Learning-Based Automatic 2D-to-3D Image Conversion Using Local Point Transformation and Depth Estimation

Miss Simantinee Shinde[1], Prof. (Dr.) J.A. Shaikh[2]
[1]*Student , Padmabhooshan Vasantraodada Patil Institute of Technology, Budhgaon*
[2]*Faculty, Padmabhooshan Vasantraodada Patil Institute of Technology, Budhgaon*

Abstract: **The work presents a simple automatic 2D to 3D image conversion using Local Point Transformation Learning. The method uses the NY Depth Dataset for training and testing. Different local image features such as grayscale intensity, edge information and mean filtered values are extracted from the image. These features are used to train a regression model for depth prediction. The predicted depth map is improved using bilateral filtering to reduce noise and preserve edges. Using depth information, a right stereo image is generated from the original image. Finally, 3D anaglyph image is created by combining the left and right views. The performance of the method is evaluated using RMSE, MAE, PSNR and SSIM parameters. The results show that the proposed method can generate good quality depth maps and realistic 3D images with low computational.**

*Keywords :***2D to 3D conversion, depth estimation, Local Point Transformation, NYU depth dataset, stereo image generation.**

I. INTRODUCTION

Today's there is an urgent need to convert the existing 2D content to 3D. A typical 2D-to-3D conversion process consists of two steps: first, depth estimation for a given 2D image and second is depthbased rendering of a new image in order to form a stereo pair. The depth based rendering step is well understood and algorithms exist that produce good quality images, the challenge is in estimating depth from a single image.There are two basic approaches to 2D-to-3D conversion: one that requires a human operator's intervention and one that does not need any human operator's intervention. In the former case, the so-called semi-automatic methods have been proposed where a skilled operator assigns depth to various parts of an image or video. Based on this sparse depth assignment, a computer algorithm estimates dense depth over the entire image or video sequence.In the case of automatic methods, no operator intervention is needed and a computer algorithm automatically estimates the depth for a single image. To this effect, methods have been developed that estimate shape from shading, structure from motion or depth from defocus.

The work implemented in this paper based on learning a point mapping from local image attributes, such as color, spatial position to scenedepth at that pixel using a regression type idea. The demand for the efficient 2D to 3D image and video conversion increased due to popularity of the 3D display technologies in the entertainment, virtual reality and, multimedia applications. To generate the stereoscopic contents from conventional 2D media, several review and surveys were proposed related to the development of automatic and semiautomatic methods. The different depth estimation approaches such as motion analysis . optical flow, perspective geometry, saliency detection and machine learning based techniques for generation of realistic 3D outputs reviewed by *Shweta Patil et al.* [1] and *S. Patil et al* [2]. Similarly, *Divya K.P. et al* [3] discussed disparity map generation and the use of visual cues including color, edges, motion, and geometric features for estimating scene depth. Recent review studies by *S. Singla et al* [4] further highlighted the growing role of deep learning, triangulation, feature matching, and point cloud generation techniques in modern 3D reconstruction systems. Several researchers have focused on developing automatic depth estimation methods using local image features and machine learning techniques. *Victoria M. Baretto* [5] evaluated local point mapping and global nearest-neighbor approaches for automatic depth estimation, where pixel attributes such as color, motion, and spatial location were used to determine depth information with high computational efficiency. *Mujawar A. R. et al* [6] implemented a local point transformation algorithm capable of generating depth

information automatically while reducing computational complexity.

In another study, *N. Chahal et al* [7] proposed a manifold learning-based method using Locally Linear Embedding (LLE) for accurate depth map generation. Their approach preserved neighborhood relationships among pixels and produced depth maps that closely matched ground truth values. Recent advancements in 2D-to-3D conversion mainly focus on deep learning and depth map refinement techniques for improving 3D reconstruction quality. *J. L. Herrera et al* [8] proposed a machine learning-based system that estimates scene depth from single images by utilizing similarities between visual structures and trained color-depth databases. Furthermore, a deep learning-based 2D-3D Transformation Network using Generative Adversarial Networks (GANs) was introduced by *N. U. Islam* [9] to generate accurate 3D representations from single 2D images without requiring labelled datasets. Depth mapbased methods have also shown significant improvements in stereoscopic rendering quality. *S. Mathai et al.* [10] proposed a global depth map estimation technique using saliency and depth fusion methods, while *B. Pan et al.*[11] utilized RGB-D images along with advanced hole-filling and depth inpainting algorithms to enhance Depth ImageBased Rendering (DIBR). These studies demonstrate that modern automatic 2D-to-3D conversion systems increasingly rely on machine learning and deep learning approaches to improve depth accuracy, visual realism, and real-time performance.

II. METHODOLOGY

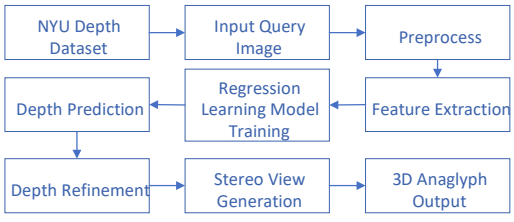

Figure 2.1 Workflow of implemented work

The workflow of the implemented work is as shown in Figure 2.1. The proposed system utilizes the NYU Depth V2 dataset consisting of RGB images and corresponding ground truth depth maps for supervised depth learning. Initially, the RGB query image is resized to 240 × 320 pixels to maintain dimensional consistency throughout the processing pipeline. During preprocessing, the RGB image is transformed into grayscale using the *rgb2gray()* function to reduce computational complexity and simplify intensity-based feature analysis. In the feature extraction stage, three local descriptors are computed for every pixel: grayscale intensity, gradient magnitude, and local mean filter response. The gradient magnitude is obtained using the *imgradient()* operator, which captures edge discontinuities and structural variations in the scene.

A 5 × 5 averaging filter generated using *fspecial(average,[5 5])* is applied through *imfilter()* to extract local neighborhood intensity statistics. These descriptors collectively form the feature vector representing local spatial characteristics of the image. For regression learning, feature vectors extracted from the first 50 training images are concatenated into the matrix *X*, while the corresponding depth values are stored in target vector *Y*. A linear regression model is then trained using the *fitrlinear()* function to establish a mapping between local image features and depth information. During inference, the same feature extraction procedure is applied to the query image, and the trained regression model predicts pixel-wise depth values. The predicted depth vector is reshaped into a 240 × 320 depth map representing the estimated scene geometry. Since regression-based estimation may introduce local inconsistencies and noise, bilateral filtering using *imbilatfilt()* is applied for depth refinement. This filtering process performs edge-preserving smoothing, which reduces noise while maintaining object boundaries and depth discontinuities. The refined depth map is normalized and converted into disparity information for stereo synthesis. Disparity values are computed using the relation between normalized depth and the predefined maximum disparity value (*maxDisparity = 30*). Each pixel in the original image is horizontally shifted according to its disparity value to generate the right stereo view. Larger disparity shifts correspond to foreground objects, while smaller shifts represent background regions, thereby simulating binocular parallax. Due to pixel displacement, holes appear in the synthesized stereo image, which are compensated using a hole filling strategy that replaces missing pixels with corresponding pixels from the original image. Finally, a stereoscopic 3D anaglyph image is generated by combining the red channel of the original left image with the green and blue channels of the synthesized right image. The generated depth maps are

quantitatively evaluated using Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index (SSIM) to measure depth estimation accuracy and reconstructed image quality.

III. RESULT AND DISCUSSION

The performance analysis of the implemented Local Point Learning Transformation method for automatic 2D to 3D image conversion is presented in this section. The visual evaluation of the generated depth maps, stereo views, and 3D anaglyph images is carried using query images from the NYU Depth V2 dataset. Based on the comparative analysis of the predicted depth map with the corresponding ground truth depth maps, the quantitative performance parameters like RMSE, MAE, PSNR, SSIM, Mean Depth and Standard Deviation are computed. Results obtained demonstrate that the regression-based depth estimation approach has the ability to generate visually realistic and structurally accurate 3D representations from 2D images. Figure 3.1 shows the input query image from the NYU Depth V2 dataset. The image contains shelves, clothes and different background objects with varying spatial distances. This image is used for feature extraction and depth prediction using the Local Point Transformation Learning method.

Figure 3.1 Input RGB query image 1

Figure 3.2 presents the actual depth map obtained from the NYU Depth V2 dataset. Different color intensities represent varying object distances from the camera, where darker regions indicate closer objects and brighter regions indicate distant areas. The ground truth depth map is used for quantitative evaluation of the predicted depth results.

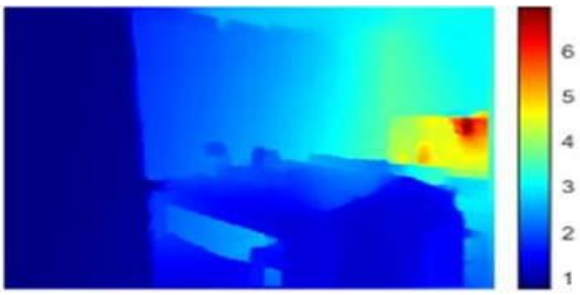

Figure 3.2 Ground Truth Depth Map for input query image

The depth map generated using regression based Local Point Transformation Learning model is as shown in Figure 3.3. The model estimates depth values using greyscale intensity, gradient magnitude and local mean filter features. The predicted depth map successfully identified major foreground and background regions of the scene.

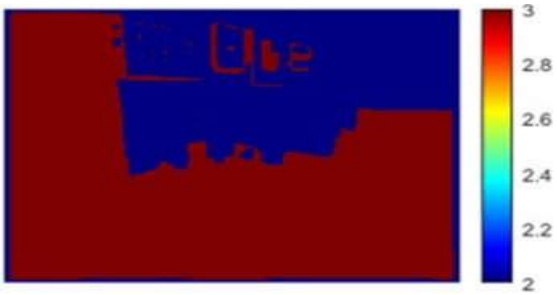

Figure 3.3 Predicted depth map using Local Point Transformation Learning for query image

Figure 3.4: Generated stereo right view for input query image

The synthesized stereo right image generated using disparity-based pixel shifting is illustrated in Figure 3.4. The disparity values are calculated from the normalized depth map using a maximum disparity of

30 pixels. The generated right view is used for stereoscopic visualization.

Figure 3.5: Final 3D anaglyph qutput for query image

The final red cyan anaglyph image created by combining the original left image with generated stereo right view is as shown in Figure 3.5.The anaglyph image provides depth perception when viewed using red-cyan 3D glasses, thereby creating a stereoscopic visualization effect.

Figure 3.6: Input RGB image for query image 2

Figure 3.6 shows the second query image selected from the NYU Depth V2 dataset. The indoor scene contains furniture, walls, and electronic objects with different depth variations. This image is used to evaluate the performance of the proposed depth estimation method on another scene.

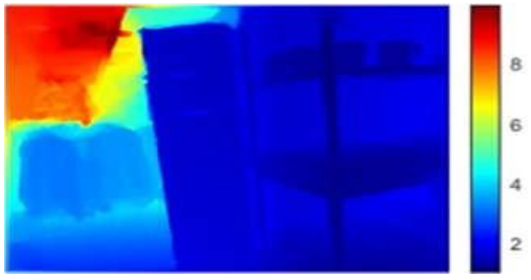

Figure 3.7: Ground truth depth map for query image 2

Figure 3.7 presents the original ground truth depth information corresponding to Query Image Index 200. The depth distribution clearly differentiates the foreground furniture from the distant background wall and surrounding objects.

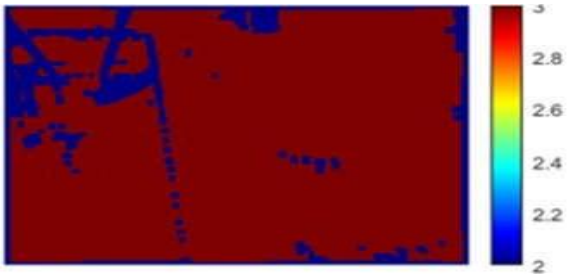

Figure 3.8: Predicted depth map using Local Point Transformation Learning for query image 2

Figure 3.8 illustrates the predicted depth map generated by the trained regression learning model. The estimated depth values preserve the major structural characteristics of the indoor scene and provide suitable depth information for stereo image synthesis.

Figure 3.10: Generated stereo right view for query image 2

Figure 3.10 presents the synthesized stereo right image produced using the estimated depth map. Pixel displacement based on depth values creates horizontal disparity required for stereoscopic image generation.

Figure 3.11: Final 3D anaglyph output for query image 2

Figure 3.11 shows the final 3D anaglyph image generated for Query Image Index 200. The produced stereoscopic image enhances depth perception and provides an effective 3D visualization of the indoor environment.

Table 1 Quantitative performance evaluation of Local Point Transformation Learning method

Query Image	RMSE	MAE	PSNR (dB)	SSIM	Mean Depth	Std Depth
Image 1	2.31	1.69	40.92	0.88	2.85	0.35
Image 2	1.50	1.31	44.85	0.87	2.61	0.48

The Table 3.1 presents the quantitative evaluation of the implemented Local Point Transformation Learning method for two different query images selected from the NYU Depth V2 dataset. The performance is analysed using standard depth estimation metrics including RMSE, MAE, PSNR, SSIM, Mean Depth, and Standard Deviation. The RMSE and MAE values are 2.3147 and 1.6922 for image 1 respectively indicates a moderate level of prediction error in depth estimation. However, the PSNR value of 40.9224 dB shows that the predicted depth map maintains good reconstruction quality compared to the ground truth depth. The SSIM value of 0.8847 confirms that the structural information of the scene is largely preserved. The mean depth of 2.8521 represents overall scene depth distribution, while the standard deviation of 0.3550 indicates relatively stable depth variation across the image.

For Image 2, improved performance is observed with lower RMSE (1.5058) and MAE (1.3148), indicating more accurate depth prediction compared to Image 100. The higher PSNR value of 44.8509 dB reflects better reconstruction quality and reduced noise in the predicted depth map. The SSIM value of 0.8796 shows strong structural similarity with the ground truth depth map. The mean depth of 2.6165 and standard deviation of 0.4862 suggest slightly higher depth variation due to more complex scene geometry.

Overall, the results demonstrate that the proposed method performs effectively for automatic depth estimation, providing reliable 3D reconstruction with acceptable accuracy and good visual consistency across different indoor scenes.

IV. CONCLUSION

The implemented Local Point Transformation Learning method successfully performed automatic 2D to 3D image conversion using local image feature s and regression-based depth estimation. The system generated accurate depth maps, stereo right views and visually effective 3D anaglyph images from single RGB images obtained from the NYU Depth V2 dataset. Quantitative evaluation showed satisfactory performance with low RMSE and MAE values, while PSNR values above 40 dB and SSIM values close to 0.88 confirmed good structural similarity between predicted and ground truth depth maps. The experimental results demonstrate that the proposed method provides a computationally efficient and practical solution for depth estimation and stereoscopic visualization

REFERENCES

[1] S. Patil and P. Charles, “Review on 2D to 3D Image and Video Conversion Methods,” in *National Conference on Emerging Trends in Advanced Communication Technologies (NCETACT-2015)*, *International Journal of Computer Applications (IJCA)*, 2015, pp. 14– 18.

[2] Sweta Patil and P. Charles, “Review on 2D-to3D Image and Video Conversion Methods,” in *2015 International Conference on Computing Communication Control and Automation (ICCUBEA)*, Pune, India, 2015, pp. 728–732.

[3] D. K. P. and S. Arun, “A Survey on 2D to 3D Image and Video Conversion Techniques,” *International Journal of Research in Advent Technology*, vol. 2, no. 12, pp. 99–103, Dec. 2014.

[4] S. Singla, P. Saini, K. Malhotra, M. Kukreja, G. Kiran, and R. Bhandari, “Exploration and Analysis of Various Techniques used in 2D to 3D Image and Video Reconstruction,” in *2024 International Conference on Emerging Innovations and Advanced Computing (INNOCOMP)*, Sonipat, India, 2024, pp. 8–15, doi: 10.1109/INNOCOMP63224.2024.00013.

[5] V. M. Baretto, “Automatic Learning based 2Dto-3D Image Conversion,” *International Journal of Engineering Research & Technology (IJERT)*, vol. 2, no. 13, NCRTS-2014, 2014. (IJERT)

[6] A. R. Mujawar and J. D. Nanaware, "Development and Implementation of Automatic 2D to 3D Image Conversion System," *International Journal of Advance Research in Science and Engineering (IJARSE)*, vol. 5, no. 8, pp. 1–6, Aug. 2016.

[7] N. Chahal, M. Pippal, and S. Chaudhury, "Automated Conversion of 2D to 3D Image Using Manifold Learning," in *2015 Fifth National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG)*, Patna, India, 2015, pp. 1–4, doi: 10.1109/NCVPRIPG.2015.7490002.

[8] J. L. Herrera, C. R. del-Blanco, and N. García, "A Novel 2D to 3D Video Conversion System Based on a Machine Learning Approach," *IEEE Transactions on Consumer Electronics*, vol. 62, no. 4, pp. 429–436, Nov. 2016, doi: 10.1109/TCE.2016.7838096.

[9] N. U. Islam and S. Lee, "Learning Typical 3D Representation from a Single 2D Correspondence Using 2D-3D Transformation Network," in *Proceedings of the 13th International Conference on Ubiquitous Information Management and Communication (IMCOM 2019)*, *Advances in Intelligent Systems and Computing*, pp. 440–455, May 2019, doi: 10.1007/978-3-030-19063-7_35.

[10] S. Mathai, P. P. Mathai, and K. A. Divya, "Automatic 2D to 3D Video and Image Conversion Based on Global Depth Map," in *2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, Madurai, India, 2015, pp. 1–4, doi: 10.1109/ICCIC.2015.7435781.

[11] B. Pan, L. Zhang, H. Yin, J. Lan, and F. Cao, "An automatic 2D to 3D video conversion approach based on RGB-D images," *Multimedia Tools and Applications*, vol. 80, no. 13, pp. 19179–19201, May 2021, doi: 10.1007/s11042-021-10662-0.