Development of Methodology for Depth Estimation for Images and Videos

R.M.Mulajkar [#1], V.V.Gohokar[#2]
Research Scholar, Research Lab, E&TC Dept. JDIET, SGBAU, Amaravati (India)
[#2] Associate Professor, E&TC Department, MIT, Pune (India)
[1] rahul.mulajkar@gmail.com
[2] vvgohokar@rediffmail.com

***Abstract*— Recent experiences with stereoscopic image/video conversion have sharply increased their demand. Although 3D stereoscopic view enhances visual quality compared to 2D, depth information which is required to generate 3D view is unavailable for existing 2D content. Therefore, there is a large requirement to generate depth information. This paper uses a fusion of monocular cues as Motion, Aerial Perspective cue (AP), Linear Perspective cue (LP), and Defocus cue to estimate the depth. The proposed system developed a mechanism to re-estimate depth map if the estimated depth map is inaccurate in a situation such as fast motion, false foreground estimation. This algorithm is tested in different conditions such as the sequence of camera motion and multiple objects, static cameras and stationary background, a highly dynamic foreground, background with less motion and when motion is behind the foreground. The experimental results show that generation of the depth map is very close to the real depth map. Thus, the algorithm can be applied for 2D-to-3D conversion. To evaluate the performance of our system its results are compared with existing algorithms. The subjective evaluation test was performed on proposed algorithm. The result shows that proposed system has good performance.**

Keyword - Depth Map, 3D Video, 2D-to-3D conversion, Monocular cue, motion cue.

I. INTRODUCTION

There are numbers of 3D hardware's available such as 3D TVs, Blu-Ray Players, smart phones, Kinect Camera, LASER cameras. The new 3D can be generated by using such devices. In this, two or more than two cameras are fixed at particular angles. Two views (left and right) are generated to get a stereoscopic view. Large number of companies are working in the same field such as Dynamic Digital depth (DDD), HD logicx, Himax Technologies, In-Tree, Legend 3D, Samsung, LG, Stereo D, JVC, IMAX, but not limited. However, researchers are still working on algorithm which gives superior quality of depth estimation. The depth estimation from video or image can be done either by monocular view or by binocular view. The binocular video consists of two view images or two video similar to left view and right view. Thus, the depth can be extracted from left view and right view by comparing different features in two views. The monocular view consists of single image or video. It is necessary to generate a left view and a right view from single view depending on various cues as Color, Aerial Perspective cue, Linear Perspective cue, motion, focus/defocus etc. Thus, the estimation of the depth from a monocular video/image is a difficult task. The conversion of 2D-to-3D video involves two steps: Depth estimation from a given 2D video and generation of two view left view and right view. The task of generation of two views, left view and right view is performed by algorithms known as Depth based image rendering [DIBR]. These algorithms were well understood and there exists an algorithm that produces good quality video. Therefore, there is a requirement to develop a device (efficient algorithm) for extraction of scene depth information for conversion of a 2D-to-3D video. This Paper focuses on depth estimation from given monocular 2D video or image, which plays an important role in conversion from 2D-to-3D video.

Currently, there are two types of 2D-to-3D video conversion techniques. 1] Semi-automatic method: - It involves an operator to indicate the exact position of objects in the individual frames, placing them in suitable depth and final rendering. These methods are time-consuming but provide more accurate depth maps. Vary Camp et. al. propagated rendering by assigning manually disparity from key frame. Agnol et. al. used cross-bilateral filtering but it works only for simple images, not for complex images [2],[3]. 2] Automatic method: - In this operator is not needed. Consumer- electronics manufacturers developed different techniques, but they depended on some perfect assumption and failed on more challenging scenes. Over recent years, a number of algorithms are developed to get depth map automatically such as depth from motion [4], defocus or focus [5],[6], Linear perspective, edge [8], color [9]. These methods use a single depth cue to generate depth.

The rest of the paper is organized as follows: Section II describes the proposed algorithm in details. Section III shows evaluation and experimental results. Finally, Section IV concludes the paper.

II. EXPERIMENTAL METHODOLOGY

This paper uses four different monocular depth cues: motion cue, aerial perspective cue, defocus cue, and linear perspective cue to estimate depth. Initially, key frames are extracted from a scene. Four different cues discussed earlier are extracted. A multicue depth estimation module is proposed to generate the depth map. This paper utilizes the foreground and background separation to avoid complexity. After this, Depth Image Based Rendering (DIBR) algorithm gives 3D views. The workflow of our proposed system consists of following steps:

- Key frame and Non-Key frame Extraction
- Estimation of depth from four monocular cues such as motion, LP Cue, AP Cue and Defocus.
- Multi-cue fusion and depth refinement
- Stereo view frame synthesis.

A. Motion Monocular Cue

This algorithm always focuses on moving objects from the scene. The moving object is considered as foreground and background object is considered as non-moving objects. It is essential to separate this foreground and background object. Thus, Pixel (Frame) based method is employed to extract the background and foreground object. The formula is as below:

$$Diff\ (i, j) = \left| p_t(i, j) - p_{t-1}(i, j) \right| \tag{1}$$

Where $p_t(i, j)$ is present frame and $p_{t-1}(i, j)$ is the past frame. The difference is calculated, and this difference is compared to a threshold value. The threshold is calculated as follows:

$$\sigma_t = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(p_i - \overline{p})^2}$$
$$Diff\ (i, j) = \begin{cases} 1, & Diff\ (i, j) \geq \sigma_t \\ 0, & otherwise \end{cases} \tag{2}$$

Where σ_t - standard deviation, N is the number of pixels, $\overline{p}$ mean of the pixel values for the frame.

B. Linear Perspective Monocular Cue

To estimate depth map from linear perspective monocular cue, vanishing point is used. The vanishing point is the point of intersection of parallel in an image plane that appears to be vanishing but actually it is not vanishing. Thus to point a accurate vanishing point is the important task to estimate depth. Following steps are used to estimate depth map from the linear perspective monocular cue:

- Extraction of edges by canny edge detector.
- Remove the noise from edge detected image.
- Apply the Hough Transform to identify straight lines in an image.
- Locate the intersection of point of the curve.
- Find the vanishing point.
- Estimate depth map from vanishing points.

C. Aerial Perspective Monocular Cue

Aerial perspective is a result of scattering of light. When seen at a distance, the vapour, dust in the atmosphere leads the light to bend which creates a landscape of blurred outlines scene. This paper uses Dark channel prior based algorithm for removal of haze in the scene. Aerial perspective cue [30] of pixel p is given as-

$$D_{aerial} = 1 - \theta^{-t(p)} \tag{3}$$

Where θ is a variable used to adjust irradiance, $t(p)$ is a medium transmission, which is calculated using dark channel prior explained in [30].

$$t(p) = 1 - \min_{p \in \Omega}\{\min(I_{RGB}(p)\} \tag{4}$$

Where $\{\min(I_{RGB}(p)\}$ is the lowest intensity at each pixel in RGB channels, Ω is a window size of the minimum filter.

D. Defocus Monocular Cue

The mechanism provides an interactive system to add defocus cue to get depth map that will be close to real depth map. The depth value using defocus cue is calculated as-

$$d_{defocus} = s \cdot W_I \cdot \left\{ \frac{\alpha(p) - \alpha_{\min}(I)}{\alpha_{\max}(I) - \alpha_{\min}(I)} - r \right\} \tag{5}$$

Where W_I represents image width of image I, Aerial perspective cue. $\alpha_{\max}(I)$ is the maximum depth value in the images. $\alpha_{\min}(I)$ is the minimum depth value in the images. s is the control factor, in our system we consider s as 0.05. r is the variable determined by stereo effect ,we have taken r =0.5.

E. Fusion of Monocular Cues

This algorithm selected four monocular cues–motion cue, AP cue, LP cue and Defocus cue. Each of these gives different depth estimation. Each individual cue produces depth. The combination of this monocular cue removes weakness of each other and gives better depth estimation. This paper fuses the depth map of each depth cue. The proposed system interacts with user to select certain shot for depth refinement. The proposed algorithm divides a input scene into two sections: foreground and background region.

a) Background Region:

The depth information of the object which consists of background is given as-

$$D_{Fusion} = w_{LP} \times D_{LP} + w_{AP} \times D_{AP} \quad (6)$$

Where w_{LP} , w_{AP} represents weight of the linear perspective cue, Aerial perspective cue. These weight values are determined by the human visual system as explained in [22]. D_{LP} , D_{AP} are depth information of Linear perspective cue and Aerial Perspective cue. This algorithm set $w_{LP} = 0.8$, $w_{AP} = 0.2$ as a parameter. These parameters can be varied based on the stereoscopic effect, but it is required that addition of these parameters should be 100% i.e. sum should be 1.

b) Foreground Region:

The depth information about the object which consists of foreground is given as-

$$D_{Fusion} = w_{motion} \times D_{motion} + w_{AP} \times D_{AP} \quad (7)$$

Where w_{motion} , w_{AP} represents weights of motion cue, Aerial perspective cue.

F. 3D Warping

The original image points at location (P_m , y) are transformed to left points (P_{left} , y) and right points (P_{right} , y) [29] calculated as-

$$\left.\begin{aligned} P_{Left} &= P_m + \frac{B \cdot f}{2 \cdot z} \\ P_{right} &= P_m + \frac{B \cdot f}{2 \cdot z} \end{aligned}\right\} \quad (8)$$

Where f is focal length of camera, B is distance between left and right cameras. Z represents depth value of each pixel, and it is calculated as-

$$Z = \frac{1}{\left\{\frac{d}{255}\left(\frac{1}{z_{\min}} - \frac{1}{z_{\max}}\right) + \frac{1}{Z_{\max}}\right\}} \quad (9)$$

Where d the depth level from 0-255, $z_{\min}$, $z_{\max}$ are the minimum and maximum depth values.

III.EXPERIMENTAL RESULTS

The experimental platform is a PC I5 2.90 GHz and 4GB RAM. This algorithm can be applied to input image sequences as well as video sequences.

A. Simulated and Experimental Result of image sequences

Four depth maps based stereoscopic image sequences- *interview, orbi, Break dance,* and *Ballet* were used in the experiment which shows different conditions. *Interview* is captured with a static camera, and background is static, *Orbi* has camera motion and multiple objects; the sequences *Break dance* and *Ballet* are obtained from the multiview image sequences [33], *Break dance* contains a highly dynamic break dancer in the foreground, and a number of supporters are with less motion at the background. *Ballet sequence* has a stationary observer in the foreground, and a *Ballet dancer* is behind the foreground observer. These sequences are captured by a stationary camera. These dataset have true depth map so we can compare the performance of our 2D-to-3D conversion system. These are generated by the Interactive Visual Media group at Microsoft Research [33]. Fig. 1 and Fig.2 shows the experimental evaluation of estimated depth map. Foreground object and depth estimation may be not correct if the object in fast motion, mechanism provides a user interface to estimate the depth as per the requirement. It should be noted that it is not possible to estimate depth map exactly to true depth map, but researcher's goal is to tries to approximate estimated depth map to true depth map. Our system uses four monocular cue-motion, LP, AP and defocus. In images, motion cue doesn't exist, hence it is not used.

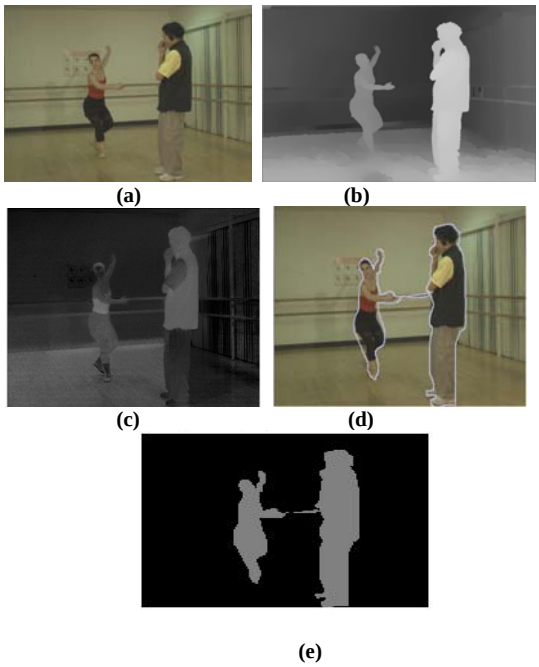

Figure 1. Experimental result of Ballet Sequence: a) input image (b)True Depth map (c)Estimated depth map(our Proposed algorithm)(d)Updating with defocus cue (e)Final estimated depth map

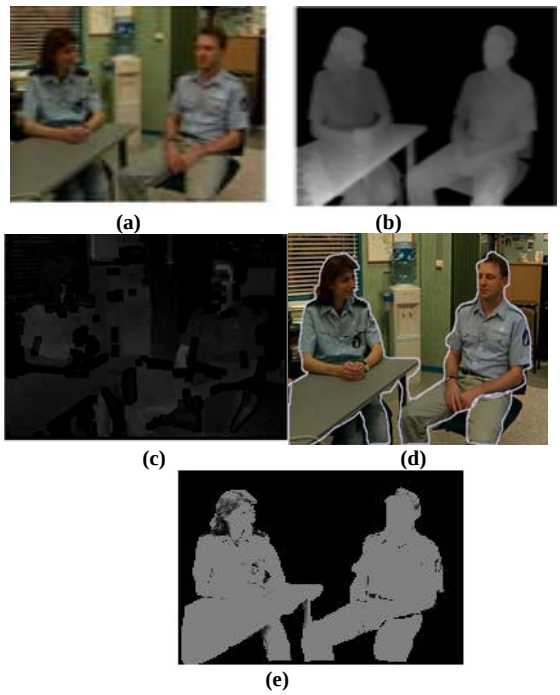

Figure 2. Experimental result of Interview Sequence: a) input image (b) True Depth map (c)Estimated depth map(our algorithm)(d) Updating with defocus cue (e) Final estimated depth map.

To evaluate performance of proposed algorithm, the algorithm is compared with 3 different algorithms [36], [37] and [38] as shown in Fig.3.

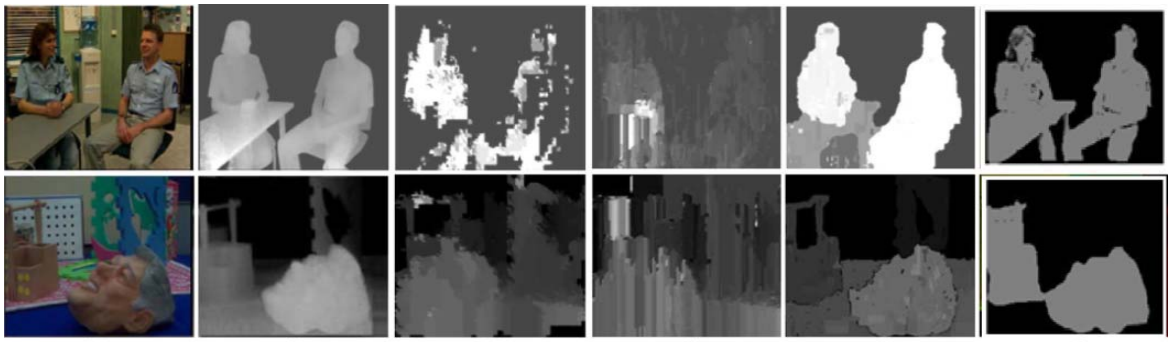

Figure 3. Experimental Evaluation of estimated depth –*Interview* and *Orbi* Sequences.a)Input sequences b)true depth map c) & d) results of algorithm in[36] and [37] e) depth by [38] and f) depth by Proposed algorithm.

B. Simulated and Experimental Result of Video sequences

We have demonstrated that proposed algorithm can also give stereoscopic view on real-time 2D video as well as video test sequences. Three depth maps based stereoscopic video sequence- *Poznan Street*, *Kendo*, and *Akko&Kayo* video sequences were used in the experiment which shows different conditions. These dataset have the true stereoscopic view. The result of proposed experiment is compared with stereoscopic view. It compares the video captured by the stereoscopic camera. Initially we selected four monocular cues as motion, AP Cue, LP Cue and defocus cue. The depth from each individual cue is estimated and then fuses these depth maps to estimate depth map.Fig.4 shows the converted 3D Left view and right view (stereoscopic view) which can show 3D effect by wearing red-cyan glasses.

Figure 4. Experimental Result of our proposed system :Converted 2D to stereoscopic View

C. Subjective Evaluation of Proposed Algorithm

The subjective evaluation Test was performed on the proposed system. The experimental result obtained by proposed system is compared with true stereoscopic view generated by stereoscopic camera. This thesis selected Poznan Street, Kendo, and Akko Kayo video sequences for subjective evaluation. This evaluation test was done on these three video sequences. The depth map was generated using the left view of stereoscopic view available. In this parameters were taken as depth quality and visual comfort as described in ITU-R BT.500-13. The synthesized two-view were displayed on 3D monitoring display with active-shutter glasses. 15 volunteers were selected for subjective evaluation. These volunteers are professor, student and non-teaching faculty. The volunteers watched the stereoscopic video captured by the stereoscopic camera and stereoscopic view generated by the proposed algorithm. The volunteers watched the videos in random order and were informed to rate each video. The parameters were given as depth quality and visual comfort. The depth quality was accessed by five-segment scale as Excellent (80-100), Good (60-80), Fair(40-60), Poor(20-60) and Bad(0-20) [34]. The visual comfort analysis was performed by comparing generated depth map (proposed algorithm) and true depth map and scaled as very comfortable(80-100),Comfortable(60-80), Mildly comfortable(40-60), Comfortable (20-40), and Extremely uncomfortable (0-20). The video sequences were shown randomly and repeated multiple times randomly to ensure a better result.

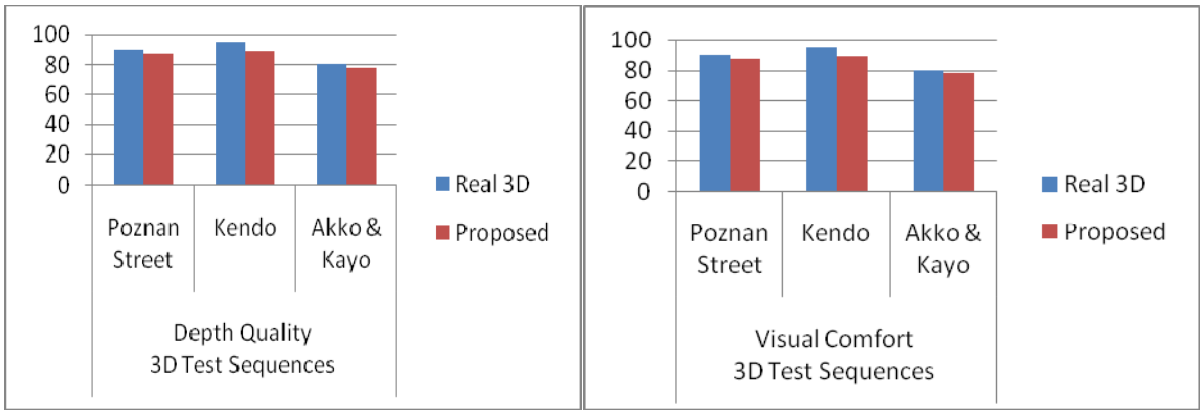

Figure 5. Subjective Evaluation Results

IV. CONCLUSION

This paper has presented an effective semi-automatic 2D-to 3D conversion algorithm. The proposed algorithm uses selected four monocular cues as AP, LP, Motion, and Defocus cue. Each cue has its own advantages: Motion cue gives information to differentiate the foreground and background information, defocus cue is better to estimate the mid-range depth, the aerial perspective cue can give the estimation of depth at long distance object, Linear perspective cue plans the basic depth to the scene. The performance of proposed algorithm is evaluated by comparing with existing algorithms. The experimental results shows that estimated depth map are very close to true depth map. The subjective assessment method was performed on proposed algorithm. The subjective evaluation test result shows that proposed system has good performance. In future, we may integrate a system such that 3D can be viewed with naked eyes.

ACKNOWLEDGMENT

Authors wishes to acknowledge to Dr.S.M.Gulhane, Head, E&TC, JDIET, Yavatmal and Dr. A.W. Kolhatkar, Principal, JDIET, Yavatmal.

References

[1] Zhebin Zhang,Chen Zhou, “Interactive Stereoscopic Video Conversion”,IEEE Transaction on Circuits and Systems for Video Technology,Vol. 6,Jan.2013.
[2] Yeong-KangLai, Yu-FanLai, Ying-Chang Chen, “An effective Hybrid Depth-Generation Algorithm for 2D-to-3D Conversion in 3DDisplays”, Journals of display technology, Vol.9, no.3, March 2013.
[3] Chung-Te Li, Yei-Chieh Lai,”Brain-inspired framework for fusion of multiple depth cues, accepted for publ., IEEE journal, 2012.
[4] I. A. Ideses, L. P. Yaroslavsky, B. Fishbain, and R. Vistuch , “3D from compressed 2D video,” in Proc. SPIE: Stereoscopic Displays and Applications XIV, vol. 6490, 64901C, 2007.
[5] Shaojie Zhuo and Terence Sim, "Recovering Depth from a Single Defocused Image", Pattern Recognition, 2011.
[6] P. Favaro and S. Soatto, “A geometric approach to shape from defocus,”IEEE Trans. Pattern Anal. Mach. Intell., vol. 27, no. 3, pp. 406–417, 2006.
[7] J. Park and C. Kim, “Extracting focused object from low depth-of-field image sequences,” in Proc. SPIE Visual Communications and Image Processing, vol. 6077, pp. 607710-1–607710-8, Jan. 2006.
[8] W. J. Tam, A. S. Yee, J. Ferreira, S. Tariq, and F. Speranza, “Stereoscopic image rendering based on depth maps created from blur and edge information,” in Proc. SPIE: Stereoscopic Displays and Applications XII,vol. 5664, pp. 104–115, 2005.
[9] W. J. Tam, C. Vázquez, and F. Speranza, “Three-dimensional TV: A novel method for generating surrogate depth maps using color information,” in Proc. SPIE: Stereoscopic Displays and Applications XX,vol. 7237, 2009.
[10] Lian Zhang, Carlos Vazquez, and Sebastian Knorr, “3DTV Content Cration: Automatic 2D-to-3D Video Conversion,”IEEE Trans. on broadcasting, Vol.57, No.2, June 2011.
[11] Janusz Lonrad, Meng Wang, “Learning –based, automatic 2D-to-3D image and Video Conversion,”accepted for publication in IEEE journals, 2013.
[12] J.Elder, S.Zucker, “Local scalecontrol for edge detection and blur estimation”, IEEE Trans.Pattern Anal. Mach.Intell.20 (7), 19998, 699-716.
[13] S.Bae,F.Durand,“Defocus magnification “,Proc.Eurographics,2007,pp. 571-579
[14] Y.W.Tai, M.S.Brown, “Single image defocus map estimation using local contrast prior, Proc ICIP, 2009.
[15] Diego Cheda, “Monocular depth cues in computer applicaation”, PhD Thesis, 2012.
[16] Q.Wei, “Converting 2D to 3D:A Survey”,Research Assignment,2012
[17] Rafael C. Gonzalez. Richard E, M, Woods, Pearson education, "Digital Image processing", 2009.
[18] S.Jayram,McGraw Education, "Digital Image Processing”,2013
[19] M. Dimiccoli, J. Morel, and P. Salembier, “Monocular Depth by Nonlinear Di˜ usion,” Indian Conf. Comput. Vision, Graphics & Image Process., pp. 95–102, 2008
[20] A.Criminisi, “Single-View Metrology: Algorithms and Applications,” in DAGM Symp. on Pattern Recognit., pp. 224–239,2002
[21] S. Battiato, S. Curti, M. La Cascia, M. Tortora, and E. Scordato, “Depth Map Generation by Image Classif cation,” in Soc. of Photo-Opt. Instrum. Eng. Conf.Ser., vol. 5302, 2004, pp. 95–104.
[22] G. Wang, Z. Hu, F. Wu, and H. Tsui, “Single View Metrology from Scene Constraints,” Image Vision Comput. vol. 23, no. 9, pp. 831–840, 2005.
[23] J. Huang and B. Cowan, “Simple 3D Reconstruction of Single Indoor Imagewith Perspective Cues,” in Can. Conf. on Comput. and Rob. Vision, 2009, pp.140–147.

[24] H. Chen and T. Liu, "Finding Familiar Objects and their Depth from a Single Image," in IEEE Int. Conf. Image Proc., 2007, pp. 389–392.
[25] A. Torralba and A. Oliva, "Depth Estimation from Image Structure," IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, no. 9, pp. 1226–1238, 2002.
[26] K. He, J. Sun, and X. Tang, "Single Image Haze Removal Using Dark Channel Prior," in IEEE Conf. Comput. Vision and Pattern Recognit., 2009, pp. 1956–1963.
[27] D. Hoiem, A. Efros, and M. Hebert, "Automatic Photo Pop-Up," ACM Trans.on Graphics, vol. 24, no. 3, pp. 577–584, 2005.
[28] A. Saxena, M. Sun, and A. Ng, "Make3D: Learning 3D Scene Structure from a Single Still Image," IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, no. 5, pp. 824 –840, 2009.
[29] V. Nedovic, A. Smeulders, A. Redert, and J. M. Geusebroek, "Stages As Models of Scene Geometry," IEEE Trans. Pattern Anal. Mach. Intell., vol. 32, no. 9,pp. 1673–1687, 2010.
[30] M. Dimiccoli, J. Morel, and P. Salembier, "Monocular Depth by Nonlinear Diffusion," Indian Conf. Comput. Vision, Graphics & Image Process., pp. 95–102, 2008.
[31] Baoping Li,Long Ye, "Multi-cue Fusion Based Depth Map generation from 2D video Frames", Journals of Information & Computational Scoences,2015
[32] Sung-Fang Tsai,Chao-Chung Cheng, "A Real-Time 1080p 2D-to-3D Video Conversion System," IEEE Transaction on Consumer Electronics,Vol.57,No.2,May 2011.
[33] C.L. Zitnick, S.B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski, "High-quality video view interpolation using a layered representation," ACM SIGGRAPH and ACM Trans. on Graphics, Los Angeles, CA, Aug. 2004, pp. 600-608.
[34] C. T. E. R. Hewage, S. T. Worrall, S. Dogan, S.Villette, A. M. Kondoz, Quality evaluation of color plus depth map-based stereoscopic video, IEEE J. Selected Topics in Signal Process., Vol. 3,No. 2, Apr. 2009, 304-318
[35] "Methodology for the Subjective assessment of the quality of television pictures," International Telecom union/ITU Radio Communication Sector,ITU-R BT.500-13,2012.
[36] M.T.Pourazad, P.Nasiopoulos, "H.264-based scheme for 2D to 3D Conversion",IEEE Transaction on consumer Electronics ,Vol.55,No.2,May 2009
[37] Yue Feng, Jinchang Ren and Jianmin Jiang, "Object-based 2D-to 3D Conversion for effective Stereoscopic content Generation in 3D-TV Applications", IEEE Transaction on Broadcasting, Vol.57, No.2, June 2011.

AUTHOR PROFILE

Mr. R. M. Mulajkar received his BE & ME degree in Electronics Engineering in 2006 & 2011 resp. He is currently a Research scholar in Research Lab, E&TC department, JDIET, Yavatmal, (Sant Gadge Baba Amravati University, Amravati), India. His research interests include Image and Video Processing.

Dr.(Mrs.) V. V. Gohokar is working as Associate Professor in Department of E&TC, M.I.T., Pune, India. She has teaching experience of more than 24 years in Engineering colleges. Her areas of interest include Image and Video Processing, Microwave Communication and Internet of Things.