High-Speed Multi-Camera Depth Estimation

Hao Yang, and Chu King Kung

Abstract—The integration of high-speed imaging and depth estimation for real-time applications remains underexplored, despite significant advancements in both fields. This study introduces a novel approach using pixel-wise coded exposure cameras to address challenges in high-speed depth estimation, including computational complexity, hardware constraints, and synchronization issues. By enabling pixel-level exposure configuration and employing cost-effective techniques like temporal multiplexing, the proposed system facilitates the seamless integration of high-speed imaging with conventional depth-estimation algorithms. Results from both static and dynamic scenes demonstrate subframe-level disparity map generation, showcasing the system's capability for agile, real-time depth estimation. The source code implementing the proposed methodology is publicly available at https://github.com/Hao111y/passive-high-speed-stereo.

Index Terms—High-speed Imaging, Depth Estimation, Pixel-wise Coded Exposure, Passive Stereo

1 Introduction

High-speed depth estimation is a cornerstone of modern autonomous navigation and robotics, enabling rapid decision-making and precise maneuvering in dynamic environments. As these fields continue to advance, the need for efficient, accurate, and scalable depth perception systems has become increasingly critical. However, several challenges must be addressed to achieve real-time performance and reliability in high-speed applications.

One significant obstacle is computational complexity. Traditional frame-based stereo methods often struggle to meet the real-time demands of dynamic scenarios, such as obstacle avoidance or high-speed navigation. The intensive processing requirements of these methods can lead to unacceptable delays, undermining their effectiveness in fast-paced environments.

Hardware limitations further complicate high-speed depth estimation. High-speed cameras, while capable of capturing rapid movements, are often bulky, expensive, and resource-intensive. These constraints make them impractical for applications like drones and mobile robotics, where reducing weight, size, and power consumption is paramount.

Synchronization poses another critical challenge. Accurate depth estimation requires precise synchronization across multiple cameras, particularly at elevated frame rates. However, many modern high-speed devices rely on asynchronous readout mechanisms to enhance speed, leading to potential timing discrepancies that degrade depth estimation accuracy and compromise autonomous system functionality. Alternative approaches, such as time-of-flight cameras and active sensors, offer promising solutions but are often limited by range and cost.

To address these challenges, this work proposes the use of pixel-wise coded exposure cameras integrated into a synchronized dual-camera setup. This configuration, paired with off-the-shelve imaging processing and depth-estimation techniques, offers the potential to deliver high-speed, cost-effective depth estimation. Figure 1 highlights the proposed system's design, featuring a fan to create dynamic motion and two synchronized cameras capturing high-speed depth information. This visualization underscores the system's ability to tackle computational, hardware, and synchronization challenges effectively.

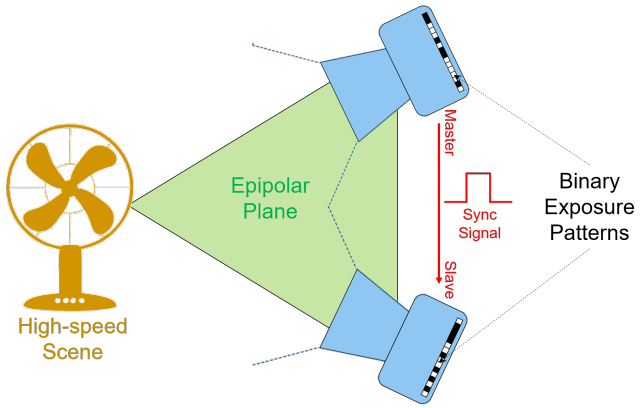

Fig. 1. Illustration of the proposed high-speed depth estimation setup, featuring two synchronized cameras and a rotating fan to simulate dynamic motion.

2 Related Work

Research in stereo vision has historically emphasized accuracy over speed. For example, one method enhances depth measurement accuracy by analyzing error patterns and employing weighted least squares for 3D reconstruction, achieving a 3% improvement [1]. Similarly, another approach integrates semantic segmentation with stereo calibration for depth estimation across multiple cameras [2]. Recent advancements, such as those presented in [3] and [4], apply deep learning techniques to improve multi-camera depth estimation. However, the challenge of real-

- *Hao Yang and Chu King Kung are with the Department of Electrical and Computer Engineering, University of Toronto.*

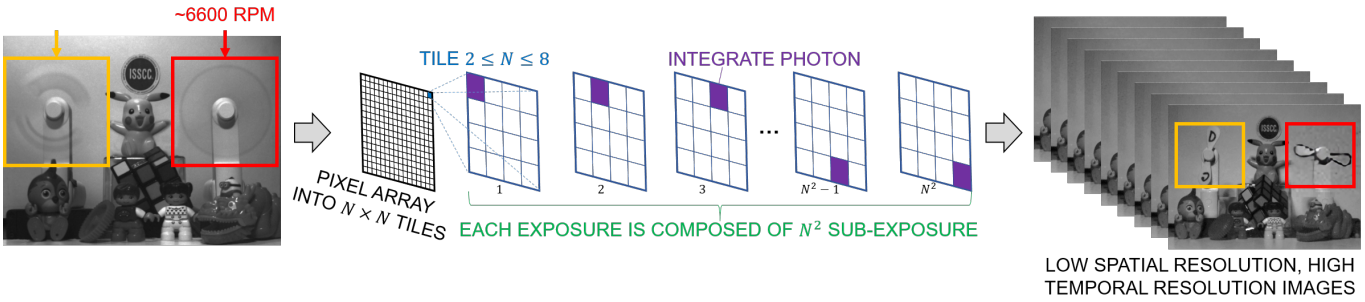

Fig. 2. Temporal Multiplexing: Operational Principles and Experimental Results from High-Speed Imaging Scenarios.

time, high-speed stereo vision integration remains largely unaddressed.

Efforts to achieve high-speed depth estimation have primarily focused on event-driven systems, which demonstrate real-time efficiency and outperform traditional methods [5], [6]. Despite these advantages, such systems encounter significant challenges, including motion-induced asynchrony, timestamp inaccuracies, and variability in event rates, limiting their practical utility.

Active sensors like time-of-flight cameras provide precise depth measurements but are constrained by limited operational range, rendering them less effective in dynamic, high-speed environments with rapid changes.

Conventional frame-based methods remain prevalent due to their established reliability and ease of use. However, they exhibit latency issues and struggle to manage rapid movements, resulting in diminished performance in applications requiring quick responses.

Despite advancements across various approaches, a notable gap persists in integrating high-speed imaging with depth estimation, particularly in cost-sensitive and resource-constrained applications. Addressing this gap represents a critical avenue for future research and innovation.

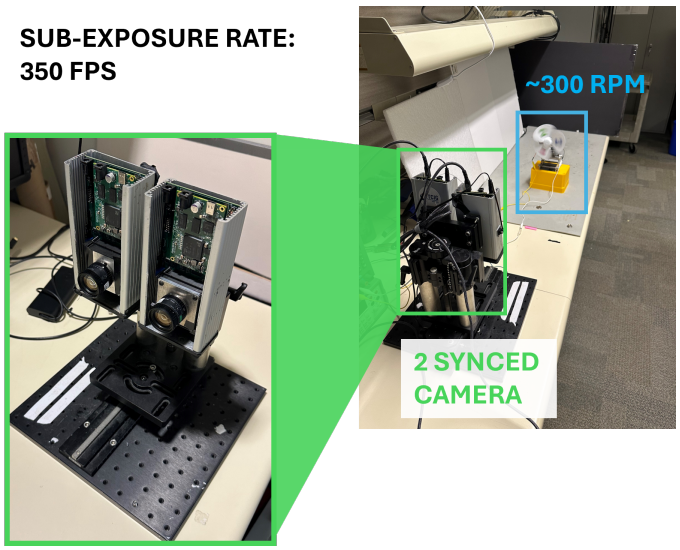

Fig. 3. Hardware setup for passive stereo imaging, designed to capture hybrid scenes featuring both static and dynamic objects.

3 Proposed Method

This proposed method leverages pixel-wise coded exposure cameras and sophisticated processing techniques to significantly enhance high-speed depth estimation capabilities while addressing common limitations faced by other methods. It incorporates several key features that enhance both performance and efficiency.

3.1 Hardware setup

Figure 3 illustrates the hardware and environment used in this project. The two cameras, sourced from the Intelligent Sensory Microsystems Laboratory (ISML) at the University of Toronto, support high-speed coded exposures. We can synchronize the stereo camera pair at rates up to 10 kHz. However, maintaining such a high frame rate necessitates extremely strong illumination and a rapidly moving scene to fully demonstrate our system's high-speed stereo imaging capabilities.

This approach imposes certain trade-offs. Implementing high-speed burst imaging with the CEP sensor reduces spatial resolution and can cause subframe-by-subframe shifts in object position. To achieve a practical balance between temporal resolution and image quality, we selected a 3×3 super-pixel tile size. At this configuration, we attain an exposure rate of approximately 350 Hz, offering a good compromise between speed and clarity.

3.2 System Workflow

The images in this project were captured using a pair of synchronized cameras. To ensure precise alignment and minimize differences between the two cameras (such as fabrication inconsistencies or vibrations during capture), checkerboard calibration was performed at the sub-exposure level. This process produces rectified images, improving overall system accuracy. These rectified images were then subjected to conventional Lanczos upscaling to restore their original resolution, which had been reduced to accommodate high-speed imaging requirements. Non-local means denoising was subsequently applied to improve image quality. Finally, the rectified images were processed using Semi-Global Block Matching (SGBM). We also verify the result with a state-of-the-art MoCha-Stereo [7] matching algorithm recently introduced at CVPR.

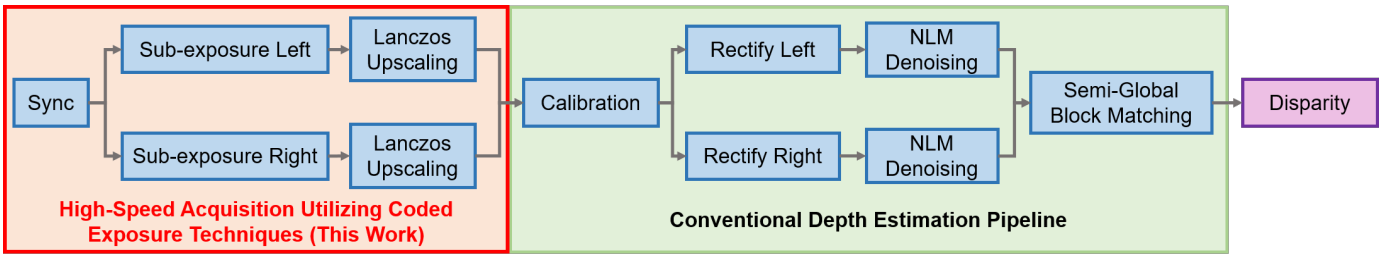

Fig. 4. A block diagram illustrating a high-speed depth estimation system utilizing a conventional image processing pipeline.

3.3 Pixel-Level Exposure Configuration

A fundamental aspect of this method is the pixel-level exposure configuration, which allows for the implementation of various techniques, including temporal multiplexing [8], as shown in Figure 2. This technique enables a strategic trade-off between spatial resolution and speed, facilitating faster image capture without necessitating extensive hardware upgrades. By optimizing exposure settings at the pixel level, the system can effectively manage hardware constraints while maintaining cost-effectiveness.

3.4 SubFrame-Level Disparity Map Generation

The proposed system employs a depth-estimation pipeline designed to process high-speed images, utilizing either traditional methodologies or advanced deep learning techniques. This integration ensures compatibility with existing algorithms, thereby facilitating seamless incorporation into current workflows and making it particularly well-suited for cost-effective, high-speed applications. As illustrated in Figure 4, the proposed dual-camera system can be effectively integrated with a conventional image processing pipeline to achieve high-speed depth estimation.

Among traditional algorithms, Semi-Global Block Matching (SGBM) is widely recognized for its effectiveness in depth estimation tasks. SGBM extends the conventional block-matching approach by optimizing disparity maps through a cost function that balances local and global constraints [9]. An implementation of SGBM is readily available in OpenCV, which slightly diverges from the original algorithm in terms of its cost function, providing a practical solution for many applications.

3.5 Synchronization Techniques

To achieve precise synchronization during high-speed operations, two cameras are configured in a master-slave arrangement. This configuration is facilitated through hardware methods, made possible by our custom-designed cameras. Such a setup enables accurate timing coordination at the subframe level, which is essential for capturing dynamic scenes without motion blur or artifacts. The synchronization techniques implemented ensure cohesive operation among all cameras.

4 Experimental Results

4.1 Static Scene

Figure 5 demonstrates a representative outcome of this study, highlighting the disparity map generated from a static scene. The generated disparity maps with SGBM show excellent correspondence between the left and right grayscale images, with clear delineation of depth levels. The system successfully achieved subframe-level disparity generation, a testament to its efficiency and precision.

4.2 Dynamic Scene

The system's capability to handle rapid movements was evaluated using three dynamic scenarios, as shown in Figure 6. Notably, the fan depicted in the first row of the figure was rotating at a speed of 300 revolutions per minute (rpm). Despite the rapid changes occurring within the scene, no motion blur was detected. The cameras effectively captured these fast-paced dynamics, enabling precise depth estimation of the environment. The second row shows a toy car being pushed forward with some initial speed, while the last row features a candle rolling on the desktop. Both scenes are hybrid, containing static elements (the desk and a box) alongside dynamic ones (a hand, the moving toy car, and the rolling candle). Despite the motion, the outlines of all objects—both stationary and moving—remain well-defined and exhibit no motion blur.

4.3 Verification Results with a State-of-the-Art Stereo Algorithm

To further demonstrate the capabilities of passive stereo imaging in high-speed scenarios, we evaluated our system using the state-of-the-art MoCha-Stereo algorithm [7], recently published at CVPR 2024. To highlight the rapid motion in the scene, we present an image sequence captured within a single frame, subdivided into nine subframes (achieved using a 3×3 tile size).

Figures 7 and 8 show the rectified images from the left and right cameras, respectively. Figure 9 illustrates the disparity maps produced by the SGBM method for comparison, while Figure 10 presents the disparity maps for each subframe generated by MoCha-Stereo.

4.4 Results Discussion

The experimental results highlight our system's effectiveness in handling both static and dynamic scenes. It achieves high-speed, cost-effective, and low-latency depth estimation with a high degree of accuracy. Its robustness under rapid motion makes it well-suited for real-time applications in robotics, autonomous vehicles, and augmented reality.

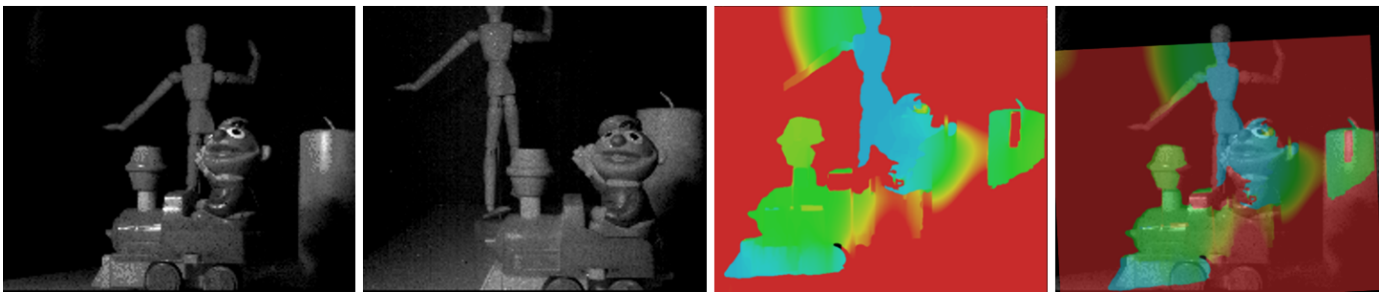

Fig. 5. Disparity map of a static scene: The sequence from left to right includes the left grayscale image, the right grayscale image, the corresponding disparity map, and the disparity map overlaid on the right grayscale image.

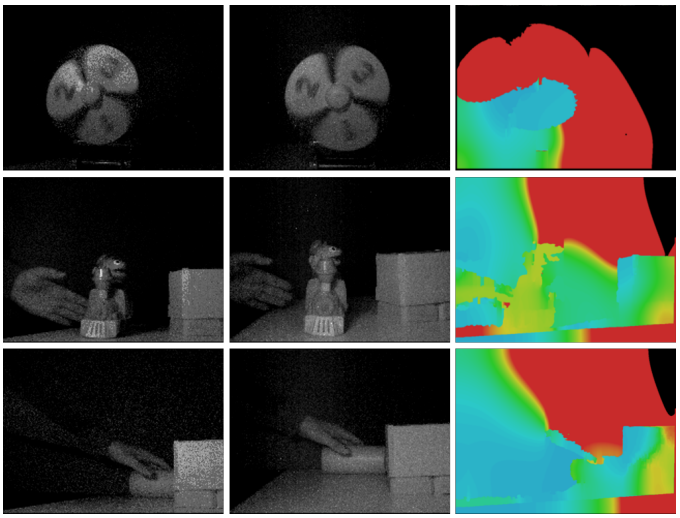

Fig. 6. Dynamic Scene Results: The top row displays a fan rotating at 300 RPM, the middle row features a toy car approaching the camera, and the bottom row depicts a candle rolling toward the camera.

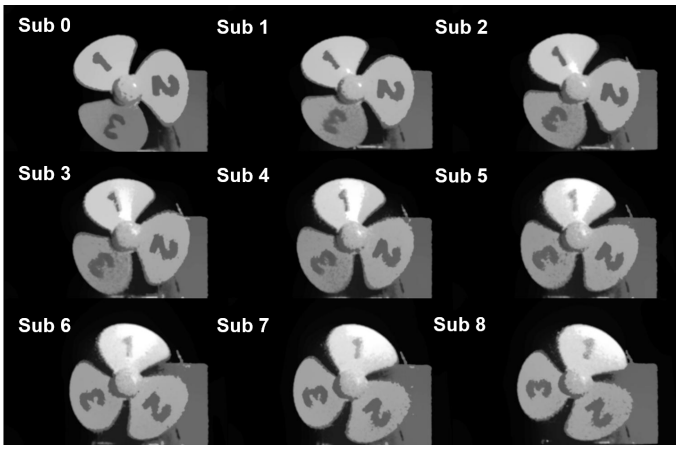

Fig. 7. Image sequence captured within a single frame using a 3×3 super-pixel tile size, resulting in nine subframes. The sequence presented here is from the left camera.

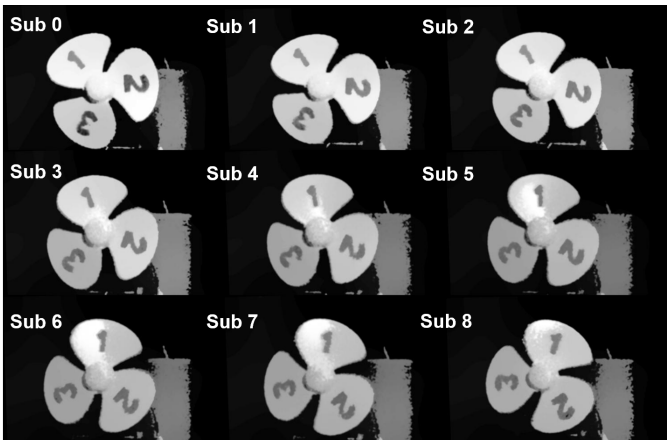

Fig. 8. Image sequence captured within a single frame using a 3×3 super-pixel tile size, resulting in nine subframes. The sequence presented here is from the right camera.

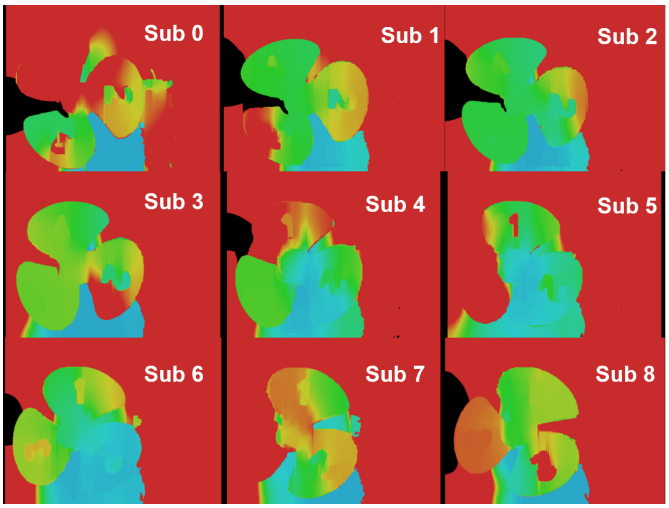

Fig. 9. Calculated disparity map obtained using Semi-Global Block Matching (SGBM). Subframe indices are marked to facilitate matching between left and right views.

Using the SGBM method, most subframes exhibit clearly defined blade edges of the dynamic object (a fan) and maintain recognizable details of static elements. Although some subtle segmentation challenges arise—for instance, the temporary appearance of unexpected black regions representing ambiguous areas—the overall disparity maps remain coher-

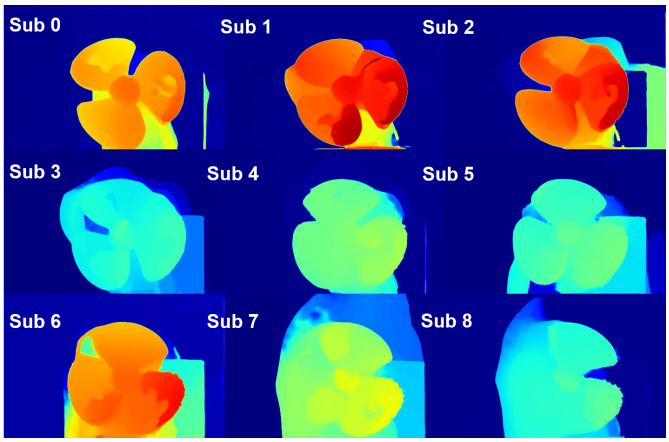

Fig. 10. Calculated disparity map obtained using MoCha-Stereo [7]. Subframe indices are marked to facilitate matching between left and right views.

ent, demonstrating the strong general performance.

By contrast, when integrated with a state-of-the-art method like MoCha-Stereo [7], our high-speed stereo imaging system demonstrates even more refined object delineation across multiple sub-exposures. Certain frames reveal exceptionally clear outlines of both the moving fan blades and the static candle—down to details like the candle wick. While minor inconsistencies still appear, the increased clarity and definition in these key frames highlight the potential of our system to deliver superior results, even when operating under challenging, high-speed stereo imaging conditions.

4.5 Potential Improvements

4.5.1 Adding Realistic Textures for Background Segmentation

To improve object-background segmentation, we introduced checkerboard patterns to add texture to the background.Figures 11, 12, and 13 show, respectively, a sample image captured by the left camera, a sample image captured by the right camera, and the resulting disparity map. However, this approach did not yield significant benefits. Instead, the checkerboard pattern seemed to draw undue attention from the block-matching algorithm, likely due to non-uniform illumination and the presence of shadows. Additionally, imbalanced sensor responses may have exacerbated these issues. Although we conducted thorough gain calibration on both cameras at the outset—striving for a consistent and linear dynamic range—these measures did not fully mitigate the observed challenges.

To address these limitations, we plan to conduct further experiments using textured backgrounds derived from real objects with varying depths, rather than flat checkerboard patterns. This approach aims to introduce natural variations and depth cues that are more representative of real-world scenes, potentially improving the segmentation performance.

4.5.2 Subframe-Level Checkerboard Calibration Method

To successfully implement stereo block searching at the subframe level, checkerboard calibration is applied individually to each subframe. The procedure involves an initial pixel reshuffling step to divide the image into 9 subframes, based on a 3×3 tile size, followed by checkerboard calibration for each subframe. This approach is expected to yield the best results. Figure 10 illustrates the disparity map obtained after applying checkerboard calibration to each subframe.

Despite these results, the observed improvement is marginal when compared to the following previously tested methods:(a) performing checkerboard calibration at the frame level prior to reshuffling pixels into 9 subframes, and (b) reshuffling pixels to generate 9 subframes, followed by applying checkerboard calibration to only the first subframe.

While small improvements are noted, the differences between these approaches are limited, even though subframe-level checkerboard calibration was anticipated to provide superior results.

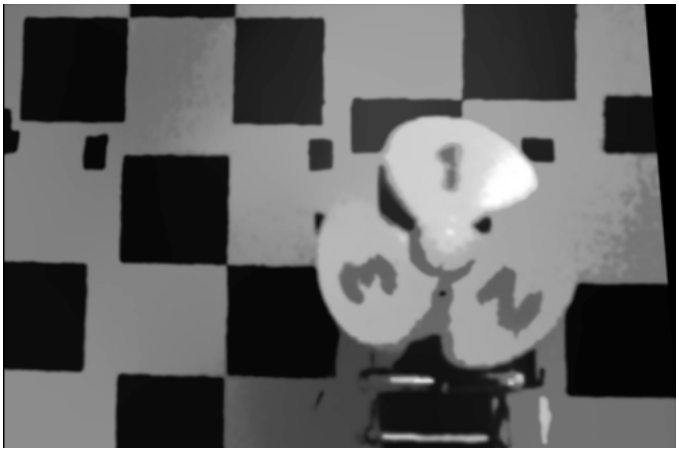

Fig. 11. One image captured within checkerboard background. The image presented here is from the left camera.

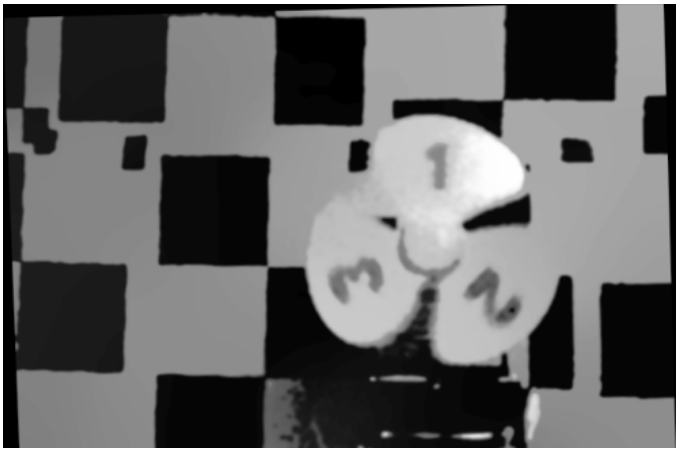

Fig. 12. One image captured within checkerboard background. The image presented here is from the right camera.

Fig. 13. An example of disparity map calculated with checkerboard background using MoCha-Stereo [7].

5 Conclusion

This work presents a novel approach to high-speed depth estimation, leveraging pixel-wise coded exposure cameras and innovative processing techniques to address the critical challenges of computational complexity, hardware limitations, and synchronization. The proposed method integrates seamlessly with existing depth estimation algorithms, enhancing their performance in dynamic environments.

In comparison to existing methods, the proposed approach bridges the gap between cost-effectiveness and performance. Unlike event-driven and active sensor-based methods, which face synchronization or cost-related constraints, the pixel-wise coded exposure methodology offers a balanced trade-off between spatial resolution and speed. Additionally, the system's compatibility with off-the-shelf components broadens its applicability across various domains.

In conclusion, the proposed system provides a promising solution for high-speed depth estimation, paving the way for advanced real-time applications in dynamic settings.

Acknowledgments

The authors wish to express their sincere gratitude to Professor David Lindell and Professor Kyros Kutulakos for their invaluable insights and guidance throughout this research.

ChatGPT Statement

Some paragraphs of this report are polished with ChatGPT.

References

[1] S. Bi, Y. Gu, Z. Zhang, H. Liu, C. Zhai, and M. Gong, "Multi-camera stereo vision based on weights," in *2020 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, 2020, pp. 1–6.

[2] F. Yang, Z. Liming, Z. Yi, and K. Hengyang, "Multi-camera system depth estimation," in *2022 IEEE 6th Information Technology and Mechatronics Engineering Conference (ITOEC)*, vol. 6, 2022, pp. 1202–1207.

[3] Y. Shi, H. Cai, A. Ansari, and F. Porikli, "Ega-depth: Efficient guided attention for self-supervised multi-camera depth estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 119–129.

[4] Y. Wei, L. Zhao, W. Zheng, Z. Zhu, Y. Rao, G. Huang, J. Lu, and J. Zhou, "Surrounddepth: Entangling surrounding views for self-supervised multi-camera depth estimation," in *Conference on robot learning*. PMLR, 2023, pp. 539–549.

[5] J. N. Martel, J. Müller, J. Conradt, and Y. Sandamirskaya, "An active approach to solving the stereo matching problem using event-based sensors," in *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2018, pp. 1–5.

[6] H. Kim, S. Lee, J. Kim, and H. J. Kim, "Real-time hetero-stereo matching for event and frame camera with aligned events using maximum shift distance," *IEEE Robotics and Automation Letters*, vol. 8, no. 1, pp. 416–423, 2022.

[7] Z. Chen, W. Long, H. Yao, Y. Zhang, B. Wang, Y. Qin, and J. Wu, "Mocha-stereo: Motif channel attention network for stereo matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 27 768–27 777.

[8] R. Rangel, X. Sun, A. Barman, R. Gulve, S. Bajic, J. Wang, H. Wang, D. B. Lindel, K. N. Kutulakos, and R. Genov, "23,000-exposures/s 360fps-readout software-defined image sensor with motion-adaptive spatially varying imaging speed," in *2024 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)*. IEEE, 2024, pp. 1–2.

[9] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 30, no. 2, pp. 328–341, 2007.