Depth Extraction from 2D Images

Ananth Bharathwaj T A, 21BCE1079 (GitHub)
Hitesh S D, 21BCE5051
Sundar Karthikeyan, 21BCE1025
Faculty: Dr. Geetha S

VIT Chennai
GitHub Repository

Abstract

Depth extraction from 2D images is a key problem in computer vision and multimedia processing, with many applications in 3D content creation, augmented reality, and autonomous systems. This paper presents a comprehensive literature review of the current state-of-the-art methods in depth extraction, focusing on three recent papers that use Convolutional Neural Networks (CNNs) and machine learning approaches to achieve real-time 2D/3D registration, novel 2D to 3D video conversion, and fully automatic 2D-to-3D video conversion. The paper also discusses the role of naturalism and realism in enhancing depth extraction processes, based on insights from another paper. The paper aims to provide a clear overview of the methodologies, contributions, challenges, and future directions of depth extraction from 2D images, as well as the implications of these technologies for various domains. The paper serves as a valuable resource for researchers, practitioners, and enthusiasts interested in the field of depth extraction.

1 Introduction

Depth extraction from 2D images is a crucial task in computer vision and multimedia processing, with widespread applications ranging from 3D content creation to augmented reality and autonomous sys-

tems. This document aims to provide a comprehensive review of the current state-of-the-art methodologies in depth extraction, with a particular focus on recent advancements highlighted in three influential papers: "A CNN Regression Approach for Real-time 2D/3D Registration," "A Novel 2D to 3D Video Conversion System Based on a Machine Learning Approach," and "Deep3D: Fully Automatic 2D-to-3D Video Conversion with Deep Convolutional Neural Networks." Additionally, we will explore insights from "Toward Naturalistic 2D-to-3D Conversion" to understand how naturalism and realism contribute to enhancing depth extraction processes.

In the pursuit of advancing depth extraction techniques, these papers employ cutting-edge technologies such as Convolutional Neural Networks (CNNs) and machine learning approaches. The focus of this literature review is to elucidate the methodologies proposed in these works, analyze their contributions, and synthesize the collective knowledge to understand the evolution of depth extraction from 2D images. As we delve into each paper, we aim to identify common themes, challenges, and potential directions for future research in this dynamic and rapidly evolving field.

Through this exploration, we hope to offer a valuable resource for researchers, practitioners, and enthusiasts interested in the intricacies of depth extraction from 2D images. The synthesis of these papers will provide insights into the advancements made, challenges faced, and the potential impact of these technologies on various applications, ultimately contributing to the broader conversation on the state and future directions of depth extraction in the realm of computer vision and multimedia processing.

2 Review of Literature

2.1 Paper 1

Title: An LSTM-Based Approach for Understanding Human Interactions Using Hybrid Feature Descriptors Over Depth Sensors [1]
Author(s): Manahil Waheed, Ahmad Jalal, Mohammed Alarfaj, et. al.
Topics Discussed: The paper discusses a novel approach for understanding human interactions by combining machine learning and deep learning techniques. It focuses on feature extraction from both 2D human silhouettes and 3D meshes using depth sensors. The proposed system utilizes LSTM (Long Short-Term Memory) networks for classification tasks.
Description: The paper presents a comprehensive human interaction recognition (HIR) framework de-

signed for various real-world applications such as behavior monitoring, security, and smart homes management. It outlines the process of feature extraction from 2D human silhouettes and 3D meshes, followed by classification using LSTM and softmax-based classifiers. The system achieves high accuracies across different datasets, demonstrating its effectiveness in understanding human interactions.

Relevance: Although the paper primarily focuses on human interaction recognition, it provides valuable insights into feature extraction from 2D images and depth sensors. The techniques discussed could potentially be adapted or extended to address problems related to depth extraction from 2D images, contributing to advancements in depth sensing technologies.

2.2 Paper 2

Title: Automatic 2D-to-3D Image Conversion Using Learning-Based Depth Estimation [2]

Author(s): Janusz Konrad, Meng Wang, Prakash Ishwar, Chen Wu, Debargha Mukherjee

Topics Discussed:

1. Comparison of different methods for 2D-to-3D image conversion. Introduction of two novel methods based on learning from examples: a local method and a global method.
2. Evaluation of proposed methods against state-of-the-art algorithms using performance metrics.
3. Discussion on the computational complexity and efficiency of the proposed methods.
4. Presentation of results and analysis, including comparisons with existing algorithms and insights into the limitations and advantages of the proposed approaches.

Description: The paper introduces two new methods for automatic 2D-to-3D image conversion, based on learning from examples. One method focuses on learning a point mapping from local image attributes to scene depth, while the other estimates the entire depth field globally using nearest-neighbor-based regression. The authors evaluate the performance of these methods against existing algorithms using various datasets and performance metrics. They discuss computational complexity and efficiency, presenting results that demonstrate the effectiveness of their approaches compared to state-of-the-art methods.

Relevance: This paper is highly relevant to the problem of depth extraction from 2D images. It introduces innovative approaches for automatically converting 2D images into 3D

representations, which inherently involve estimating depth information from 2D visual data. By comparing the proposed methods with existing algorithms and discussing their performance and computational efficiency, the paper provides valuable insights and potential solutions for improving depth extraction techniques from 2D images.

2.3 Paper 3

Title: Hierarchical Search and Learning in a Clustering-based Framework for 2D-to-3D Image Conversion [3]
Author(s): José L. Herrera, Carlos R. del Blanco, Narciso García
Topics Discussed:

1. 2D-to-3D image conversion techniques
2. Machine learning-based approaches
3. Database clustering for efficient similarity search
4. Depth map estimation and refinement
5. Comparative evaluation with state-of-the-art methods
6. Rendering of stereoscopic views

Description: The paper proposes a novel approach for converting 2D images to 3D using machine learning techniques. It introduces a hierarchical search method combined with database clustering to efficiently find similar images in a training dataset. Depth maps are estimated by fusing depth information from structurally similar images and refining them with segmentation-based filtering. The paper evaluates the proposed method against existing techniques and presents qualitative results of rendered stereoscopic views.
Relevance: This paper addresses the problem of depth extraction from 2D images by presenting a comprehensive method that leverages machine learning and database clustering for improved efficiency and accuracy. It offers insights into techniques for depth map estimation and refinement, which are crucial components in solving our problem. Additionally, the comparative evaluation with state-of-the-art methods provides valuable context for assessing the effectiveness of different approaches in the field of 2D-to-3D image conversion.

2.4 Paper 4

Title: Depth Transfer: Depth Extraction from Video Using Non-Parametric Sampling [4]
Author(s): Kevin Karsch, Ce Liu, and Sing Bing Kang
Topics Discussed:

1. Depth extraction from videos using non-parametric sampling.

2. Utilization of depth information for generating stereoscopic videos and 3D viewing.

3. Comparison with existing methods based on motion parallax and structure from motion.

4. Challenges and limitations in depth estimation from videos, including handling moving objects and airborne objects.

5. Discussion on algorithm robustness, motion segmentation, and error propagation.

6. Demonstration of the technique's applicability to single images and dynamic scenes.

7. Generation of time-coherent stereo sequences from inferred depth maps.

8. Application potential for converting 2D feature films into 3D.

Description: The paper presents a novel technique for extracting depth information from videos using non-parametric sampling, overcoming limitations of existing methods based on motion parallax and structure from motion. The approach is demonstrated to be effective for both single images and dynamic scenes, with the ability to generate visually pleasing stereoscopic videos from conventional 2D footage. The method relies on robust feature matching and depth inference, although it may encounter challenges with moving or airborne objects and error propagation.

Relevance: This paper addresses the problem of depth extraction from 2D images, offering a technique that extends to video content. It introduces a novel approach that could potentially enhance existing methods for depth estimation, particularly in scenarios where traditional methods based on motion parallax and structure from motion fail. The technique's ability to handle dynamic scenes and generate stereoscopic content aligns with the goal of extracting depth information from 2D images for various applications, including augmented reality, scene understanding, and image editing.

2.5 Paper 5

Title: A CNN Regression Approach for Real-time 2D/3D Registration [5]

Authors: - Shun Miao, Member, IEEE
- Z. Jane Wang, Senior Member, IEEE
- Rui Liao, Senior Member, IEEE

Topic Discussed: This paper explores a Convolutional Neural Network (CNN) regression approach for real-time 2D/3D registration. The authors aim to enhance the accuracy

and efficiency of the registration process by leveraging the capabilities of deep learning.

Description: The paper introduces a novel CNN regression method designed to predict transformation parameters directly for the real-time alignment of 2D and 3D data. The network architecture is tailored to capture both local and global features crucial for accurate registration. Training involves providing pairs of registered 2D and 3D data along with ground truth transformation parameters, enabling the CNN to learn the mapping for efficient and precise registration. Comprehensive experiments on diverse datasets, including medical images and computer-generated 3D models, showcase the superior performance of the CNN regression method in terms of accuracy and computational efficiency.

Relevance: This research is highly relevant in the context of 2D/3D registration, addressing the limitations of traditional methods and introducing a deep learning approach for real-time processing. The implications of this work extend to various domains, such as medical imaging, robotics, and augmented reality, where accurate and efficient registration is crucial. The paper's findings contribute to advancing applications in these fields and open up new possibilities for leveraging CNNs in dynamic and real-world scenarios.

2.6 Paper 6

Title: A Novel 2D to 3D Video Conversion System Based on a Machine Learning Approach [6]

Authors:

- José L. Herrera
- Carlos R. del-Blanco
- Narciso García

Topic Discussed: This paper introduces a novel 2D to 3D video conversion system that relies on a machine learning approach. The authors explore the application of machine learning techniques to enhance the process of converting two-dimensional videos into three-dimensional formats.

Description: The paper presents an innovative system designed for the conversion of 2D videos to immersive 3D experiences. The core of the system is based on machine learning methodologies, leveraging advanced algorithms to analyze and understand the depth information within 2D scenes. The authors detail the architecture of the system, including the training process, where the machine learning model learns to infer depth and

create a corresponding 3D representation from 2D video frames. Experimental results and performance evaluations are discussed, showcasing the effectiveness and efficiency of the proposed approach in generating high-quality 3D content from conventional 2D videos.

Relevance: The research holds significant relevance in the field of multimedia and entertainment technology, offering a solution to enhance the viewing experience by converting existing 2D video content into immersive 3D formats. The application of machine learning in this context addresses the challenges of depth perception and scene reconstruction. This novel approach not only contributes to the advancement of 3D video conversion systems but also has potential applications in virtual reality, augmented reality, and other areas where creating immersive visual content is paramount. The paper provides valuable insights into the capabilities of machine learning for video processing and lays the groundwork for further developments in the field.

2.7 Paper 7

Title: Deep3D: Fully Automatic 2D-to-3D Video Conversion with Deep Convolutional Neural Networks [7]

Authors:

- Junyuan Xie (University of Washington, Seattle, USA)
- Ross Girshick (University of Washington, Seattle, USA)
- Ali Farhadi (University of Washington, Seattle, USA; Allen Institute for Artificial Intelligence, Seattle, USA)

Topic Discussed: This paper introduces Deep3D, a fully automatic system for converting 2D videos into 3D format using deep convolutional neural networks (CNNs). The authors delve into the development and application of deep learning techniques to achieve seamless and automated 2D-to-3D video conversion.

Description: Deep3D is presented as an innovative solution for transforming conventional 2D videos into immersive 3D content without manual intervention. The core of the system relies on deep CNNs, specifically designed to automatically infer depth information and generate corresponding 3D representations. The paper provides insights into the architecture of Deep3D, detailing the training process where the CNN learns to discern depth cues from 2D frames. The authors showcase the system's performance through experimental results, emphasizing its ability to produce high-quality 3D videos autonomously.

Relevance: The research is highly relevant in the context of multimedia and computer vision, addressing the demand for automated methods in 2D-to-3D video conversion. The use of deep CNNs signifies a departure from manual or rule-based approaches, offering a more sophisticated and adaptive solution. The implications of this work extend to various applications, including entertainment, virtual reality, and augmented reality, where immersive visual content is paramount. The paper contributes to the growing body of research leveraging deep learning for video processing and underscores the potential of deep CNNs in automating complex tasks in the multimedia domain.

2.8 Paper 8

Title: Toward Naturalistic 2D-to-3D Conversion [8]

Authors:

- Weicheng Huang
- Xun Cao, Member, IEEE
- Ke Lu
- Qionghai Dai, Senior Member, IEEE
- Alan Conrad Bovik, Fellow, IEEE

Topic Discussed: This paper explores advancements in naturalistic 2D-to-3D video conversion, aiming to create immersive three-dimensional content from two-dimensional sources. The authors delve into techniques and methodologies that contribute to achieving a more realistic and visually appealing conversion process.

Description: The paper focuses on the development of a system that moves beyond traditional 2D-to-3D conversion methods to create more naturalistic and visually pleasing results. The authors investigate techniques that consider depth perception, scene structure, and visual comfort, enhancing the overall realism of the converted 3D content. The approach incorporates insights from computer vision and image processing to improve the spatial and perceptual qualities of the generated 3D videos. Experimental results and evaluations showcase the effectiveness of the proposed methods, providing a step forward in achieving naturalistic 2D-to-3D conversion.

Relevance: The research is highly relevant in the field of computer vision, multimedia, and immersive technologies, as it addresses the challenge of creating 3D content that closely mimics the natural visual experience. The paper's contribution lies in advancing the state-of-the-art in 2D-to-3D conversion by in-

corporating more sophisticated techniques that consider not only depth perception but also the overall visual comfort for the viewers. The findings have implications for applications in virtual reality, augmented reality, and 3D cinema, where the quality of the converted content is critical for user experience. This work contributes to the ongoing efforts in making 2D-to-3D conversion more natural and visually appealing.

3 Proposed Idea

1. Data Preparation: Assemble a diverse dataset of 2D images with corresponding depth maps.
2. Network Architecture: Design a CNN architecture inspired by the presented papers, emphasizing efficiency and accuracy.
3. Loss Function: Define a loss function to penalize differences between predicted and ground truth depth maps.
4. Data Augmentation: Apply data augmentation techniques to enhance model generalization, including rotations, flips, and changes in lighting.
5. Training: Split the dataset into training, validation, and test sets. Train the model, validate for performance, and avoid over-fitting.
6. Hyperparameter Tuning: Fine-tune hyperparameters systematically, optimizing learning rate, batch size, and regularization terms.
7. Evaluation: Assess model performance using metrics like MAE and RMSE on the test set. Verify visually against ground truth depth maps.
8. Deployment: Deploy the trained model in the target application environment, monitoring real-world performance and considering online fine-tuning if needed.

4 Dataset Description

The GTA5 dataset contains 24966 synthetic images with pixel level semantic annotation. The images have been rendered using the open-world video game Grand Theft Auto 5 and are all from the car perspective in the streets of American-style virtual cities. There are 19 semantic classes which are compatible with the ones of Cityscapes dataset.

References

[1] M. Waheed, A. Jalal, M. Alarfaj, Y. Y. Ghadi, T. A. Shloul, S. Kamal, and D.-S. Kim, "An lstm-based approach for understanding human interactions using hybrid feature descriptors over depth sensors," *IEEE Access*, vol. 9, pp. 167 434–167 446, 2021.

[2] J. Konrad, M. Wang, P. Ishwar, C. Wu, and D. Mukherjee, "Learning-based, automatic 2d-to-3d image and video conversion," *IEEE Transactions on Image Processing*, vol. 22, no. 9, pp. 3485–3496, 2013.

[3] J. L. Herrera, C. R. del Blanco, and N. García, "Automatic depth extraction from 2d images using a cluster-based learning framework," *IEEE Transactions on Image Processing*, vol. 27, no. 7, pp. 3288–3299, 2018.

[4] K. Karsch, C. Liu, and S. B. Kang, "Depth extraction from video using non-parametric sampling," in *Computer Vision – ECCV 2012*, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 775–788.

[5] S. Miao, Z. J. Wang, and R. Liao, "A cnn regression approach for real-time 2d/3d registration," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1352–1363, 2016.

[6] 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, 2016.

[7] J. Xie, R. Girshick, and A. Farhadi, "Deep3d: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 842–857.

[8] W. Huang, X. Cao, K. Lu, Q. Dai, and A. C. Bovik, "Toward naturalistic 2d-to-3d conversion," *IEEE Transactions on Image Processing*, vol. 24, no. 2, pp. 724–733, 2015.