Human-Inspired Camera: A Novel Camera System for Computer Vision

Shubham Kumar*, Jonathan Mi*, Qingyuan Zhang, Benjamin Chang, Hao Le, Ramsin Khoshabeh, Truong Nguyen
Electrical and Computer Engineering Dept., University of California, San Diego

Abstract— **Computer vision models aim to emulate biological design so that systems can perform meaningful tasks. We believe that the underlying processes of the human visual system hold the keys to further improving the performance of such computer vision solutions. This exploratory paper investigates the swaying motion of human vision when walking to develop a novel camera system. We successfully demonstrate that this design is able to improve performance in computer vision tasks, such as monocular depth estimation.**

Keywords: computer vision; depth estimation; human vision

I. INTRODUCTION

Stereoscopic vision, 3D scene understanding and representation, and depth estimation have been subjects of academic interest long before computer vision became a discipline [1][2]. In this work, we aim to demonstrate a camera system inspired by the human visual system. We observe that as humans perceive the world, it is not simply that they perceive depth and understand the scene because they have two eyes. When a person sees the world, it is often while in motion or having accumulated a mental model of the world while in motion. By studying the movement patterns of humans as they walk, we have developed a more robust camera system that is capable of emulating that behavior. We demonstrate that by using such a camera, we are able to more closely model what a person sees. Therefore, we hope that our approach will yield much better algorithmic performance than traditional cameras for vision-based tasks. The applications of this work are far-reaching – from surveillance to 3D capture and reconstruction and even to the vision systems of self-driving cars.

II. HUMAN-INSPIRED CAMERA SYSTEM (HICS)

A. Human Motion Modeling: In order to develop a human-inspired camera system (HICS), we first identify the motion displacement pattern of the eyes as a person is walking. To record this motion while walking on a treadmill, three colored trackers are placed on the subject's face. Video footage is collected with the camera positioned directly in front of the subject. We subsequently extract motion data from people of various heights and gaits and observe that the motion is roughly a semi-circular parabolic movement (see Fig. 1).

B. Design of HICS: Having collected the motion data, we designed and built two HICS prototypes to mimic the approximate motion. The prototypes contain an Intel RealSense D435 camera [3] for its raw video footage as well as depth data for ground truth analysis and model verification.

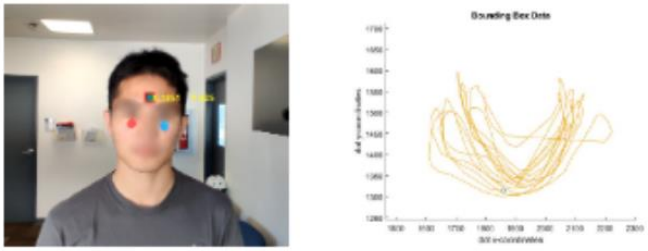


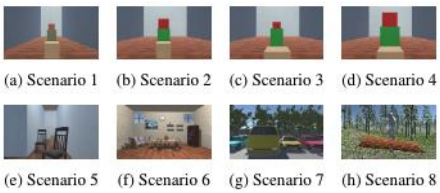
Figure 1. Human motion data collection

The first HICS uses a virtual four bar mechanism to move the camera in a semi-circular pendulum-like motion and is designed to match the human motion as closely as possible. The second unit utilizes a rack and pinion mechanism to move the camera horizontally in a linear motion. This second design is developed to evaluate the significance of the vertical motion in human perception. Using these setups (see Fig. 2), we collected footage from two different scenes with multiple variations of colored boxes and chairs.

C. Synthetic Dataset: Prior to evaluating the HICS designs, we simulate data from a virtual environment and compare algorithmic performance with ground truth data. We generate a synthetic dataset with Unity [4], a 3D game engine, and 3D assets available on the Unity Asset Store. The eight different scenarios are shown in Fig. 3, where the virtual cameras follow both linear and semi-circular movements with parameters similar to the HICS (e.g., filed-of-view and focal length). Additionally, high precision depth values are used as ground truth and are stored as depth images. Both real and synthetic datasets will be made publicly available.

(a) HICS 1 (b) HICS 2

Figure 2. Oscillatory (right) and linear (left) HICS



(a) Scenario 1 (b) Scenario 2 (c) Scenario 3 (d) Scenario 4

(e) Scenario 5 (f) Scenario 6 (g) Scenario 7 (h) Scenario 8

Figure 3. Scenarios created from Unity

* Denotes equal contribution.

III. RESULTS

In this work, we use an out-of-the-box spatio-temporal monocular depth network (ST-CLSTM) proposed by Zhang *et al* [5]. We measure the performance of the ST-CLSTM model on HICS video frames and compare it to the performance on static image frames. The simulation is based exclusively on the synthetic Unity dataset. Evaluation of real data is left to future work.

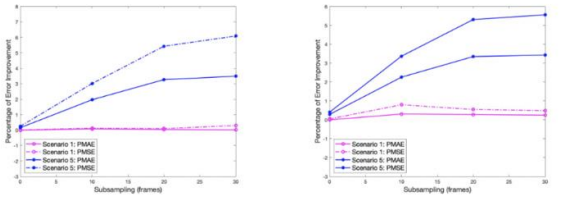
Our camera system is configured to collect data at 60 fps. In order to simulate larger baselines, we evaluate the HICS with 0 (none), 10, 20, and 30 frame sub-sampling. The following simulation tests investigate the effect of the HICS-like motion in various situations. We extract the mean squared error (MSE) and mean absolute error (MAE) of the prediction compared to the ground truth depth values and report the percent error difference (PMAE \& PMSE) between the static camera case and the four HICS motion cases. In the following subsections, we evaluate the results on: 1) basic scenarios, 2) depth masking, and 3) a comparison of linear and oscillation motion.

A. Full Image Evaluation: We begin with full image evaluation of the basic scenarios (Scenarios 1-5 in Fig. 3), meaning that we report PMAE and PMSE based on the whole frame. No significant improvements are observed for Scenarios 1-4, but Scenario 5 yields a notable decrease in error.

Fig. 4a reveals that performance is only improved with the HICS when there is severe pixel occlusion in the scene. Scenario 5 features a chair blocking a box, and the box only comes into view during the semi-circular motion. This movement in the horizontal and vertical plane gives the temporal ST-CLSTM more information about occluded pixels, facilitating a decrease in error. In Scenarios 1-4, the motion does not reveal much background information, limiting the effect of the HICS on the ST-CLSTM estimate.

Furthermore, we observe that error tends to decrease with larger sub-sampling (which simulates a larger baseline), meaning that the HICS is able to capture more background information from occluded scene objects. A larger baseline allows for more occluded pixels to become visible through the swinging motion, improving ST-CLSTM's performance.

B. Depth Masking Evaluation: Masking isolates certain areas of the image to help us understand where the HICS motion has the greatest impact on the estimator. In depth masking, we included pixels that were at or near the depth of the farthest foreground object when calculating error. This effectively allows us to evaluate the performance of the model on mostly foreground objects. In general, we notice increases in PMAE and PMSE in all scenarios with lower sub-sampling rates. The masking results (Fig. 4b) suggest that most of the improvement from the HICS's motion is in depth prediction on foreground objects.



(a) Full Image Scenario (b) Depth Masking Scenario

Figure 4: PMAE & PMSE of Scenario 1 & 5 for two experiments

C. Linear vs. Semi-Circular Motion: The final point of monocular evaluation is the comparison between the semi-circular and linear motion. While the parabolic motion benefits from displacements in both the vertical and horizontal planes, linear motion only has movement in the horizontal plane.

In Scenario 5 and 7, we observe that when parabolic motion improves depth estimation, linear motion does so to a lesser extent. We also note that linear motion has less of an increase of error when parabolic motion increases error, as seen in Scenario 8. This can be important for designing the next iteration as we observe more prominent improvement with parabolic motion but more stability with linear motion

IV. CONCLUSION

This paper explores a new Human-Inspired Camera System (HICS) and provides evidence that the system is able to increase performance in monocular depth estimation. Most current work tries to improve these methods by focusing on new models and algorithms. However, our approach differs in that we move the camera to capture more information from the scene. There is great potential for such a camera system, and there are many more open questions to investigate. One objection to our system is that a stereo camera system can accomplish the same thing, but we argue that the motion of the monocular HICS system brings value to computer vision algorithms, along with cheaper costs. Future avenues to explore include stereo estimation from a single, monocular HICS system and 3D scene reconstruction.

REFERENCES

[1] Charles Wheatstone, "Xviii. contributions to the physiology of vision.—part the first. on some remarkable, and hitherto unobserved, phenomena of binocular vision," Philosophical Trans. of the Royal Society of London, no.128, pp. 371–394, 1838.

[2] Ramsin Khoshabeh, Jason Juang, Mark Talamini, Truong Nguyen, "Multiview glasses-free 3-d laparoscopy, "IEEE Trans. on Biomedical Engineering, vol. 59, no. 10, pp. 2859–2865, 2012.

[3] Intel realsense d435 depth camera. [Online]. Available: https://www.intelrealsense.com/depth-camera-d435/.

[4] "Unity 3d." [Online]. Available: http://unity3d.com.

[5] Haokui Zhang, Chunhua Shen, Ying Li, Yuanzhouhan Cao, Yu Liu, Youliang Yan, "Exploiting temporal consistency for real-time video depth estimation," in International Conference on Computer Vision (ICCV), Seoul Korea, Oct. 2019, pp. 1725-1734.