Snapbot : Enabling Dynamic Human Robot Interactions for Real-Time Computational Photography

Chanyeok Choi
angledsugar@hanyang.ac.kr
Hanyang University
Ansan, South Korea

Jeonghan Kim
kimjh9813@hanyang.ac.kr
Hanyang University
Ansan, South Korea

Yunjae Nam
ujma1234@hanyang.ac.kr
Hanyang University
Ansan, South Korea

Youngmoon Lee
youngmoonlee@hanyang.ac.kr
Hanyang University
Ansan, South Korea

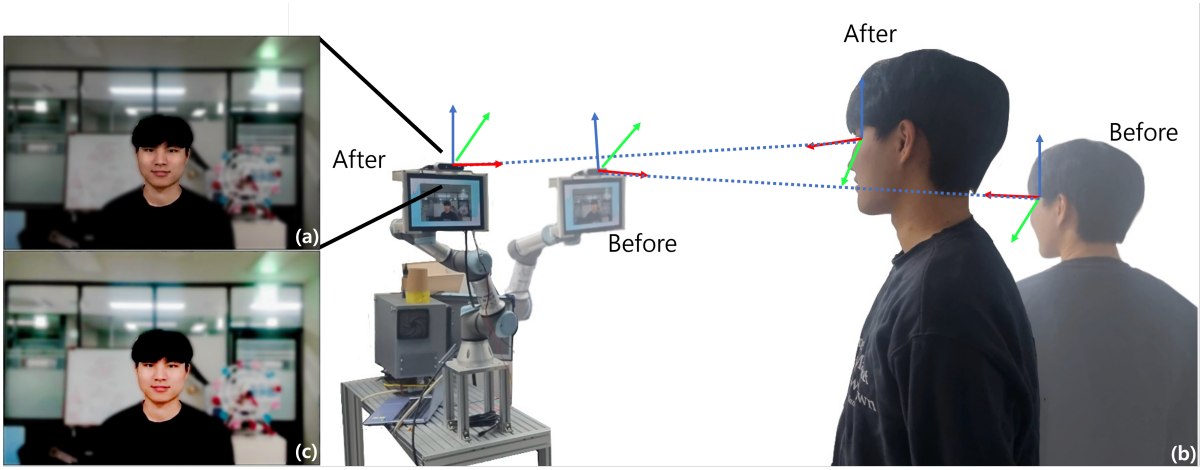

Figure 1: Snapbot leverages state-of-the-art computational photography in human robot interaction system to (a) dynamically capture human subjects (§2.1), (b) interactively adjust camera composition (§2.2), and (c) generate enhanced images (§2.3).

ABSTRACT

Photography remains an expert area requiring right focus, exposure, composition, and even post-processing. Yet, robotic automation can enable precise camera manipulation, focus and exposure adjustment, camera composition, and post-processing by leveraging state-of-the-art computational photography. Existing proposals for robotic photography focus on adjusting camera angles for static portraits or developing image evaluation metrics, thus falling short in capturing dynamic human robot interactions. This paper describes the design and implementation of Snapbot, a human robot interaction system designed specifically for computational photography. Snapbot dynamically detects face and pose for exposure and focus and interactively controls robot arm for camera composition to perform image scoring and enhancing. As perception, control, and computational photography form an end-to-end pipeline, Snapbot promises a new future in which image focus, exposure, composition, and generation can be jointly optimized as a unified process. We have implemented and deployed Snapbot on a UR3 demonstrating the mean image quality score is 1.51× compared to aesthetic visual analysis dataset. We also perform ablation study to analyze the impact of each stage of Snapbot both visually and quantitatively.

HRI '24 Companion, March 11–14, 2024, Boulder, CO, USA

ACM ISBN 979-8-4007-0323-2/24/03…$15.00
https://doi.org/10.1145/3610978.3640712

CCS CONCEPTS

• **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability.

KEYWORDS

Human-Robot interaction, Robotics, Computational Photography

ACM Reference Format:
Chanyeok Choi, Jeonghan Kim, Yunjae Nam, and Youngmoon Lee. 2024. Snapbot : Enabling Dynamic Human Robot Interactions for Real-Time Computational Photography. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI '24 Companion), March 11–14, 2024, Boulder, CO, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3610978.3640712

1 INTRODUCTION

Computational photography is widely used for image-based relighting, image enhancement, image deblurring, geometry/material recovery and so forth. The main reason for the popularity is simple: while photography remains an expert area requiring right focus, exposure, composition, and even post-processing, anyone can generate better images. However, human photographers cannot

leverage computational photography at best when capturing dynamic scenes, manipulating camera, and aforementioned settings in real-time.

There are two primary factors in leveraging computational photography: (i) humans are not efficient in real-time computational tasks and (ii) machines are not efficient in visual representation and appreciation. Real-time computational photography integrated with robotic automation can enable precise camera manipulation, focus/exposure adjustment, image composition, and even post-processing at once on the fly. Yet, human can intuitively adjust subjects and interactively supervise machines for better composition, where human robot interaction (HRI) comes into the picture. Our goal is to marry human robot interaction with state-of-the-art computational photography to dynamically capture human subjects and interactively compose and enhance images in real-time using perception, control, and generative models.

In this paper, we present Snapbot, a new human robot interaction solution that enables effective robotic photography by advancing computational photography. It detects human, performs autofocus/exposure, estimates head pose to control robot arm for camera composition, and then scoring and enhancing images in perception, control, and computational photography pipeline. Snapbot, unlike existing computational photography, brings a new opportunity to jointly optimize focus/exposure, camera composition as well as image generation all together as a unified process.

Proposals for robotic photography [8–10, 16], acknowledge this opportunity and aim to operate robots to capture images. Recent studies on computer vision [2, 3] and graphics [5, 17] suggest that real-time pose estimation and image scoring can meet the image quality requirements for photography datasets [11, 12] in accordance with the human experts. However, existing proposals focus on static portraits [9, 10] requiring 20-30s [8] or even 60s [16] lead time, making them infeasible for real-time HRI. To address this, Unlike existing solutions, Snapbot needs to capture real-time human movement and interaction together that causes the camera sensory depth noises, then dynamically and safely operate robot arm. The main challenge is to i) track dynamic subjects in real-time, ii) control robot arm capturing interactions, and iii) selectively enhance real-time generated images for computational photography.

Snapbot addresses these challenges via three primary stages: perception, control, and computational photography. Perception stage performs depth estimation, auto exposure, and DNN-baesd autofocus to capture real-time dynamics of subjets and noises. Control stage manipulates camera composition and closed-loop inverse kinematics to capture human robot interactions. Computational photography stage runs image score model and enhanced image generative model to select and enhance images from burst shooting. We have implemented Snapbot using UR3 and RealSense D435 camera demonstrating that the mean image quality score is 0.5189 vs. 0.7852 when comparing Snapbot with the aesthetic visual analysis (AVA) dataset, improving score by 1.51×.

2 SNAPBOT DESIGN

Snapbot is a human robot interaction solution for snap that leverages state-of-the-art computational photography. Figure 2 illustrates a novel approach to robotic photography, designed to integrate the dynamic movements of manipulators and DNN-based

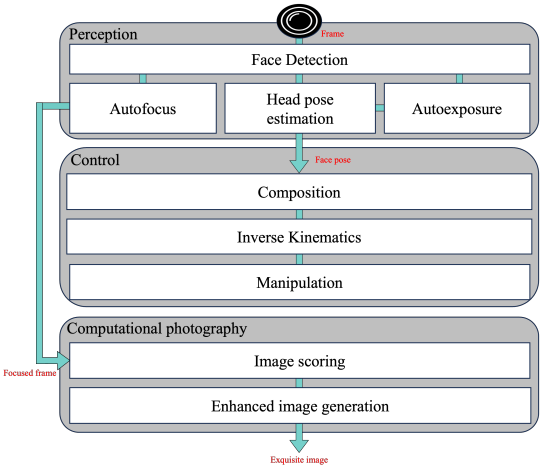

Figure 2: Overview of Snapbot. In the perception section, recognise a human face in the camera frame. Autoexpose and autofocus on a human face and human pose estimation. Control (§2.2) section, the system computes the camera composition by leveraging the direction of the subject and controls manipulators. The computational photography is that the photo to Snapbot taken scoring and enhancing image generation.

computational photography. For the latter, we introduce three primary stages: perception, control and computational photography

2.1 Perception

In the context of dynamic interactions in photography, three perception techniques have been implemented for real-time photography quality: Autoexposure for precise depth extraction, Autofocusing for photographic quality, and Head pose estimation for dynamic composition.

Autoexposure. Snapbot uses lower resolutions to alleviate real-time processing overhead, albeit causing deterioration of overall quality in depth during dynamic interactions. In the Snapbot's HRI scenarios, The quality of depth is exclusively pertinent in the detected facial region rather than the entire frame. Autoexposure which maintain the average the intensity of all the pixels inside of detected facial region prevents the deterioration in restricted region despite the overall lower resolution. By automatic adjustment throughout dynamic interactions, Snapbot achieves highly precise position estimation of the face, even in real-time.

Autofocus. In cases of Autoexposure in particular region, it often leads to a notable degradation in the performance of RGB frames. In perception stage, Snapbot leverages image focus model [15] for recovering/focusing the degraded image because professional-grade photography should be required high-quality image. Using the bounding box obtained through face detection as the focus target drives background defocusing and foreground focusing. This allows the process to ensure the professional-grade focusing capabilities with recovering quality.

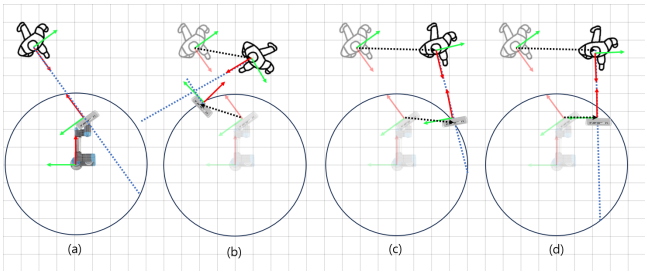

Figure 3: Camera Composition, depicting the movement of the camera in response to the orientation and position of the human. (a) : camera position relative to the initial human pose, (b), (c), (d) : variations in Camera Composition in response to human movement.

Head Pose Estimation. For determining the composition of Snapbot, real-time and precise head pose estimation should be required. we apply an algorithm-based head pose estimation approach [1]. This involves a real-time landmark localization process [7] generating 2D-facial landmarks with facial image. Using a quadtree-search method with 2223 pre-constructed samples of head poses, we estimates the orientation with the most similar sample as head pose. This estimation technique enable us to achieve both accurate and real-time head pose estimation.

2.2 Control

Camera composition in conjunction with a robotic manipulator can significantly improve the quality of captured images. We propose a novel system that automatically computes an optimal camera composition leveraging the perceived human's pose, enabling both dynamic subject tracking and real-time interactive control.

Composition. The system dynamically tracks the human subject by leveraging the perceived facial direction vector. This vector is combined with the transformation matrix details aligning the manipulator and facial coordinate systems, along with pose data, to encode the desired manipulation into the transformation matrix H, as shown in Eq. (1). Additionally, a constraint ($y^* = 0$) is imposed to exclude undesired translational movements of the human face along the y-axis.

$$H = \begin{bmatrix} R & P \\ 0 & 1 \end{bmatrix}, X = \begin{bmatrix} x & y & z \end{bmatrix}^T, X^* = \begin{bmatrix} x^* & y^* & z^* \end{bmatrix}^{-1}$$
$$\begin{bmatrix} X^* \\ 1 \end{bmatrix} = H \begin{bmatrix} X \\ 1 \end{bmatrix} \quad (1)$$

In Algorithm 1, The parameters O, C and d represent the robot's base frame, the camera's frame, and range of manipulator motion. The illustration depicted in Figure 3-(b), (c), and (d) correspond to the output lines 9, 15, and 17 within Algorithm 1. The output ensures(perform, allow) position of the camera composition throughout dynamic subject. Furthermore, the orientation of composition is determined by the orientation vector within the subject's coordinate system. This methodology enables dynamic tracking of the subject while optimizing the search for an optimal camera composition.

Algorithm 1 Camera-composition

1: **function** Minimize(H, Y)
2: $X^* = [x^*, 0, z^*]^T, X = H^{-1}X^*$
3: $k = \min_{x^*,z^*} \|H^{-1}X^* - Y\|$
4: **return** X, k
5: **end function**
6: **Input**: H^{-1}, O, C, d, **Output**: P
7: $A, \alpha \leftarrow$ Minimize(H, O), $B, \beta \leftarrow$ Minimize(H, C)
8: **if** $\alpha > d$ **then**
9: $P \leftarrow A$
10: **else**
11: **if** $\|B - O\| > d$ **then**
12: $\|H^{-1}X^* - C\| = d$
13: $D \leftarrow H^{-1}X^*$
14: min $\|D - B\|$
15: $P \leftarrow D$
16: **else**
17: $P \leftarrow B$
18: **end if**
19: **end if**

Inverse Kinematics. The target position of the manipulator undergoes dynamic changes in each moment, and its orientation is required to consistently align with the human during this process. To fulfill these criteria while minimizing computational overhead, we introduced a closed-loop inverse kinematics system employing the damped pseudo-inverse method. Detailed insights into the employed methodologies can be found in [4] and [13]

$$J^* = J^T(JJ^T + \lambda^2 I)^{-1} = \sum_{i=1}^{n} \frac{\sigma_i}{\sigma_i^2 + \lambda_i^2} v_i u_i^T \quad (2)$$

The manipulator encounters singularity when its determinant approaches zero. To avoid this challenge, the Pseudo Inverse is applied as a avoidance method for the zero determinant problem. Nevertheless, the use of the Pseudo Inverse may introduce instability concerns in close proximity to the singularity. Motivated by the potential for instability near singular points, the Damped Pseudo Inverse is implemented to improve stability. This technique utilizes a damping term, λ, incorporated into the standard Pseudo Inverse, as defined in Eq. (2)

Manipulation. Given the dynamic nature of human movement being tracked by the system, the target point of the end effector experience substantial fluctuations, leading to peaks in robot control. These oscillations pose a challenge to the precise motion of the robot and contribute to a shortened operational lifespan. To address these issue, we introduce a control input filtering mechanism employing a moving average filter [14]. Through the integration of an inverse kinematics system and the application of a moving average filter, we successfully implement a sophisticated camera composition methodology that accommodates the dynamic movements of a human subject while maintaining control stability. This enables real-time interactive control, allowing adjustments to the composition based on changes in the pose of the subject.

2.3 Computational Photography

In §2.1 and §2.2, we discuss the robot's method for making sure our system takes image of human. However, we run Snapbot system into a problem because we are receiving images in low quality to achieve real-time in this process. We need to provide human with images that are high quality and look like they were edited by a professional photographer. To provide human with the best possible images, we introduce the Computational Photography method.

Image Score Model (ISM). Like a professional photographer, we need to filter out the photos taken by the robotic system that are out of focus and those that don't capture the human properly. Trained on the AVA portrait dataset, ISM evaluates photos based on aesthetic criteria like composition and lighting, effectively filtering out lower-quality images. Through this method, the robotic system prioritizes high-scoring images, ensuring that only the most aesthetically pleasing images, free from common issues like blurriness, are selected for final presentation.

Enhanced Image Generative Model (EIGM). We need to convert one image selected in ISM to high quality, and make it look like it was retouched by a professional photographer. EIGM primary function is to enhance key digital imaging processes, including exposure compensation, hue and saturation adjustment, color space conversion, tone mapping, and gamma correction. These procedures, traditionally requiring significant manual input and expertise from photographers, are streamlined by EIGM to ensure consistent, high-quality outcomes across various scenes and conditions.

Snapbot has developed a novel computational photography system that leverages a robotic arm for capturing and subsequently enhancing images of dynamically moving subjects. This system integrates the ISM and the EIGM to effectively improve image quality. It is adept at successfully capturing low-resolution images of subjects in real-time motion and significantly elevating the final output's quality through expert-level image correction techniques.

3 EXPERIMENTS

Configuration and Setup. Snapbot uses UR3 manipulator and CN0364 touchscreen LCD for HRI interfacing. Human subject was captured using a RealSense D435 camera, and object estimation was performed using the yolov8-face model [6]. The live demonstration and experiment comprehensively recorded the entire Snapbot scenario, executed on an Intel i5-6600K CPU with a Titan RTX 24GB GPU. AVA dataset [11] used to train the ISM by extracting the human category from over 250,000 images containing metadata such as aesthetic scores and labels for over 60 categories.

Result. Balancing elements, depth of field, lighting, and object presence are crucial factors in Snapbot's image scoring system, significantly impacting the quality of images captured during dynamic real-time human tracking. Without the aid of ISM and EIGM, the images presented in Figures 4-(a),(b),(c) remain unfiltered and unenhanced, resulting in Snapbot scores consistently falling below all categories of the AVA dataset. Therefore, it's essential to exclude images whether the human is out of focus or not. By using only the ISM, the image shown in upper Figure 4-(d) results in a score of 0.5521, higher than the AVA dataset score of 0.5189. This score demonstrates that Snapbot's perception and control are effective in capturing human. Finally, utilizing the entire system including the EIGM, the image shown in lower Figure 4-(d) achieves a score of 0.7852, representing a 1.51× improvement compared to the AVA dataset. Snapbot demonstrates the ability to capture humans in dynamic interactions, resulting in images that rival those taken and enhanced by professional photographers.

	Score	Balancing Elements	DoF	Light	Object
AVA Dataset	0.5189	0.0341	0.1152	0.0568	0.0941
Snapbot	0.7852	0.8085	0.7318	0.7799	0.6360
w/o EIGM	0.5521	0.7002	0.5910	0.5832	0.4217
w/o ISM, EIGM	0.3544	-0.062	0.1144	-0.1943	-0.1030

Table 1: Overall mean score and individual evaluation metrics for Aesthetic Visual Analysis dataset vs. Snapbot with and without ISM, EIGM.

Figure 4: Comparative Analysis of System Applications. Column (a) illustrates images before and after the application of Autofocus (Sec. 2.1). Column (b) contrasts photographs taken by a fixed camera without composition with those captured by the robotic arm post-composition application (Sec. 2.2). Column (c) displays the effect of applying the ISM, showcasing images before and after ISM application (Sec. 2.3). Column (d) compares images before and after the EIGM.

4 CONCLUSION

The Snapbot is a novel human robot interaction system for computational photography achieving professional-grade photography in dynamic settings. Snapbot accurately tracks and determines the positioning of a subject's face, dynamically adjusting camera focus and composition. This capability is pivotal for capturing diverse angles and adapting to varying lighting conditions. Snapbot achieves professional-grade photographic configurations, elevating the resultant images to a professional standard through its automated control systems and refined image processing algorithms. Our in-depth evaluation has demonstrates its advantages in substantially improving photography both quantitatively and visually.

ACKNOWLEDGMENTS

This work was supported in part by the National Research Foundation of Korea (NRF) grant 2022R1G1A1003531, 2022R1A4A3018824 and Institute of Information and Communications Technology Planning and Evaluation (IITP) grant IITP-2024-2020-0-01741 funded by the Korea government (MSIT).

REFERENCES

[1] Andrea F Abate, Paola Barra, Carmen Bisogni, Michele Nappi, and Stefano Ricciardi. 2019. Near real-time three axis head pose estimation without training. *IEEE Access* 7 (2019), 64256–64265.

[2] Niaz Ahmad, Jawad Khan, Jeremy Yuhyun Kim, and Youngmoon Lee. 2022. Joint human pose estimation and instance segmentation with poseplusseg. In *AAAI*.

[3] Niaz Ahmad, Jawad Khan, Jeremy Yuhyun Kim, and Youngmoon Lee. 2022. MultiPoseSeg: Feedback Knowledge Transfer for Multi-Person Pose Estimation and Instance Segmentation. In *ICPR*.

[4] Samuel R Buss. 2004. Introduction to inverse kinematics with jacobian transpose, pseudoinverse and damped least squares methods. *IEEE Journal of Robotics and Automation* 17, 1-19 (2004), 16.

[5] Heng Huang, Xin Jin, Xinning Li, Shuai Cui, and Chaoen Xiao. 2022. Aesthetic evaluation of Asian and Caucasian photos with overall and attribute scores. *Computers and Electrical Engineering* 103 (2022), 108341.

[6] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. 2023. *YOLO by Ultralytics*. https://github.com/ultralytics/ultralytics

[7] Vahid Kazemi and Josephine Sullivan. 2014. One millisecond face alignment with an ensemble of regression trees. In *CVPR*.

[8] Kai Lan and Kosuke Sekiyama. 2019. Autonomous robot photographer with KL divergence optimization of image composition and human facial direction. *Robotics and Autonomous Systems* 111 (2019), 132–144.

[9] Pei-Chun Lu and Kai-Tai Song. 2021. Interactive Motion Planning for Autonomous Robotic Photo Taking. In *International Conference on Control, Automation and Systems (ICCAS)*.

[10] Ren C Luo, Wai Un Chan, and Po-Jen Lai. 2014. Intelligent robot photographer: Help people taking pictures using their own camera. In *International Symposium on System Integration (SII)*.

[11] Naila Murray, Luca Marchesotti, and Florent Perronnin. 2012. AVA: A large-scale database for aesthetic visual analysis. In *CVPR*.

[12] Miriam Redi, Nikhil Rasiwasia, Gaurav Aggarwal, and Alejandro Jaimes. 2015. The beauty of capturing faces: Rating the quality of digital portraits. In *International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*.

[13] Bruno Siciliano. 1999. The Tricept robot: Inverse kinematics, manipulability analysis and closed-loop direct kinematics algorithm. *Robotica* 17, 4 (1999), 437–445.

[14] MT Tham. 1998. Dealing with measurement noise. Moving average filter. *Chemical Engineering and Advanced Materials, University of Newcastle upon Tyne* (1998).

[15] Chengyu Wang, Qian Huang, Ming Cheng, Zhan Ma, and David J Brady. 2021. Deep learning for camera autofocus. *IEEE Transactions on Computational Imaging* 7 (2021), 258–271.

[16] Taisei Yokomatsu and Kosuke Sekiyama. 2022. Optimal Viewpoint Selection by Indoor Drone Using PSO and Gaussian Process With Photographic Composition Based on KL Divergence. *IEEE Access* 10 (2022), 69972–69980.

[17] Hui Zeng, Jianrui Cai, Lida Li, Zisheng Cao, and Lei Zhang. 2020. Learning image-adaptive 3d lookup tables for high performance photo enhancement in real-time. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 4 (2020), 2058–2073.