differential elements from both visions and correctly estimate the depth, the monocular parameters are also capable of reproducing the desired information using similar properties and inducting the differentials. Such cues are: dynamic parallax, planar proportional measures, the relative texture changes, and focal distances. The implication of the discussion is that monocular images/videos can be used to produce 3D visuals. Experimental support and practical application of Visual parameters and disparities that provide the mentioned cues have been provided by different studies to generate 3D from conventional images.

The various methods that have attempted 2D to 3D rendering use two main types: Single-frame Method and Multiple-frame Methods. Notable among Single-frame Methods is the attempt by Battiato [4] who made use of multiple cues, a progression over some other methods that concentrated on single cues – focal proximity [5], machine learning algorithms [6-7], and 2 D proportional dimensions (height and width ratio) and texture changes [8] and geometric measurements. A significant development in Multi-frame Method is the successful hybridization of MTD (Modified Time Difference) and CID (Computed Image Depth) achieved by Iinuma et al., [9] that uses triangulation successfully to generate close to real 3D images. Recent approaches [18]-[20] that showed commendable results when utilized in a different application but can also be deployed to solve for 2D to 3D rendering. Stereoscopic footage and depth assessment from motion provide the required data for triangulation [10–13]. Analysis of Motion parallax offers the temporal correspondence brought about by the existing disparity. The motion vector (MV) thus extracted is converted to disparity vector (DV). In an alternate method, the depth map is not produced, instead a stereoscopic vision comprising of right-eye and the left eye vision. However, such methods require the deployment of motion camera. As stated earlier, a combination of monocular and binocular strategies appear to have the potential to deliver the desired quality efficiently.

1.1 Single-frame Methods

Park et al., [5] laid emphasis on the clarity content of different areas measured by the distance from the focal plane to estimate the relative depth of elements in the 2 D image. The image is divided into rectangular blocks that form a matrix and the blurriness of each block is considered to estimate the depth of that part of the image from the viewer. The limitation of such estimation is clearly the number of discernible blocks that can be accommodated leading, in turn to inconsistency in read-off values amongst non-adjacent patches of same shades, and the accuracy of the measuring device to measure gradually changing hues that may vary by infinitesimally small amounts. This method is also called the CID (Computed Image Depth). The machine learning methodology was used by Hoiem [7] who mapped the image to assess the depth and then added the extracted parameter to the 2D image to obtain the 3D visual. In machine learning methodology, the objects and distances are assigned progressively through trained classes, and pose the potential skipping of cases in the initial training phase causing perceptible inaccuracy in the final rendition, as the progressive stages depend heavily on the preceding feed. Thus, missed cases can lead to unacceptable disruption from the desired outcome. The geometric method used by Tsai [14] employs the 'vanishing point' concept to add the third dimension from single frames. Jung et al., [8] used the information accrued from the edges and tried to consolidate the process with the help of known data (hue against depth) to estimate the required quantity (depth) - perhaps a first step towards predictive extraction. The above methods may be successfully deployed in certain specific cases individually and limited to shorter depth estimation in images used. Loss of other parametric values than the ones under consideration makes the methods using single parameters prone to overlooking other significant properties that add to the depth perception. Consequently, the generated 3D images are most likely to suffer from loss of temporal coherence and spiking.

1.2 Multi-frame Methods

The work on 2D to 3D conversion using Multi-frames has to contend with the issue of depth within an object in the frames and relational depth information of all objects in the frames. A methodology of grouping of pixels with same hue and spatial orientation is the basic need for obtaining desired uniformity in depth estimation. Some earlier works have used motion parallax as the primary cue to allow for grouping of pixels within the frames. However, some pixels carry different self MV's leading to error-prone depth information. Thus the contention seems to revolve around the methodology needed to group pixels depth values to predict the depth of individual objects and for relational depth estimation. This paper attends to the

issue of depth prediction through a novel scheme of referential depth assignment presented in the following section. Such initialization value helps accelerate the follow-up iterative estimation process. Experimental result measurement and feedback from consumers of the video feed supports the contention that the output is visually pleasing and comparable to synthetic video-game experience, which is itself rated as very close to 'real'. Section 2 presents the proposed system. Section 3 presents the outcome of the experimental rigor and discusses the complexity and accuracy as compared to other contentions regarding depth estimation. Finally section 4 and 5 summarize the paper and suggests possible areas of further work.

2. PROPOSED SYSTEM

Commercial viability and time consumed to add the depth perception through iterative predictions are not practical when streaming live events. Most of the 'pixel grouping' methodologies seem to serve at best academic interests. Though such methodologies of 3D rendering are potentially very close to 'real' perception of observations to the human eye, the time lag between image capture and final output delivered for consumption is rarely within acceptable limits. The stages and sequence (starting from left: left gradient, left bottom, bottom-up, right bottom, right) of hue grading to assess depth that converts 2D to 3D usually follows the sequence depicted in Figure 1.

Figure 1. [15]

The proposed system attempts to improve the estimation process by using existing referential depth cue instead of trying to estimate the same from the frames that have to be converted. Such reference hue gradient helps in accelerating the estimation procedural speed by providing the difficult first estimation step.

Use of existing good quality 3 D stereoscopic videos can be used to provide the referential depth cue for the images that need to be processed to deliver 3D for live streaming. Appropriate 3 D video games may be used as referential input that corresponds most closely to the images under consideration. In this work live streaming of soccer games is attempted.

The most acceptable quality video games provide the most appropriate depth perception cue through the texture gradient used to develop such games. The rest of the process utilizes appropriate iterative algorithms based on a sound initial value. Thus, this process utilizes the existing high-quality in video games (that resemble real 'feel' very closely) and can provide appropriate referential hue gradient cues for depth estimation. The rest of the process is automatic as in other methods, too. This process improves the conversion process mainly by reducing the processing time through spatial (height and width) assignment of depth gradient inferred from the 'manufactured' referential video games. As this process uses existing stereoscopic 3D frames as reference cues, the proposed system is domain-specific. Hence, the algorithm is not generic. The proposed algorithm will have to be modulated based upon the contextual data. It also implies that each frame may need a corresponding data frame for reference. Thus for each generation, the process would refer to a database of similar 3D video games; assuring that the rendering is of good quality and the output is visually pleasing. This is an improvement over existing generic methodologies that may suffer from latency issues to deliver acceptable quality stereoscopic visual, whether static (images) or dynamic (live video streams).

3. METHODOLOGY

As is evident from Figure 2, the two main components that help create the depth dimension from the 2D images for proper gradient hue are: 1) the S-RGBD (synthetic- RGB and Depth) database from similar images elsewhere can be used as a reference, using the ones that most closely matches with the input image sectors; 2) Object masking that would help de-lineate edge profiles and create sharper depth perception, which otherwise would create blurriness at the edges due to overlapping depth values. The accumulation of S-RGBD imagery could be a daunting task, as many different collections would be required this issue can be

handled by categorizing images based on situation, scenery-based value, event-based images/videos and other such combinations in addition to video games that best incorporate different imagery with quality depth and hue cues. Masking involves separating different object boundaries that is achieved by preprocessing the input images based on appearance, hue, and motion. For close-ups, local features such as matting-based approach are used that involves initiation using trajectory segmentation [16], [17] for non-close-up images the GMM (Gaussian Mixture Model) processing delivers the required content of fine quality.

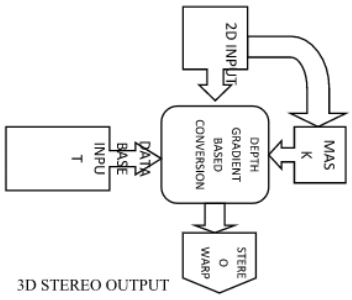

Figure 2. The conceptual process sequence is depicted in Figure 2

The overarching concept is that of Gradient-based conversion. The gradient in the hue corresponds to depth of the objects and can be used to assign corresponding values. In order to achieve accurate assignment, the input image is differentiated in n x n blocks and then compared with similar images that assign the depth values according to the hue in the reference image (S-RGBD database). The masked images with clear differentiation of objects add to the clarity, both in hue and depth.

The picture quality and computational difficulties form the basis of the trade-off points where the conversion process is required to achieve a balance between these two parameters. The computational difficulty arises from the number of blocks created to achieve a visually pleasing 3D image. The duration required for more number of blocks is $O(e^{5E-5n})$, where, n is the number of blocks. Figure 3 below illustrates the computational duration.

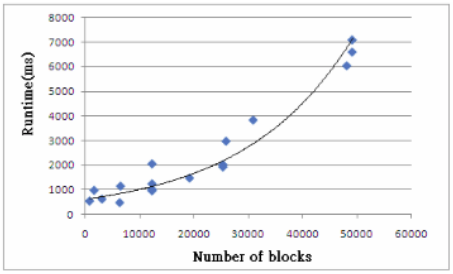

Figure 3. Computational complexity [15]

4. DISCUSSION

The basic benchmark categories for 3D perception in images are based on three categories: visual pleasantness, depth quality, and picture quality. Visual pleasantness/discomfort is the physiological discomfiture such as visual strain, headache on continuous viewing and other symptoms. The depth quality is measured as a perception of envisaged depth. The transmission parameters and the encoding framework affect the picture quality. Since depth quality and visual pleasantness are qualitative parameters, assessment should involve showing different 3D images generated by different techniques to the same set of viewers/respondents to get a fairer grading. Such images/videos can be shown for short durations in a random sequence, and allow the viewers to rate them on ascending order of quality, marked by distinct numbers/grades (bad, poor, fair, good, excellent) for depth quality and on a similar grading that starts at very

uncomfortable and ascends to very comfortable for visual comfort. In a random study carried out thus, the images produced by the presented technique scored very high in the quality assessment on both counts.

5. CONCLUSION

This paper has introduced an innovative method for 2D to 3D conversion of images/videos. The strength of the proposal lies in faster conversion rate. The quality of the rendered 3D image is also rated highly by viewers. The main processes involved are those of S-RGBD database collated from similar videos or from video games and edge information. The overarching schema is the square matrix block segmentation that is eventually modified following depth hue gradation cues and comparison with referential database. The main contribution of the paper is that of domain centered initial prediction values. Matrix based edge information removes the boundary discontinuities, and use of existing, quality images, that adds to the gradient discernment. Thus the method resolves two major issues – that of latency and investment required for 2D to 3D conversion.

A possible limitation could be the diverse and quality database required to implement the process presented in this work. Secondly, this system is not generic - a sacrifice for quality and reduction in latency.

REFERENCES:

[1] C. R. Madan, "Creating 3D visualizations of MRI data: A brief guide," *F1000Research*, vol. 4, Aug. 2015.

[2] A. Zelener, "Survey of object classification in 3d range scans," *Tech. Rep*., 2015.

[3] W. J. Tam and L. Zhang, "3D-TV Content Generation: 2D-to-3D Conversion," in 2006 *IEEE International Conference on Multimedia and Expo*, 2006, pp. 1869–1872.

[4] S. Battiato, S. Curti, M. L. Cascia, M. Tortora, and E. Scordato, "*Depth map generation by image classification*," presented at the Three-Dimensional Image Capture and Applications VI, 2004, vol. 5302, pp. 95–105.

[5] J. Park and C. Kim, "*Extracting focused object from low depth-of-field image sequences*," presented at the Visual Communications and Image Processing 2006, 2006, vol. 6077, p. 60771O.

[6] P. V. Harman, J. Flack, S. Fox, and M. Dowley, "*Rapid 2D-to-3D conversion*," presented at the Stereoscopic Displays and Virtual Reality Systems IX, 2002, vol. 4660, pp. 78–87.

[7] D. Hoiem, A. A. Efros, and M. Hebert, "*Automatic Photo Pop-up*," in ACM SIGGRAPH 2005 Papers, New York, NY, USA, 2005, pp. 577–584.

[8] Y. J. Jung, A. Baik, J. Kim, and D. Park, "*A novel 2D-to-3D conversion technique based on relative height-depth cue*," presented at the Stereoscopic Displays and Applications XX, 2009, vol. 7237, p. 72371U.

[9] T. Iinuma, H. Murata, S. Yamashita, and K. Oyamada, "*54.2: Natural Stereo Depth Creation Methodology for a Real-time 2D-to-3D Image Conversion,"* SID Symp. Dig. Tech. Pap., vol. 31, no. 1, pp. 1212–1215, May 2000.

[10] J. y Chang, C. c Cheng, S. y Chien, and L. g Chen, "*Relative Depth Layer Extraction for Monoscopic Video by Use of Multidimensional Filter*," in 2006 IEEE International Conference on Multimedia and Expo, 2006, pp. 221–224.

[11] Y. L. Chang, C. Y. Fang, L. F. Ding, S. Y. Chen, and L. G. Chen, "*Depth Map Generation for 2D-to-3D Conversion by Short-Term Motion Assisted Color Segmentation*," in 2007 IEEE International Conference on Multimedia and Expo, 2007, pp. 1958–1961.

[12] C.-H. Choi, B.-H. Kwon, and M.-R. Choi, "A real-time field-sequential stereoscopic image converter," *IEEE Trans. Consum. Electron.* vol. 50, no. 3, pp. 903–910, Aug. 2004.

[13] C.-C. Cheng, C.-T. Li, P.-S. Huang, T.-K. Lin, Y. M. Tsai, and L. G. Chen, "*A block-based 2D-to-3D conversion system with bilateral filter*," in 2009 Digest of Technical Papers International Conference on Consumer Electronics, 2009, pp. 1–2.

[14] Y. M. Tsai, Y. L. Chang, and L. G. Chen, "*Block-based Vanishing Line and Vanishing Point Detection for 3D Scene Reconstruction*," in 2006 International Symposium on Intelligent Signal Processing and Communications, 2006, pp. 586–589.

[15] C. C. Cheng, C. T. Li, and L. G. Chen, "A novel 2D-to-3D conversion system using edge information," *IEEE Trans. Consum. Electron.* vol. 56, no. 3, pp. 1739–1745, Aug. 2010.

[16] A. Levin, D. Lischinski, and Y. Weiss, "A Closed-Form Solution to Natural Image Matting," *IEEE Trans. Pattern Anal. Mach. Intell*., vol. 30, no. 2, pp. 228–242, Feb. 2008.

[17] P. Ochs, J. Malik, and T. Brox, "Segmentation of Moving Objects by Long Term Video Analysis," *IEEE Trans. Pattern Anal. Mach. Intell*., vol. 36, no. 6, pp. 1187–1200, Jun. 2014.

[18] Lee, Jongwon, Hyunju Lee, Donggyun Yu, and Hoekyung Jung, "Body information analysis based personal exercise management system." *International Journal of Electrical and Computer Engineering (IJECE*) 8, no. 1, 2018.

[19] *Zakaria Boulouard, Amine El Haddadi, et al,* **"**Bat-Cluster: A Bat Algorithm-Based Automated Graph Clustering Approach", *International Journal of Electrical and Computer Engineering (IJECE)* 8, no. 1, 2018.

[20] *Madhu Chandra*, "Framework for Contextual Outlier Identification using Multivariate Analysis approach and Unsupervised Learning", *International Journal of Electrical and Computer Engineering (IJECE)* 8, no. 1, 2018.