ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 10, No. 5, 11-19 (2025)
www.astesj.com

ASTES Journal
ISSN: 2415-6698

3D Facial Feature Tracking with Multimodal Depth Fusion

Jenna Snead[1,2], Nisa Soltani[3], Mia Wang[2], Joe Carson[1,4], Bailey Williamson[4], Kevin Gainey[4], Stanley McAfee[4], Qian Zhang[*,3]

[1]*Department of Physics and Astronomy, College of Charleston, Charleston, South Carolina, 29424, USA*

[2]*Department of Computer Science, College of Charleston, Charleston, South Carolina, 29424, USA*

[3]*Department of Engineering, College of Charleston, Charleston, South Carolina, 29424, USA*

[4]*Pensievision Inc., San Diego, California, 92130, USA*

ARTICLE INFO

Article history:
Received: 30 June, 2025
Revised: 28 August, 2025
Accepted: 30 August, 2025
Online: 15 September, 2025

Keywords:
Convolutional neural networks (CNNs)
Shape-from-focus
Facial feature tracking
Human Robot Collaboration (HRC)

ABSTRACT

As models based in artificial intelligence increase in sophistication, there is a higher demand for the integration of hardware components to heighten real-world implementations. Both facial feature tracking and shape-from-focus are known techniques in computer vision. However, the combination of these two elements, particularly in a cost-effective configuration, has not been extensively explored. In this study, a 64 megapixel (MP) autofocus Arducam camera module collected images of participants at various focal lengths and the Laplacian of the Gaussian (LoG) identified the frame of maximum sharpness each second. The image data was then processed by two convolutional neural networks (CNNs) from Google MediaPipe that identified the bounding box for the face and the coordinates of facial features. These coordinates, in conjunction with a shape-from-focus calculations, were fused to measure facial feature depths relative to the camera system. The depths, aggregated across a working period contributed another metric for total participant effort in a Human Robot Collaboration (HRC) experiment without introducing significant additional costs or logistical modifications. Inheriting the constraints of an existing HRC configuration, this methodology achieved consistent 2D tracking of facial features and coarse 3D facial motion trends from a singular, static imaging system.

1. Introduction

The application of artificial intelligence (AI) to human interactions has greatly increased in sophistication. In the case of the human face, the ability to automatically locate individual facial features enables higher granularity analysis into emotion, movement, and posture. Increases in precision for facial feature analysis has been shown to reduce subjectivity in emotion-based research and aid the modeling of neural activity [1, 2]. Active appearance models (AAMs), which take in image key-points as training and predict their locations in novel images, are a viable algorithmic option for facial feature tracking [3]. By training a model on labelled facial key points, an AAM can automatically record the movement of a face over time. In conjunction with the use of Haar Cascades, a scale-invariant object detection algorithm, studies have achieved real-time facial feature tracking on devices as compact as mobile phones [4]. While these models have been largely successful in theoretical and idealized implementation, AAMs still experience a decline in performance when applied to real-world, unpredictable situations [5].

The tracking of facial features gains a degree of freedom in complexity when transitioning from two dimensions (2D) to three dimensions (3D). However, this additional dimension unlocks capabilities related to posture, motion, and enhanced positional information, making it a valuable level of abstraction. One way to accomplish 3D tracking is to fit the subject to a pre-existing facial model, such as with *Candide* [6]. This is beneficial for applications where the relative tip and tilt of the face are required without needing absolute distance information.

It can also be valuable to track the distance from the camera to the human subject. One way of accomplishing this is through shape-from-focus. This method, which is described in detail in Section 2.1, relies on an initial calibration to map each camera focal length with a unique distance to the in-focus plane of the target. Once aligned, this relationship can be used to predict the distance to the target across a range of focal lengths during the experiment. Live movement can cause image blurring, making it challenging to determine whether the lack of focus is due to the target being out of focus or an in-focus target in motion. When used to track

[*]Corresponding Author: Qian Zhang, 66 George St. Charleston, SC 29424, 716-598-9621 & zhangq@cofc.edu

the natural movement of humans, this degeneracy can often lead to significant uncertainties, as demonstrated in the literature on overall face detection [7].

While the prioritization of decreasing computational cost is a widespread concept, the idea of similarly decreasing hardware burden in the field of data science is relatively under-discussed. The use of low-cost imaging equipment such as those produced by Arducam (Arducam, China, Nanjing), however, is shown to assimilate well to advanced research environments [8]. Along with academic applications, low-cost technologies have also been explored in the manufacturing process, particularly for the use of the identification of defective parts [9, 10]. This inclusion of low-cost methods is crucial for increasing the accessibility of the technologies to a wider range of applications.

The promise of low-cost technologies, combined with the expanding capabilities of AI, motivated this study to incorporate the advantages of both hardware and software growth to a realistic application. In response to the consistent gaps in model performance when introduced to non-experimental environments, this study sought to maximize the capability of a 3D facial feature tracking system while inheriting the constraints of the existing HRC data collection set-up. While the components of the system may not be individually novel, the methodology in this paper presents a practical way to augment the capabilities of a low-cost hardware configuration and estimate its corresponding uncertainties in the absence of ground-truth data.

The rest of the paper will be organized as follows:

1. A review of the current state-of-the-art, and its corresponding gaps, in facial feature tracking and HRC in Related Work.
2. A summary of the hardware and software components of the experimental set-up in Methodology.
3. A description of the connection of the experimental set-up to a larger HRC study in HRC Application.
4. An overview of the results of the experiment in Validation.
5. A comparison of the results to the state-of-the-art in Discussion.
6. A succinct recap of major points in Conclusion.

2. Related Work

Live facial feature tracking becomes particularly relevant in applications of Human-Robot Collaboration (HRC). The sounds generated and realism of general robotic appearance can shape or even decisively alter the human comprehension of its movements [11]. The ability of users to recognize human personality traits in robot collaborators can influence their perception and trust in performance [12]. These postures can be improved by more in-depth tracking of human participant motion patterns [13]. Through the collection of accurate, automated evaluations of the movement and emotions of the human collaborator, much information about the state of human during the interaction can also be gleaned. Convolutional neural networks (CNNs) have been used for emotional analysis of human participants through treating the problem as a 2D classification [14]. Along with distance considerations, facial feature tracking has also been shown to drastically improve facial recognition model performance by allowing the model to use previously identified faces as context for subsequent predictions [15]. When combined with other HRC tasks such as hand motion detection, facial feature recognition can verify that the user has the proper permissions associated with their role [16].

This feat has also been achieved in 3D, allowing for a nuanced analysis of human facial postures. In this case, 3D imaging was accomplished using a stereoscopic, multi-camera system, where the multiple images are mapped together to acquire depth information [17]. Another multi-sensor study utilized rotation of the camera to create an unprecedented augmentation of 3D data, and compiled such images using point clouds. Using multiple Asus Xtion sensors, a depth uncertainty of 16-23 millimeters at a distance of 0.8 meters was recorded [18]. While powerful in the precision achieved, these previous studies utilize either multiple cameras, static objects, premium cameras, Time-of-flight (TOF) sensors, or a depth-sensing laser. With predictable decreases in resolution, this methodology seeks to expand on the applications of these previous studies by evaluating the suitability of 3D facial feature tracking in the presence of additional practical constraints and limiting image collection to a single low-cost camera module.

Shape-from-focus, facial feature tracking, and the use of Arducam variable focus camera modules are all individually well-known techniques across the field of computer vision. However, this study represents a novel investigation through the combination of these relevant elements for a realistic HRC application. By generating absolute distance to the face through shape-from-focus, the distances traveled in the XY plane are evaluated by an independent measure from the distances traveled by the Z axis, representing a fusion of measurement techniques from a singular instrument. This inclusion of absolute 3D information, when combined with low-cost technology and HRC applications, represents a practical and repeatable component integration. With these in mind, the goals of this project were to: (1) track the position of facial features, (2) track the depth of the face as a whole, and (3) evaluate the granularity to which the depth of the individual facial features can be tracked. These goals and their corresponding set-up were configured such that overall participant motion can be measured for HRC experimentation without modifying any of the study's preexisting constraints.

3. Methodology

The set-up for this experimentation includes the coordination of inexpensive, portable hardware components with open-source processing software.

3.1. Hardware components

As seen in Figure 1, the Arducam Hawkeye camera module represents a compact, low-cost variable focus measuring tool. While there exist other imaging systems that offer higher precision, the Arducam module's open-source, inexpensive features make it favorable to portable and inclusive applications. This was the sole

camera module used to capture data in the HRC experiment, and its specifications can be found in Table 1. The Hawkeye camera uses a motor to change the focal length, which correspondingly changes the distance to the in-focus object plane.

Table 1: The hardware specifications for the Arducam "Hawkeye" 64MP Autofocus camera. This device is driven by a Raspberry Pi computer using the Bullseye OS[19].

Part Name	Resolution	Focus
Arducam 64MP Autofocus Camera	9152x6944	8cm-INF

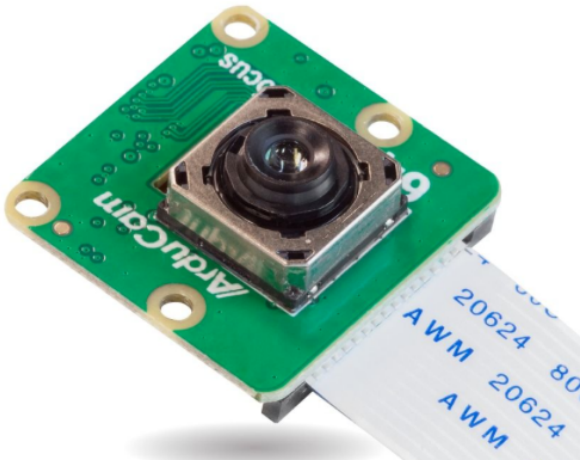

Figure 1: The 64 MP Arducam Hawkeye camera module uses a motor to change the focal length, and thus the depth in the image that comes into focus. Image Source: https://www.arducam.com

3.2. Shape from Focus

As a single camera was employed to minimize both cost and spatial burden, depth estimation had to be conducted from a single viewpoint. One method to accomplish this, as mentioned in Section II, is shape-from-focus. This is mathematically justified using the thin lens equation 1:

$$\frac{1}{f} = \frac{1}{o} + \frac{1}{i} \tag{1}$$

where f is the focal length, o is the object distance, and i is the image distance, the relationship between the focal length of a lens and the distance to a target can be established [20]. Using this concept, iterated across small steps of distance, focal length is converted to the depth of each piece of an image. When all of the pieces are aggregated together, the overall shape of the object can be mapped, known as "Shape from Focus" [21]. To successfully implement shape-from-focus in practice, there are two especially important steps. The first is the creation of an accurate 'sharpness map' for each digital image, for purposes of evaluating which focal length brings a given target feature into sharpest focus. The second is the application of iterative filtering of the sharpness map to reduce overall noise and also to identify-and-reject outlier sharpness values. An outlier value can result, for example, from target movement, excessive glare, or other factors.

In this investigation, the sharpness map creation relies on second spatial derivatives, a version of a Laplacian transformation. The optimized filtering and outlier rejection employs parameterized Gaussian functions. The overall strategy is therefore referred to as a Laplacian of the Gaussian (LoG), and has overlapping properties with versions of LoG filtering described in standard image processing literature. The specific shape-from-focus strategy used was contributed by the Pensievision team, and represents a version of the strategy documented in U.S. Patent No. 20190090753 [22]. Later figures demonstrate example curves from the shape-from-focus strategy, where sharpness-of-focus is represented on the y-axis as the standard deviation of the aforementioned Laplacian version.

A distance calibration is required to describe the measured surface in true depth units. The Hawkeye camera's motor physically modifies the camera focal length; therefore, for each motor position, or "step count", a unique distance is brought into camera focus. To determine the distance corresponding to each step count, a flat target at a known distance was imaged across the range of motor positions. The sharpness evaluation revealed which step count achieved the sharpest image for the given object distance. By repeating this process for a range of object distances, a physical distance was determined for each motor step count. In Figure 2, step values are plotted against their corresponding distance values to create a regression for conversion of step count to physical distance.

In the case of finding the depth for individual facial features, the step value of the most in-focus image was converted to a physical distance, as calibrated through the above process. This distance was then combined with a synthetic relative facial feature depth to generate a fusion-based absolute distance estimate across the entire face.

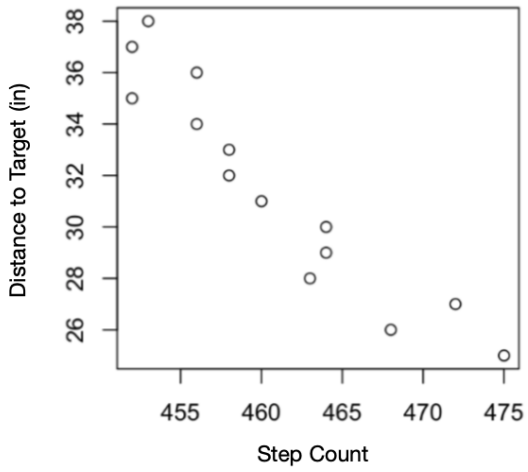

Figure 2: Distance in inches was plotted against steps from the camera module during the calibration process to generate an approximate conversion between the two measurements.

3.3. Software System

Given that facial feature detection is a well-studied task, two complementary state-of-the-art models were employed: Google MediaPipe's FaceDetector and FaceMesh-V2 algorithms. The FaceDetector, based on the BlazeFace detection model, was first used on

every image gathered to find the bounding box of the face in the image for cropping [23]. These cropped images were used for the estimation of overall face depth instead of the general images, as they allowed for a more consistent field-of-view with the removal of the background. The FaceDetector is based on a single shot detector (SSD) structure, resulting in a low computational cost of the facial cropping step [24].

Once the image with the most in-focus face was discerned, the FaceMeshV2 algorithm, which is based on the Attention Mesh face mapping model, was used to generate (x,y,z) coordinates of 478 facial landmarks per image [25]. The participants were facing within 90 degrees of the camera field-of-view at all times, centered, and positioned relatively in the plane of the camera system, making the full image data fall within the requirements of successful implementation of the FaceMesh-V2 algorithm.

The following is a description of the general procedure, repeating once per second of image collection:

1. Detect the bounding box for the face in each image using Google MediaPipe's FaceDetector model and crop
2. Calculate the standard deviation of the LoG measure for each cropped image
3. Return the step value of the image with the largest LoG and convert to centimeters (cm) to find depth to face
4. Input full image corresponding to the facial region of highest focus into Google MediaPipe's FaceMesh-V2 model, which returns x, y, and z positions of 478 facial features. The x and y positions are recorded in pixels, whereas the relative z "depth" is recorded as a distance of the feature from the face's center of mass, as normalized with respect to the face width. Like the absolute distance value, the z depth is measured perpendicular to the plane of the face.
5. Convert the relative facial feature z position from normalized measure to centimeters based on the mean facial width, as stratified by biological sex
6. Record facial features' x and y coordinates in pixels, relative feature depth z in centimeters, total distance to face in centimeters, as well as the net change of these measurements from the second before.

Overall, the input to the software pipeline was a series of variable focus images taken over the course of a second, and the output was 478 individual facial feature positions and depths at each unit time. In total, this represents 1435 total measurements per second of the experiment. The specific source code used to process these measurements can be found in the linked repository[1].

4. HRC Application

The variable focus camera module and facial feature tracking software was implemented as part of a HRC experiment. In this procedure, a participant organized blocks with assistance from two robot arms. A total of 30 participants completed the HRC tasks. An image of the experimental set-up is found in Figure 3.

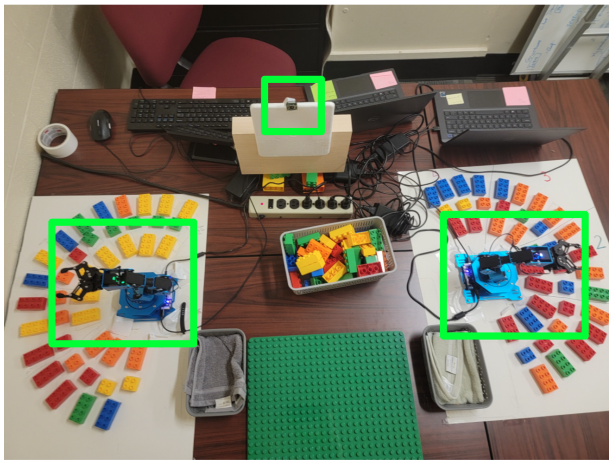

Figure 3: Participants worked with a variable number of robotic arms at different speeds to stack blocks, with their stress levels both visually and electrodermally monitored. The robotic arms and camera module are indicated by the light green boxes.

There were three different independent variables being iteratively changed: number of robots (1 or 2), robot speed, and robot orientation. For each of these variations, the participant was imaged by the Arducam camera module and had their electrodermal activity (EDA) and wrist acceleration measured by an Empatica E4 wristband (Empatica Inc., Boston, MA). These two separate measurements, along with self-reported assessment, formed the basis for the evaluation of the participant's stress levels while working with the robots on each task.

The 3D facial feature tracking using the variable focus camera module is relevant to assessing participant stress levels. This technology enables the calculation of the net movement of the participant's facial features during tasks and facilitates the approximation of changes in posture. The fusion of relative depth estimations from synthetic fitting by the FaceMesh-V2 CNN with the physical measurement from depth-from-focus calculations permits an additional dimension of data collection of the participant, in which all axes of motion can be tracked. This enhanced facial and posture analysis provides a valuable additional metric to diversify stress monitoring analysis in human robot collaboration without changing any of the conditions for the participant. By eliminating any further requirements on behalf of the participant, the application of this methodology to unconventional or more varied setups is enabled. In conjunction with biological and self-reported data, facial analysis can provide significantly more insight into participant stress levels than previous stress monitoring systems in this field. The inclusion of depth information in facial analysis is crucial due to the widespread use of 3D pose estimation in studies involving worker fatigue [26, 27].

As a subset of a larger HRC study, this methodology intends to seamlessly integrate into the participant sessions and evaluate the granularity to which facial motion can be measured. The aggregation of such metrics with EDA and self-reported data falls outside

[1] https://zenodo.org/records/15713616

the scope of this approach and is left for analysis by the greater HRC experiment.

5. Validation

The HRC experiment provided a valuable opportunity for the entire algorithm and hardware set-up to be tested in a way that is not replicated in testing accuracy of a static, idealized data set.

5.1. Evaluation of Software Performance

Based on the seated position of the participant and the relatively continuous nature of the task (i.e. not actively reacting to immediate stimulus), successful position tracking would produce relatively smooth curves that gradually change with respect to subject motion. While jumps across select seconds of movement are anticipated, the overall motion is approximated as stable.

Figure 4, which depicts the standardized x and y positions of the participant's tip of nose, demonstrates a smooth trajectory across both dimensions of movement, particularly for $t > 15$ seconds.

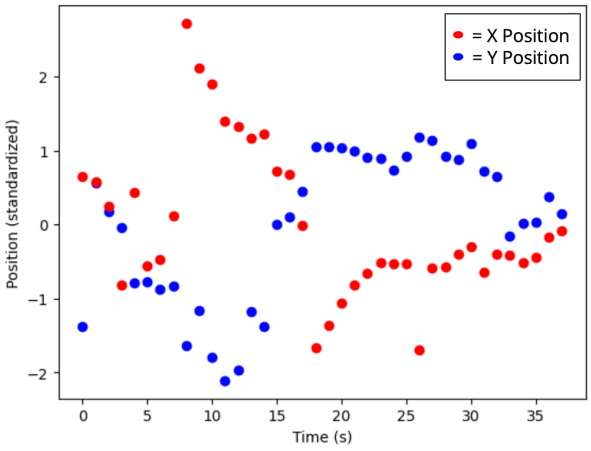

Figure 4: The X (red) and Y (blue) positions of the tip of the participant's nose over the course of a 38 second baseline session.

Notably, this trend is not replicated in the Z position measurement, in which the standardized depths are scattered without clear slope between -1.5 and 1.25 in Figure 5. When aligned with the timescale of the previous figure, it is clear that although there is some continuity between 25 and 35 seconds, the variability in the depth measurements is largely inconsistent with smooth, continuous motion. This implies that the depth calculated from the fusion technique is not precise enough to meaningfully capture motion on the time-scale of seconds.

Without benchmarks for the known 3D positioning of the participant throughout the experiment, the xy predictions of facial feature predictions were treated as a proxy for ground truth, as the mean absolute error normalized by inter-ocular distance (IOD MAE) of the model ranges from 2.67-3.85%. Given that human annotators generally average an IOD MAE of 2.62%, the model's classification represents a valid, higher fidelity benchmark to anchor results in the absense of true ground truth[2]. This was also verified by a qualitative examination of the model's predictions on each image from the sessions depicted in Figures 4 and 5, fn which the model successfully identified the pixel location of each major facial landmark. In previous works introducing the FaceMesh model, textural plausibility and qualitative confirmation of 3D renderings have served as the validation of the technique [28]. The correlation with 2D motion combined with the manual examination of the model's performance represents this methodology's surrogate for ground truth given the experimental constraints.

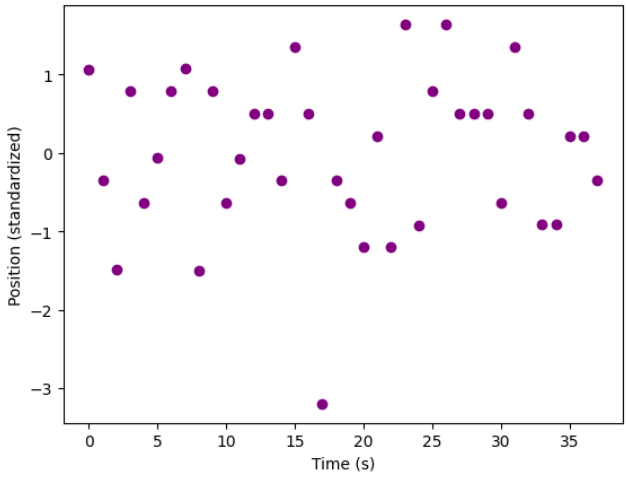

Figure 5: The Z positions of the tip of the participant's nose over the course of the same 38 second baseline session as Figure 4. The demonstrated lack of a smooth trajectory suggests that the depth resolution is larger than the actual changes in depth over second time scales.

Without a ground truth depth to each facial feature or distance to the participant in general, the aggregation of changes in Z was compared to changes in the XY plane, which had been visually verified. Once again taking the seated position of the participant into account, the body position is largely constrained to leaning (no shuffling, squatting, or other uni-dimensional motions). Thus, the motion in one dimension is assumed to be inextricably linked with motion in the other two dimensions in terms of timing. This allows for the utilization of the accuracy of motion in the XY plane to flag moments of high motion. A successful depth tracker, in this set up, would tend to report large changes in depth on the same timescales as large changes in XY position, with the exception of occasional erratic motion.

To carry out this evaluation, the net change in Z position by second (Figure 6) and by session (Figure 7) was plotted against the net changes in XY position on the corresponding time scale. The Pearson correlation coefficient and associated p-value were calculated using the relationship between the covariance matrices of the two dimensions, with significance based on a t-distribution at $n - 2$ degrees of freedom [29, 30].

The correlation between these trends helped inform whether there was a significant pattern of mutual change across the 3 dimensions of measurement.

[2]Google FaceMeshV2 Model Card: https://storage.googleapis.com/mediapipe-assets/ModelCardMediaPipeFaceMeshV2.pdf

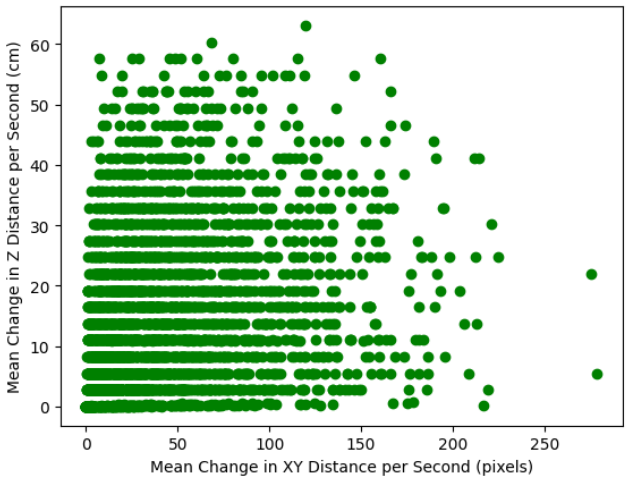

Figure 6: The changes in Z distance in centimeters, as aligned temporally with corresponding changes in the XY positions, demonstrate no clear pattern. The presence of data points representing changes in depth of over 40 centimeters in a single second also likely indicate a complete breakdown of depth as calculated by shape-from-focus. Each point corresponds to a single second of data collection.

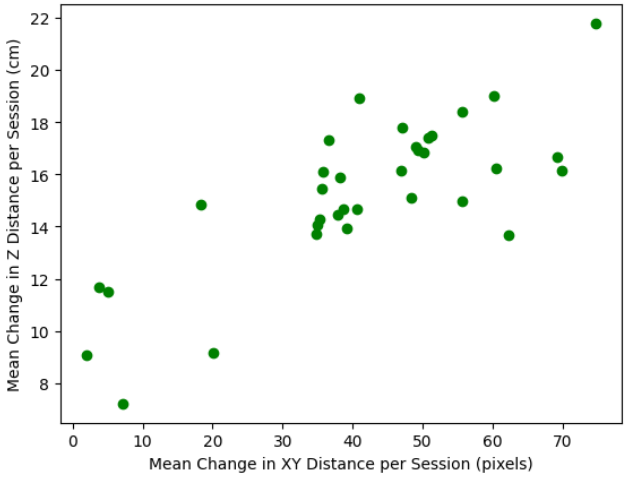

Figure 7: The changes in Z distance in centimeters, as aligned by aggregate session with corresponding changes in the XY positions, demonstrate a meaningful pattern. Each point here corresponds to a session consisting of between 30-120 seconds.

The net change in depth by the second, when compared to changes in XY position reiterates the lack of per-second depth precision from Figure 5. Additionally, the presence of measured changes in depth reaching values greater than 60 centimeters in a single second highlights instances of complete breakdown of the shape-from-focus measurements, likely caused by significant motion blur in the image. The Pearson correlation coefficient calculated between these two variables was $r(908) = 0.1993$, $p < 0.00001$, which demonstrates that the depth perception is not precise enough to accurately track motion by the second, even in terms of the simple identification of motion. However, when this metric is aggregated across a session, which consists of a period between approximately 30 and 120 seconds (depending on the HRC experiment task), the correlation is much stronger with $r(32) = 0.7775$, $p < 0.00001$.

This pattern indicates that the depth measurements from the shape-from-focus and synthetic texture fusion may be precise enough to highlight sessions of overall high aggregate motion. While the ability to track motion in the Z dimension is a low benchmark of precision, it serves as an important indicator that such motion is being effectively monitored, albeit with significantly more restrictive limitations on resolution.

Each data point in Figure 7 represents a session of the HRC experiment as collected over the course of 4 different participants. In total, the 36 points are an aggregation of 3409 seconds worth of data consisting of 1435 measurements per second. Thus, while 36 may seem like a statistically small sample size, the trend is actually being driven by the aggregated patterns of almost 5 million measurements.

5.2. *Evaluation of Hardware Performance*

The high performance of 2D facial feature tracking, without the ability to match precision in 3D demonstrated the limits at which shape-from-focus can provide useful absolute depth measurements, both in terms of time scale and general distance of the target from the camera. As seen in Figure 2, the absolute depth of a static, 2D object over the range of 445-475 steps reached uncertainties of approximately 2 inches. The resolution notable worsens in the case of a moving 3D human target, as is depicted in the differences between Figures 8 and 9.

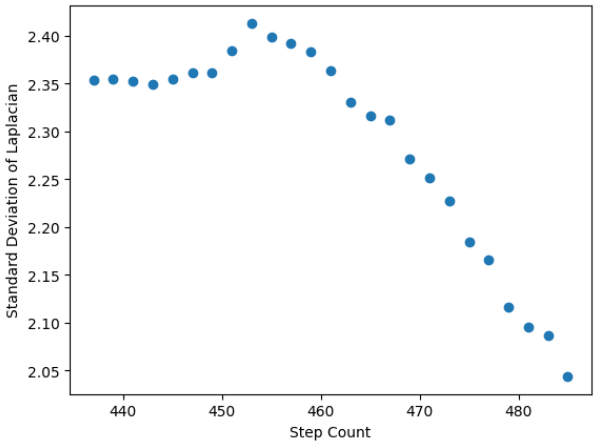

Figure 8: Time periods where the participant was relatively stationary yielded clear curves and led to more consistent depth measurements. This curve represents one second of data collection during the baseline section of the experimentation and demonstrates when shape-from-focus can more definitively identify the sharpest image.

Many human facial features lie within 2 inches of each other, completely eliminated the capabilities for individual facial feature depth resolution at the experimental distance from the target. However, overall facial movement can often exceed 2 inches, validating the use of this method for reliable 3D measurement at lower granularity.

Additionally, the camera required a 0.04 second gap between images to allow for the movement of the motor. Without this small window, the camera module would produce intermittent errors due to insufficient time to properly configure. The need for time lags over the course of an iteration introduced a trade-off between the number of steps taken between each image, and the overall amount

of time per focus range. Since the subject was engaged in a task, there was nearly constant motion. Consequently, the time window for iterating step values had to be sufficiently small to approximate the person as being static during that period. While a second is slightly lengthy for this assumption, it represented the most optimal compromise between capturing numerous images with fine depth changes and minimizing the amount of motion between images within a single time window.

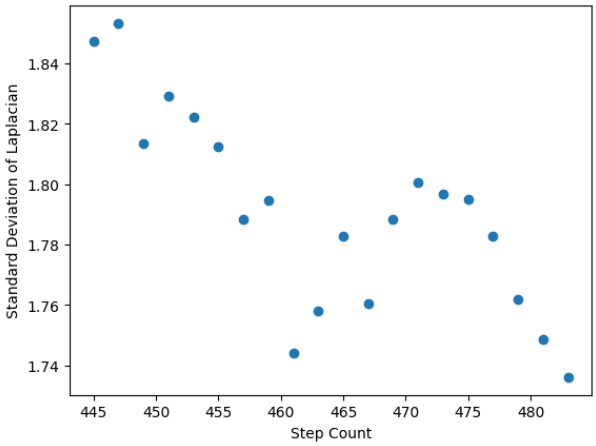

Figure 9: When compared to the graph in Figure 8, it is evident that significant movements over the course of a second cause drastic graph degradation.

6. Discussion and Future Work

As established in Section 3, the goals of this project were to: (1) track the position of facial features, (2) track the depth of the face as a whole, and (3) evaluate the granularity to which the depth of the individual facial features can be tracked, in order to measure overall participant motion for HRC experimentation.

In terms of (1), the models were highly successful tracking facial features in 2D. This supported both by the generation of smooth curves across second-based timescales, as well as the visual verification of results on representative sessions.

For goal (2), a distance-to-target calibration (Figure 2) was achieved to help quantify the hardware-based limitation of depth resolution. While this precision was insufficient for absolute distance calculations or trends of Z-axis motion by the second, it was able to consistently report periods of high motion when aggregated by the session of observation.

Lastly, in terms of (3), the uncertainties related to depth resolution were greater than the differences of depths between the average person's individual facial features. Without the necessary resolution in 3D, the tracking of the depth of individual features was highly imprecise due to imaging hardware constraints. In the cases of (2) and (3), it is apparent that the use of absolute depth from the camera without the inclusion of any assumptions of general face shape made the camera module's precision the critical ceiling for the precision of all depth measurements. Despite this inability to resolve individual facial features, the result of (3) was nonetheless a valuable assessment of the 3D capabilities of a single, low-cost camera system under stringent HRC constraints.

As described in the Introduction and depicted in Figure 9, facial feature model performance tends to sharply decline when introduced to the natural movement of the participants. This study experienced the same trend, particularly when combining facial feature location with depth measurements. In future studies using this methodology, efforts should be taken to address motion blur. Possible mitigation steps include decreasing exposure time in the camera system's driver (or correspondingly, increasing shutter speed) or utilizing deep learning networks to remove blur in post-processing [31].

Based on these results, the net movement of the participant as a whole was reliably measured in 2D on both second-based and session-based time scales. These measurements were also replicated across all 478 individual facial key-points each second to provide data on changes in face orientation over time. While the absolute depth was unable to be discerned at the necessary level of granularity, the net movement in the Z dimension was also recorded for flagging overall sessions of large posture changes as well as creating a standard for future improvement.

Without ground truth depth data as collected by a separate laser or TOF sensor, it is impossible to quantify the frame-by-frame precision of this study's methodology. The inability to resolve the depths of individual facial features on a per-second basis is definitively less precise than the 16-23 millimeter uncertainty recorded in a multi-sensor study, as many parts of the face exceed that range in distance [18].

Despite limitations in 3D precision, the results of this experiment addressed the aforementioned gaps in the state of the art through the achievement of granular 2D facial feature tracking with 3D correlations present. These 3D measurements were made from a singular, stationary camera, which is markedly distinct from stereo set-ups with multiple cameras or the rotation of a singular camera through a range of angular positions. Additionally, the participants were consistently moving throughout each session, while many studies rely on a static target. These constraints are valuable, as they prevent interference with the overall behavior of the participant and prioritize cost-effective hardware.

One potential solution to the absolute distance limitation could be a decrease in the linear assumption to the data. In this way, there would be several anchor step value points at which the distance is absolutely known, and then the amount of steps away from this anchor point would represent a linear movement from the known distance. Through this method, instead of assuming the depth values progress linearly across the entire range of steps, a few specific values would be chosen as a reference, and depth would then be approximated as a local linear regression. Since Figure 2 demonstrates up to 2 inch uncertainties across the step range at distances as close as 28 inches, this anchored method presents a way to achieve immediate 7% resolution improvements.

This limitation in absolute distance also motivates a future investigation with much closer distances. By decreasing the absolute distance to the participant, the depth of field correspondingly falls, which increases the precision in depth resolution. In conducting a close-up evaluation of these same facial feature tracking models, the realistic nature of a camera far from the subject would be sacrificed in favor of an iterative process that measures the threshold proximity needed to resolve individual facial feature depths.

Future work using this set-up spans multiple disciplines. From a 3D perspective, the experiment could be redesigned to increase granularity of the shape-from-focus calculations. Recording participants at a shorter distance from the camera could help decrease the depth-of-field limitation at the cost of a more narrow field-of-view, with a greater emphasis on centering the camera on the individuals' face. An additional sensor such as a TOF sensor or an additional camera in a stereo configuration would allow for comparison of depth results to ground truth, enabling a more robust analysis with respect to other state-of-the-art studies. In terms of HRC applications, a more in-depth analysis of the relative movement of the 478 facial features could be valuable to determine which features are most indicative of mental and physiological conditions.

7. Conclusion

In this study, a multifaceted facial detection and landmark tracking algorithm built on popular techniques in literature was successfully evaluated in the real-world conditions of a HRC experiment. While experiencing significant limitations in terms of depth resolution of individual facial features and the estimation of absolute distance to the face in cases of big movements, this study presents a valuable application of facial tracking techniques combined with low-cost technologies. The experimental pipeline successfully measured aggregate movement in all 3 dimensions and achieved higher granularity measurements when constrained to the xy-plane.

Overall, this accessible set-up enables the leveraging of AI and low-cost hardware for a wide range of future investigations in the HRC field.

Conflict of Interest The authors declare no conflict of interest.

Acknowledgments This project was graciously supported by Pensievision Inc, the College of Charleston (CofC) Office of Undergraduate Research and Creative Activities (URCA), as well as the National Institutes of Health's (NIH) South Carolina IDeA Networks of Biomedical Research Excellence (SC INBRE). Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number 5P20GM103499. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

[1] J. R. Saadon, F. Yang, R. Burgert, S. Mohammad, T. Gammel, M. Sepe, M. Rafailovich, C. B. Mikell, P. Polak, S. Mofakham, "Real-time emotion detection by quantitative facial motion analysis," Plos one, **18**(3), e0282730, 2023, doi:10.1371/journal.pone.0282730.

[2] A. Syeda, L. Zhong, R. Tung, W. Long, M. Pachitariu, C. Stringer, "Facemap: a framework for modeling neural activity based on orofacial tracking," Nature Neuroscience, **27**(1), 187–195, 2024, doi:10.1038/s41593-023-01490-6.

[3] T. F. Cootes, C. J. Taylor, "On representing edge structure for model matching," in Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, volume 1, I–I, IEEE, 2001, doi:10.1109/CVPR.2001.990655.

[4] P. A. Tresadern, M. C. Ionita, T. F. Cootes, "Real-time facial feature tracking on a mobile device," International Journal of Computer Vision, **96**, 280–289, 2012, doi:10.1007/s11263-011-0464-9.

[5] Y. Wu, Q. Ji, "Facial landmark detection: A literature survey," International Journal of Computer Vision, **127**, 115–142, 2019, doi:10.1007/s11263-018-1097-z.

[6] F. Dornaika, J. Orozco, "Real time 3D face and facial feature tracking," Journal of real-time image processing, **2**, 35–44, 2007, doi:10.1007/s11554-007-0032-2.

[7] J. Lorenzo, O. Déniz Suárez, M. Castrillon, C. N. Guerra Artal, "Comparison of focus measures in face detection environments," in ICINCO 2007-4th International Conference on Informatics in Control, Automation and Robotics, Proceedings, 2007, doi:10.5220/0001644604180423.

[8] A. Diego, M. Abou Shousha, "Portable Anterior Eye Segment Imaging System for Teleophthalmology," Translational Vision Science & Technology, **12**(1), 11–11, 2023, doi:10.1167/tvst.12.1.11.

[9] P. Minetola, M. Khandpur, L. Iuliano, F. Calignano, M. Galati, L. Fontana, "In-situ monitoring for open low-cost 3D printing," in Recent Advances in Manufacturing Engineering and Processes: Proceedings of ICMEP 2021, 49–56, Springer, 2022, doi:10.1007/978-981-16-3934-0_7.

[10] M. Leo, A. Natale, M. Del-Coco, P. Carcagnì, C. Distante, "Robust estimation of object dimensions and external defect detection with a low-cost sensor," Journal of Nondestructive Evaluation, **36**(1), 17, 2017, doi:10.1007/s10921-017-0395-7.

[11] H. Wolfe, M. Peljhan, Y. Visell, "Singing robots: How embodiment affects emotional responses to non-linguistic utterances," IEEE Transactions on Affective Computing, **11**(2), 284–295, 2017, doi:10.1109/TAFFC.2017.2774815.

[12] C. Oechsner, D. Ullrich, "Designing Dynamic Robot Characters to Improve Robot-Human Communications," arXiv preprint arXiv:2303.05219, 2023, doi:10.48550/arXiv.2303.05219.

[13] C.-L. Hwang, B.-L. Chen, H.-T. Syu, C.-K. Wang, M. Karkoub, "Humanoid robot's visual imitation of 3-D motion of a human subject using neural-network-based inverse kinematics," IEEE Systems Journal, **10**(2), 685–696, 2014, doi:10.1109/JSYST.2014.2343236.

[14] A. Chiurco, J. Frangella, F. Longo, L. Nicoletti, A. Padovano, V. Solina, G. Mirabelli, C. Citraro, "Real-time detection of worker's emotions for advanced human-robot interaction during collaborative tasks in smart factories," Procedia Computer Science, **200**, 1875–1884, 2022, doi:10.1016/j.procs.2022.01.388.

[15] A. Khalifa, A. A. Abdelrahman, D. Strazdas, J. Hintz, T. Hempel, A. Al-Hamadi, "Face recognition and tracking framework for human–robot interaction," Applied Sciences, **12**(11), 5568, 2022, doi:10.3390/app12115568.

[16] G. Tang, S. Asif, P. Webb, "The integration of contactless static pose recognition and dynamic hand motion tracking control system for industrial human and robot collaboration," Industrial Robot: An International Journal, **42**(5), 416–428, 2015, doi:10.1108/IR-03-2015-0059.

[17] C.-L. Hwang, Y.-C. Deng, S.-E. Pu, "Human-robot collaboration using sequential-recurrent-convolution-network-based dynamic face emotion and wireless speech command recognitions," Ieee Access, 2022, doi:10.1109/ACCESS.2022.3228825.

[18] M. Quintana, S. Karaoglu, F. Alvarez, J. M. Menendez, T. Gevers, "Three-d wide faces (3dwf): Facial landmark detection and 3d reconstruction over a new rgb–d multi-camera dataset," Sensors, **19**(5), 1103, 2019, doi:10.3390/s19051103.

[19] Arducam, "Pi Hawk-Eye: 64mp ultra high-RES camera for Raspberry Pi," *Arducam*, Dec. 2023. [Online]. Available: https://www.arducam.com/64mp-ultra-high-res-camera-raspberry-pi/

[20] H. Nakajima, Optical design using Excel: Practical Calculations for Laser Optical System, John Wiley and Sons, 2015.

[21] S. Pertuz, D. Puig, M. A. Garcia, "Analysis of focus measure operators for shape-from-focus," Pattern Recognition, **46**(5), 1415–1432, 2013, doi:10.1016/j.patcog.2012.11.011.

[22] J. Carson, B. Carson, S. Esener, K. Liu, D. Melnick, C. E., "Method, System, Software and Device for Remote, Miniaturized, and Three-Dimensional Imaging and Analysis of Human Lesions; Research and Clinical Applications," U.S. Patent No. 20190090753, March 2019.

[23] V. Bazarevsky, Y. Kartynnik, A. Vakunov, K. Raveendran, M. Grundmann, "Blazeface: Sub-millisecond neural face detection on mobile gpus," arXiv preprint arXiv:1907.05047, 2019, doi:10.48550/arXiv.1907.05047.

[24] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A. C. Berg, "Ssd: Single shot multibox detector," in Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14, 21–37, Springer, 2016, doi:10.1007/978-3-319-46448-0_2.

[25] I. Grishchenko, A. Ablavatski, Y. Kartynnik, K. Raveendran, M. Grundmann, "Attention mesh: High-fidelity face mesh prediction in real-time," arXiv preprint arXiv:2006.10962, 2020, doi:10.48550/arXiv.2006.10962.

[26] W. Chen, D. Gu, "Real-time physical fatigue risk assessment for construction workers using a teacher-student training paradigm," Automation in Construction, **177**, 106372, 2025, doi:10.1016/j.autcon.2025.106372.

[27] Y. Yu, H. Li, X. Yang, L. Kong, X. Luo, A. Y. Wong, "An automatic and non-invasive physical fatigue assessment method for construction workers," Automation in Construction, **103**, 1–12, 2019, doi:10.1016/j.autcon.2019.02.020.

[28] Y. Kartynnik, A. Ablavatski, I. Grishchenko, M. Grundmann, "Real-time facial surface geometry from monocular video on mobile GPUs," arXiv preprint arXiv:1907.06724, 2019, doi:10.48550/arXiv.1907.06724.

[29] NumPy Developers, "numpy.corrcoef," *NumPy v2.0 Manual*. [Online]. Available: https://numpy.org/doc/stable/reference/generated/numpy.corrcoef.html

[30] E. I. Obilor, E. C. Amadi, "Test for significance of Pearson's correlation coefficient," International Journal of Innovative Mathematics, Statistics & Energy Policies, **6**(1), 11–23, 2018.

[31] D. Gong, J. Yang, L. Liu, Y. Zhang, I. Reid, C. Shen, A. van den Hengel, Q. Shi, "From Motion Blur to Motion Flow: A Deep Learning Solution for Removing Heterogeneous Motion Blur," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, doi:10.48550/arXiv.1612.02583.