https://doi.org/10.2352/ISSN.2470-1173.2018.2.VIPC-265

Generation of Stereoscopic Image Sequences from Monocular Videos Using Epipolar Geometry

Vasundhara Goyal, Dan Schonfeld; University of Illinois at Chicago; Chicago, IL, USA

Abstract

In this paper we focus on using single frame videos from a moving camera with pure horizontal translation. We make use of the fact that "3D shape reconstruction in Euclidean space is not necessarily required, but information of dense matching points is basically enough to synthesize new viewpoint images". The scene geometry and camera motion can be inferred by factorization of feature coordinates over a series of frames. We consider zero convergence angle and unit translation for parameterization of Fundamental Matrix for pure translation as the basis for predicting the pairs. We generated a cost function that selects the best matching stereo from the given set of frames using Fundamental Matrix Estimation (FME). The predicted scenes are compared with existing methods on the basis of their Peak Signal to Noise Ratio and graphically displayed. The generated frames are also compared using a disparity map and the results are explained in the paper.

Introduction

The major difference between a regular motion film and a 3D film is perception of depth in 3D, which is caused due to projection of 2 different images of the same scene on the eyes of the viewer. The principle of 3D projection is based on binocular vision. A human eye perceives two slightly different views in real life that helps them give the necessary depth perception to be able to gauge distance and shape.

3D films have been in existence since 1915, but due to requirement of costly hardware and production equipment the production of 3D is limited. This research aims at using such existing regular motion films and converting them to 3D using the given frames and the information they provide in the form of various angles of occlusion and shadows (plenoptic structures) [16]. This also gives an ability to manually increase or decrease the vergence angle between the 2 frames [2] which is often not possible while using a conventional 3D camera.

In this paper we focus on using single frame videos with a restriction to camera motion in horizontal direction (assume intrinsic parameters if unknown, for the ease of calculation), to obtain stereopair for each frame. The existing approach for obtaining stereopairs from monocular videos are semi-automatic and require extra information in the form of depth maps, heat maps and disparity maps [3], [4], [5]. The method mentioned in [4] makes use of information from camera sensors to predict distance between the two frames. A few approaches like [6] are also prediction based approaches, but focus more on warping the image using suitable Homography. Other methods like [7], [8] are based on Markov Random Fields and useful for generating 3D models or Depth Maps for computer vision applications.

Our method uses the fact that "3D shape reconstruction in Euclidean space is not necessarily required, but information of dense matching points is basically enough to synthesize new viewpoint images" [15]. The scene geometry and camera motion can be inferred by factorization of feature coordinates over a series of frames [9]. This method is used more for calibration of stereo pairs and less for stereo pairs generation and hence is a relatively unexplored area. We take into consideration the unknown camera parameter of previously captured 2D video. With proper fine tuning of the epipolar equation, the effect of assumed intrinsic parameters on the data is reduced and restricted to predicting only the direction of motion of camera.

The rest of the paper is structured as follows. In section II we define Fundamental Matrix in terms of Rotation and Translation between two frames. In section III we generalize the equation for the case of mobile camera moving horizontally. Section IV discusses the process of estimating the best possible stereo pair. The Experimental results are exhibited in Section V and VI.

Stereo Vision

Calibration of stereoscopic images comes from a generalized theory that stereo pairs have a set of points that form Image of the Absolute Conic. This means, these set of points are invariant under Euclidean Transformation. Locii of these points gives the equation for Fundamental Matrix.

$$ax_1^2 + bx_1x_2 + cx_2^2 + dx_1x_3 + ex_2x_3 + fx_3^2 = 0 \tag{1}$$

Let x and x' represent coordinates of matched features in the left and right images respectively. Then as per two view geometry of the camera [10]

$$x'^T Fx = 0 \tag{2}$$

where F is the fundamental matrix defined as

$$F = K^T TRK' \tag{3}$$

Here, R is the pure rotation between the left and the right frame. T embodies the translation vector between left and the right frame and is a unit vector. K and K' represent the intrinsic matrix for left and the right frames respectively [11].

Since, the y axis of both the images is parallel and orthogonal to baseline, the matrices for R, T and K can be written in the form

$$R = \begin{bmatrix} cos\theta & 0 & -sin\theta \\ 0 & 1 & 0 \\ sin\theta & 0 & cos\theta \end{bmatrix} \tag{4}$$

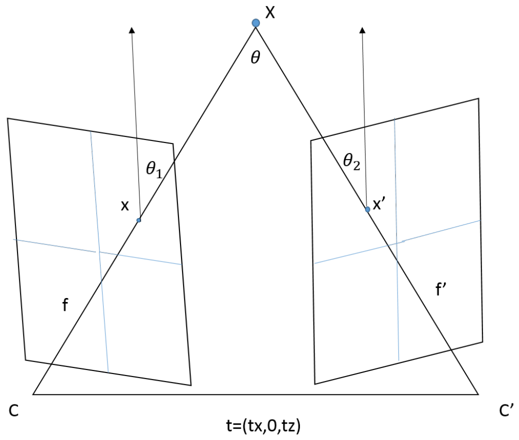


Figure 1. *A basic representation of two view geometry. θ is the convergence angle between the line joining two camera centresand the point object.*

$$T = \begin{bmatrix} 0 & -t_z & 0 \\ t_z & 0 & -t_x \\ 0 & t_x & 0 \end{bmatrix} \tag{5}$$

$$K = \begin{bmatrix} 1 & 0 & -x_0 \\ 0 & 1 & -y_0 \\ 0 & 0 & -f \end{bmatrix} \tag{6}$$

Thus the fundamental matrix using (3),(4) and (5) can be written in the form

$$F = \begin{bmatrix} 0 & -t_z & t_z y'_0 \\ \phi & 0 & -x'_0\phi - f'\sigma \\ -y_0\phi & x_0 t_z - f t_x & x'_0 y_0 \phi + y'_0(f t_x - x_0 t_z) \\ & & +f' y_0 \sigma \end{bmatrix} \tag{7}$$

where $\phi = t_z cos\theta - t_x sin\theta$ and $\sigma = -t_x cos\theta - t_z sin\theta$.

Since the absolute depth is unknown, we stick to the general idea of unit baseline between each frame. There are 8 unknown parameter in F and hence we need atleast 8 feature coordinates to form 8 equations for solving the unknown.

Convergence Angle Parametrization

The angular rotation θ can be represented as the sum of the two vergence angles θ_1 and θ_2, as shown in Figure 1. Employing the parametrization techniques of [11], we get

$$t_x = cos\theta_1 \qquad t_z = sin\theta_1 \tag{8}$$

For a Horizontal Moving Camera, the focal length of all the frames is the same. Thus the fundamental matrix reduces to

$$F_h = \begin{bmatrix} 0 & -sin\theta_1 & y'_0 sin\theta_1 \\ -sin\theta_2 & 0 & x'_0 sin\theta_2 + f' cos\theta_2 \\ & (x_0 sin\theta_1 & (-x'_0 y_0 (sin\theta_1 + sin\theta_2) \\ y_0 sin\theta_2 & -f cos\theta_1) & +y'_0 f(cos\theta_1 - cos\theta_2)) \end{bmatrix} \tag{9}$$

Assuming, $\theta_1 = 0$ and $\theta_1 \sim \theta_2$, we get the ideal case Fundamental matrix as

$$F_h = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & f' \\ 0 & -f' & 0 \end{bmatrix} = -f' \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{bmatrix} \tag{10}$$

This has already been stated under ”Fundamental matrix arising from special cases” in the book by Hartley and Zisserman [10]

Stereo Pair Prediction

This method proposes to generate better estimations using information obtained from camera motion.

Feature Detection and Matching

A number of feature detection methods including SIFT, SURF, BRISK, FAST etc., can be used to detect and match features of two images. We have used SURF Feature detection for the purpose of this project. SURF is partly influenced by SIFT yet more robust and computationally low cost to implement [6].

For feature matching, we use an exhaustive search match. Each pair-wise feature is calculated for the distance between the two. The feature pair with minimum distance SAD (Sum of Absolute Distance) is selected. Location of these matched features serve as image coordinates.

Camera Parameter Assumptions

We used videos captured using a SONY HDR-CX405 as the input. The camera has a sensor of 36mmx24mm and a focal length of 57mm. For ease of testing the results, we used 2 cameras of the same configuration to captue the videos with a side-by-side view which were meant for a 3D cardboard viewing. We also tested our method with video clips taken from Youtube and other online platforms with a side-by-side view.

Since the camera is moving in the horizontal direction, camera offsets are taken as the centroid of the input frame.

The final intrinsic matrix will be of the form

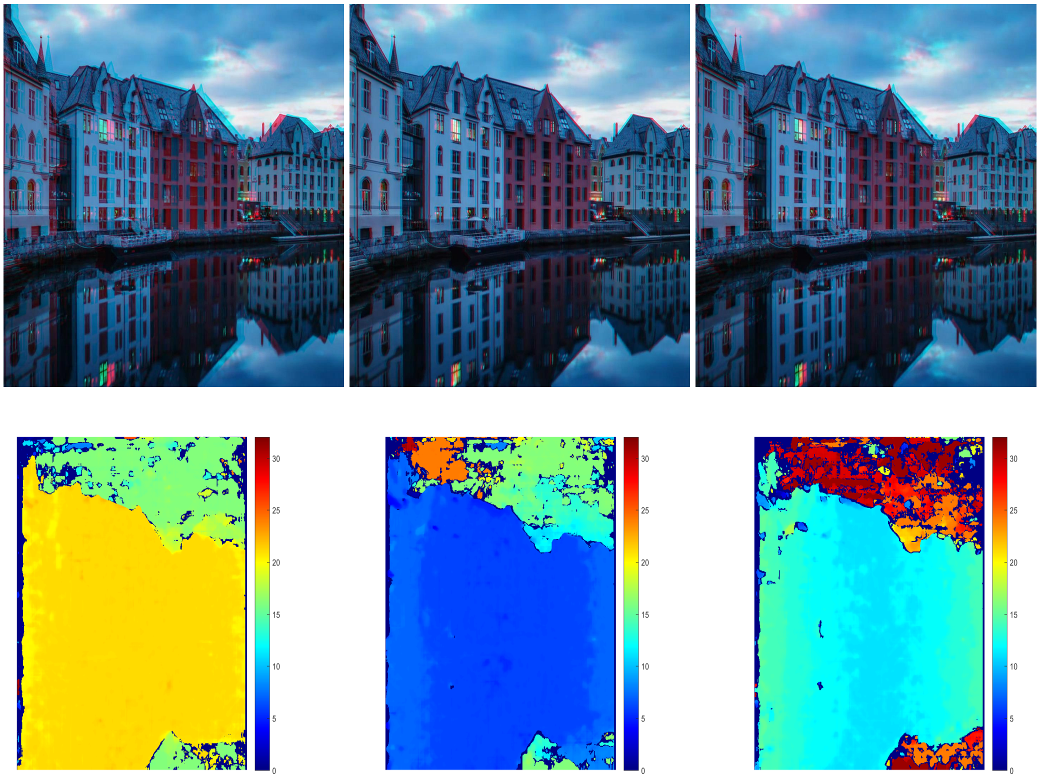
$$K = \begin{bmatrix} 1 & 0 & -size(image,2)/2 \\ 0 & 1 & -size(image,1)/2 \\ 0 & 0 & -18 * size(image,2)/57 \end{bmatrix} \tag{11}$$

Predicting the Stereo Pairs

We take the input as a sequence of 5 to 10 frames based on the camera speed. This is because a larger frame range would invariably increase the baseline which will create unwanted disparity. Also, over a large frame range the difference in the information captured is drastic, thereby causing false feature matching [12].

Let each frame be denoted by $I[i]$. We assume the first frame of the sequence to be the left frame $I^L[i] = I[i]$. The coordinates of each i^{th} frame are represented as X_i. Then the Epipolar equation between the i^{th} and j^{th} frame representing the same scene, is given by

$$X_j^T F_{i,j} X_i = 0 \tag{12}$$



Ground Truth
SSIM=1

PSNR=16.62 dB
SSIM=0.70

PSNR=17.75 dB
SSIM=0.74

Figure 2. *Anaglyphs and Disparity Maps of sequences taken from a UnivSt Dataset. Disparity in red shows closer objects and blue shows far away objects. (From Left) Stereo from Stereo Camera, Estimated Stereo Using Camera Tracking (CT) [17], Estimated Stereo Using Fundamental Matrix Estimation (FME)*

Signal to Noise Ratio and Similarity Index of Predicted Right Frames for video sequences shown in Fig. 3

PSNR of Side-by-Side YouTube Sequence Fig.3.				
Input Frame	PSNR Base	PSNR Proposed	SSIM Base	SSIM Proposed
'2454'	16.45	17.00	0.68	0.70
'2456'	16.51	17.01	0.68	0.70
'2457'	16.53	17.08	0.68	0.70
'2458'	16.56	17.71	0.68	0.73
'2467'	16.67	17.15	0.69	0.71
'2468'	16.63	17.11	0.69	0.70
'2469'	16.67	17.82	0.69	0.73
'2470'	16.69	17.85	0.69	0.73
'2472'	16.67	17.17	0.69	0.71

From (9), we can conclude that the Fundamental matrix between the i^{th} and the j^{th} frame, for a horizontal moving camera with close to no rotation is of the form,

$$F_{i,j} = \lambda_{i,j} \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{bmatrix} \tag{13}$$

Signal to Noise Ratio and Similarity Index of Predicted Right Frames for video sequences shown in Fig. 4

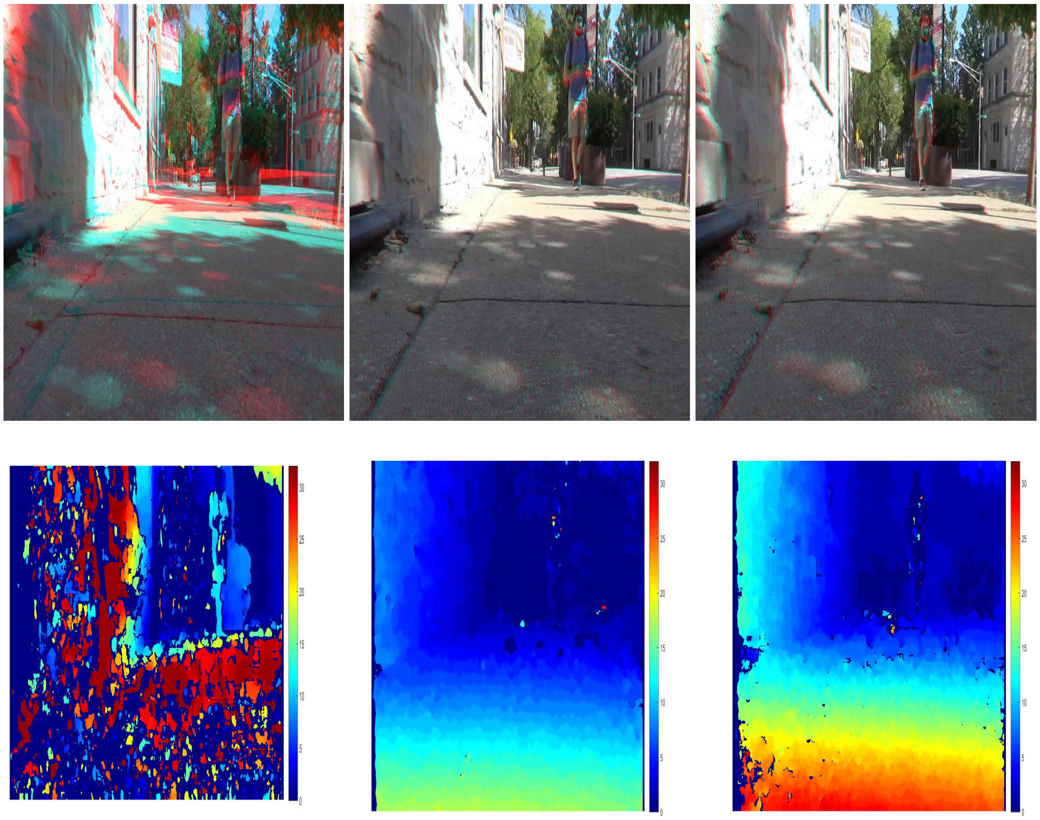
PSNR of Hand-Held StereoCamera Sequence Fig.4.				
Input Frame	PSNR Base	PSNR Proposed	SSIM Base	SSIM Proposed
'936'	13.59	13.96	0.53	0.54
'937'	13.61	13.97	0.53	0.54
'941'	13.56	13.99	0.53	0.54
'943'	13.54	13.74	0.52	0.53
'944'	13.51	13.95	0.52	0.53
'946'	13.47	13.71	0.52	0.53
'948'	13.42	13.68	0.51	0.52
'949'	13.40	13.62	0.51	0.52
'950'	13.35	13.76	0.51	0.52

such that $j = i+1, i+2 ... i+n$ and $n = 5$ or 10

The idea behind finding the perfect right frame is selecting the pair that minimizes the distance between F_h (8) and $F_{i,j}$ (12) matrix. This implies,

$$I^R[i] = I[min_j\{(-f') - \lambda_{i,j}\}] \tag{14}$$

$$I^R[i] = I[max_j\{\lambda_{i,j}\}] \tag{15}$$



Ground Truth
SSIM=1

PSNR=14.42 dB
SSIM=0.42

PSNR=14.55 dB
SSIM=0.43

Figure 3. *Anaglyphs and Disparity Maps of sequences taken from a Norway Timelapse [18]. Disparity in red shows closer objects and blue shows far away objects. (From Left) Stereo from Stereo Camera, Estimated Stereo Using Camera Tracking (CT) [17], Estimated Stereo Using Fundamental Matrix Estimation (FME))*

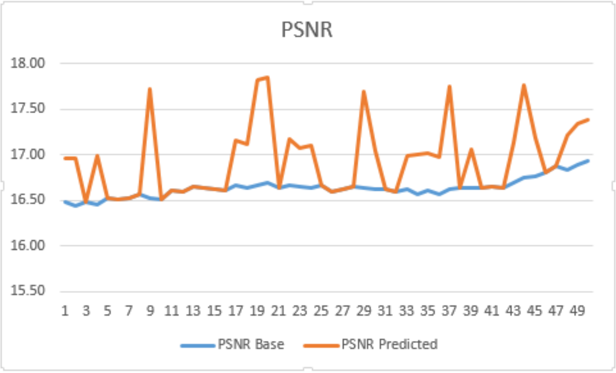


Figure 4. *PSNR of YouTube Sequence consisting of 90 frames: Red plot shows the PSNR of Proposed Method and Blue plot shows the PSNR of Existing Method.*

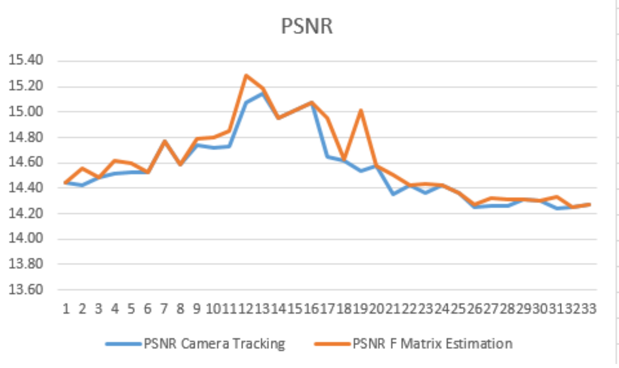


Figure 5. *PSNR of Handheld Camera Sequence consisting of 90 frames: Red plot shows the PSNR of Proposed Method and Blue plot shows the PSNR of Existing Method.*

Experimental Results

This method has been tested with video clips and image sequences from several monocular videos for visual plausibility. Besides testing it for visual plausibilty, we also tested it with different side-by-side 3D video sequence taken from YouTube [13] and stereo videos captured using two uncalibrated video cameras. The side by side view is separated in dual stream to get separate left $\Omega^L[i]$ and right streams $\Omega^R[i]$ and the left stream is selected as the input.

Figure 2 show anaglyphs created from a YouTube video clip and 3 was created from a video captured using a SONY HDR Video Camera. The left section of the image shows the original stereo anaglyph sequence and the disparity between them, the middle section shows anaglyph created using an existing method [6] and the right section of the image shows the estimated stereo anaglyph sequence using the proposed method. If you observe the original anaglyph and the proposed method anaglyph, you will notice that both images show the objects present in the scene at nearly the same distance from the viewer. A typical disparity map represents closer objects in gradients of red and farther objects in gradients of blue. The darker the blue the farther the object is. In the results shown in figure 2 and 3, the disparity map generated using the proposed method shows a better distinction between closer and far away objects. A few minor glitches can be rectified by motion estimation using particle filter [14]. This will be discussed in the future approaches to this problem.

Table 1 and 2 show the peak signal to noise ratio and structural similarity index of the predicted right frame with respect to the actual right frame (Equation 16).

$$PSNR = 10log(\frac{\Omega^L[max_j\{\lambda_{i,j}\}]^2}{\Omega^L[max_j\{\lambda_{i,j}\}] - \Omega^R[i]}) \qquad (16)$$

A good signal is indicated by a signal to noise ratio of 20db or above. Since the two frame sequences being compared for Signal to Noise Ratio are captured from different positions by different camera lenses, the slight difference in the information captured accounts for high noise index.

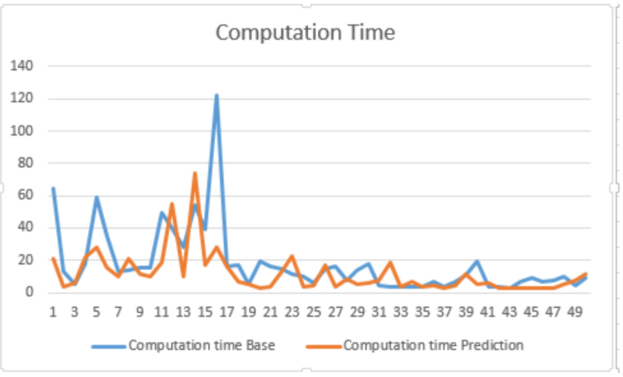


Figure 6. Computation Time: Red plot shows the time of computation per frame of Proposed Method and Blue plot shows the time of computation per frame of Existing Method.

The graph in figure 4 and 5 shows the Signal to Noise Ratio Analysis of the existing method and the proposed method.

Conclusion

In this paper, we present an automatic approach to predict stereo pairs from monocular video frames. We make use of a generalized equation of Brooks [11], on Fundamental Matrix for Horizontal Moving Cameras. The equation takes into account the convergence angle between two image frames. We further parametrize it for our case and get an ideal case solution. The most suitable stereopair is estimated using an equation that minimizes the distance between ideal case and estimated frames. The results are tested for Peak Signal to Noise ratio with respect to the original right frame and displayed. We also spoke about using particle filter as an extended approach to this problem in future.

References

[1] M. Peris, S. Martull, A. Maki, Y. Ohkawa, and K. Fukui, Towards a simulation driven stereo vision system, in Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), Nov 2012, pp. 10381042.

[2] H. Bay, T. Tuytelaars, and L. Van Gool, SURF: Speeded Up Robust Features. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 404417.

[3] P. C. N. Jr, M. Mukunoki, M. Minoh, and K. Ikeda, Estimating camera position and orientation from geographical map and mountain image, in 38th Research Meeting of the Pattern Sensing Group, Society of Instrument and Control Engineers, 1997, pp. 916.

[4] C. Holzmann and M. Hochgatterer, Measuring distance with mobile phones using single-camera stereo vision, in 2012 32nd International Conference on Distributed Computing Systems Workshops, June 2012, pp. 8893.

[5] S. Yin, H. Dong, G. Jiang, L. Liu, and S. Wei, A Novel 2D-to- 3D Video Conversion Method Using Time-Coherent Depth Maps, Sensors, vol. 15, no. 7, pp. 15 24615 264, 2015.

[6] X. Qin, Stereoscopic Video Synthesis from a Monocular Video, vol. 13, no. 4, pp. 111, 2007.

[7] A. Saxena, M. Sun, and A. Y. Ng, Learning 3-D Scene Structure from a Single Still Image, 2007 IEEE 11th International Conference on Computer Vision, pp. 18, 2007.

[8] 3-D depth reconstruction from a single still image, International Journal of Computer Vision, vol. 76, no. 1, pp. 5369, 2008.

[9] C. Tomasi and T. Kanade, Shape and motion from image streams: a factorization method. Proceedings of the National Academy of Sciences of the United States of America, vol. 90, no. 21, pp. 97959802, 1993.

[10] R. I. Hartley and A. Zisserman, ”Multiple View Geometry in Computer Vision”, 2nd ed. Cambridge University Press, ISBN: 0521540518, 2004.

[11] M. J. Brooks, L. De Agapito, D. Q. Huynh, and L. Baumela, Direct methods for self-calibration of a moving stereo head, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 1065, no. April, pp. 413426, 1996.

[12] A Mutiple-Baseline Stereo, pp. 353363, 1993.

[13] T. Borel and D. Doyen, 3D Display Technologies. John Wiley Sons, Ltd, 2013, pp. 295312.

[14] J. Yang, D. Schonfeld, C. Chen, and M. Mohamed, Online video stabilization based on particle filters, in 2006 International Conference on Image Processing, Oct 2006, pp. 15451548.

[15] Kimura, Makoto, and Hideo Saito. ”3D reconstruction based on epipolar geometry.” IEICE TRANSACTIONS on Information and Systems 84.12 (2001): 1690-1697.

[16] Adelson, E. H. and Bergen, J. R.: ”The plenoptic function and the elements of early vision.” In Computational Models of Visual Processing , pages 320. MIT Press, 1991.

[17] X. Qin, Stereoscopic Video Synthesis from a Monocu- lar Video,

vol. 13, no. 4, pp. 111, 2007.

[18] M.Rustad. NORWAY-A Time-Lapse Adventure 4K. Youtube. [Online]. Available:https://www.youtube.com/watch?v=Scxs7L0vhZ4

Author Biography

Vasundhara Goyal received her B.Tech in Electronics and Communication Engineering from the National Institute of Technology, Bhopal, India in 2012 and her Masters in Electrical Engineering from University of Illinois at Chicago in 2017. She did her Master's Thesis in Stereoscopic Vision and her work focuses on creating 3D videos/image sequences using Epipolar Geometry. Her research interests include Signal Processing, Image Processing, Video Processing and Robotics.

Dan Schonfeld received his B.S. in Electrical Engineering & Computer Science from University of California at Berkeley (1986) and his Ph.D. in Electrical & Computer Engineering from The Johns Hopkins University (1990). He is currently Professor in Department of Electrical & Computer Engineering, Department of Computer Science, Department of Bioengineering. He is also Co-Director of the Multimedia Communications Laboratory (MCL) and member of the Signal and Image Research Laboratory (SIRL). He has authored nearly 200 papers some of which include papers that won Best Paper Award in ACM 2010, Best Student Paper Award in IEEE International Conference on Image Processing 2006, 2007 and Best Student Paper Award in VCIP 2006. His current research interests are in multi-dimensional signal processing, image and video analysis, computer vision, and genomic signal processing.