https://doi.org/10.2352/EI.2026.38.6.ISS-280

Tracking Complex Gestures with Multimodal Sensors

Hsu-Wei Chen, University of California, San Diego, La Jolla, California, USA
Yang Cai, Visual Intel Studio, Pittsburgh, PA, USA

Abstract

Human gestures in the real-world are complex, ranging from sign language, to full body motion, to extremely dynamic poses such as crawling and dancing. This study examines a set of multimodal sensory fusion methods to support the real-time operation and training without the need for wearable equipment. We articulate the gesture tracking sensor modality based on the gesture tracking types, accuracy, detection latency, distance, key point requirements, and accuracy with LiDAR, webcam and inertial measurement unit (IMU) for complex gesture recognition. We applied the methodology to applications of gait detection and tracking in a high altitude, sign language detection, and background noise removal in a crewed space. Our experiments show that the usability of the multimodal interfaces can be tested in a simulated environment and measured with instruments objectively.

Introduction

Human gesture tracking has been a challenge in perceptual computing because of the complexity of structural dynamics, broader contexts, and diverse environments and applications, ranging from sign languages, surgical operations, cholygraphy, emergency responses, medical diagnoses, to physical training [1].

Conventional human pose estimation solutions apply a 'one-size-fits-all' model by training an artificial neural network and using each frame of a video as the input [2], which produces a generalized solution, but the model struggles with fringe cases like small humans in distance, low resolution, and low contrast. Furthermore, no additional information is used in the solution outside of the collected real-time data.

Existing game engines often use gamepad [5], headset-mounted Innertial Motion Unit, or RGB cameras to monitor the user's gesture [10]. However, they do not fully capture the physical motion, e.g. only the arms or a half body. The visual and physical effects are not realistic.

In this study, we explore gesture capture and interaction technology for complex gestures with multimodal sensors, including regular digital cameras, near-infrared cameras or LiDAR (Light Detection and Ranging), and other sensors such as the camera pose estimation to improve the pose estimation in altitude and in distance. We explore how to combine the multimodal sensors in terms of multimodal dual-use and multimodal sensory fusion.

We evaluate the methods with measurements of tracking latency, accuracy, coverage, keypoints, and cost. In this study, we investigate hand gesture tracking, extreme poses from standing to crawling, under normal lighting or low lighting conditions. We also explore the multimodal sensory fusion for indoor and outdoor environments with controlled and uncontrolled real-world scenarios.

Limitations of Single Model Tracking

Many devices can be used for tracking gestures. Augmented Reality and Virtual Reality goggles normally contain wearable sensors such as cameras, and Innertial Motion Unity (IMU) for detecting and tracking user's gesture, mainly hand movement. But they don't track the full body gestures, especially walking and crawling. Wearable IMU sensors can be programmed for gesture detection and tracking. However, it is challenging to use IMU sensors for detecting and tracking complex gestures such as crawling and walking backwards.

The camera-based tracking algorithms are able to detect and track complex gestures with the augmented skeleton outlines, including the joints of the limbs and landmarks on the head position and orientation, e.g. finger gestures. The computational load for the machine-learning algorithm is rather high, especially for laptops. Figure 1 shows the helmet-mounted google with the IMU-based head tracking (left) and the camera-based hand tracking (right). For specific applications, such as firefighters' navigation and medical students' anatomy training, the single modal tracking would be sufficient, but not enough for a broader spectrum of applications.

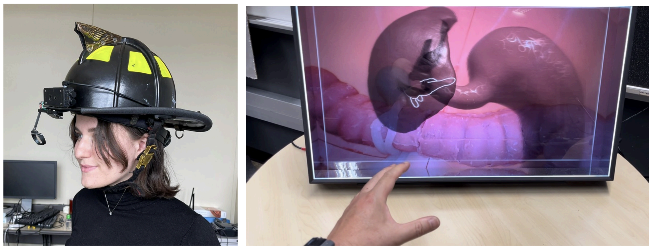

Figure 1 *The motion sensor (IMU)-based head tracking on the helmet (left) and the camera-based hand tracking on the table (right)*

The near-infrared- camera-based sensor (LiDAR), on the other hand, can deliver the similar skeletal tracking results with relatively lower computation because the LiDAR sensor fuses the data and processes the data onboard. A typical LiDAR sensor contains two NIR cameras and distance measurement that enables 3D gesture tracking.

Lighting condition is also a critical factor. IMUs and thermal cameras can track the pose without lighting. The gesture tracking with these two types of sensors are more difficult because of low-resolution sensory data. On the other hand, the near-infrared LiDAR sensors can track gestures in low-lighting. In this case, LiDAR could be a good option for the resolution and lighting considerations. Table 1 summarizes the comparisons among gesture tracking sensors in terms of computing load, lighting requirement, pose detection and tracking fidelity, and the cost.

Table 1. Comparisons among sensors

Sensors	Computing	Lighting	Fidelity	Cost
IMU	Simple	Not Required	Low	Low
Thermal	Moderate	Not Required	Low	High
LiDAR	Moderate	Required	High	Moderate
Webcam	Complex	Required	High	Low
Phone	Complex	Required	High	Low

In some special applications, multiple tracking sensors in a single mode are fused for 3D or full body tracking, for example, the 4D video capture systems [30] need 3 to 20 docent tracking sensors around the target. This requires visual feature matching and spatiotemporal registration with substantial computing power. In practice, this might be an overkill for many applications that do not need the high-resolution animated photorealistic 3D models.

Dual Modality Tracking

Gesture tracking sensors can be packaged with two or more modalities. We explore the devices that contain the due-modality of LiDAR and RGB cameras, such as RealSense and Kinect. We track the user's sign language gestures and the extreme gestures for comparisons.

Sign Language Gesture Tracking

Hand gestures, such as sign languages, can also be a useful application for training and learning purposes. We conducted an experiment with a webcam based system to test out our algorithm's ability of recognizing several Italian hand gestures. The example hand gestures we include here are "To eat", "To drink espresso" and the well-known hand sign of "*Ma che vuoi*?" which means "What do you want?", all shown in Figure 2. Often, those hand gestures pair with different facial expressions for a more accurate meaning the user likes to convey. The "*Ma che vuoi*?" gesture and the eating gestures form similar "cone" shapes, however, when the user is showing "to eat" he/she often puts his/her hand next to the mouth, so this can be an indicator of telling apart the gestures based on the reaction of other body parts, not just targeting the hand. The "cone shape" of the hand can be characterized as having the total length between four finger tips(index finger, middle finger, ring finger, and pinky finger) being much smaller than the total length of the knuckles of the four fingers.We can set a certain threshold to determine how much difference in length comparing the four finger tips tied together and their knuckles, indicating how spread out the fingers are. The greater the difference in length, the closer fingers are together.

The drinking espresso gesture is relatively simpler, with index finger and thumb being next to the mouth, and three other fingers on the side with a distance away, looking like an "OK" hand posture people see in the USA. The webcam based method is limited to 2D pixel-based input, however, a lot of Italian hand gestures require movement, sometimes with similar hand gestures but different movements can have completely different meaning. This is when a LiDAR based sensor can better accomplish, where the detected results are mapped into 3D coordinates, which includes the depth element to detect forward or backward motions(i.e. z-axis).

Figure 2 shows the skeleton overlay on the full body including detailed hand detection. Having the full body key points data we can easily include any details we need to classify different body languages with their paired hand gestures.

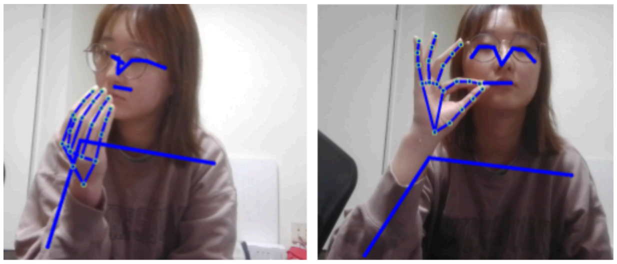

Figure 2. *Italian hand gestures of "to eat"(left) and "to drink espresso" (right) with the blue hand skeleton overlay*

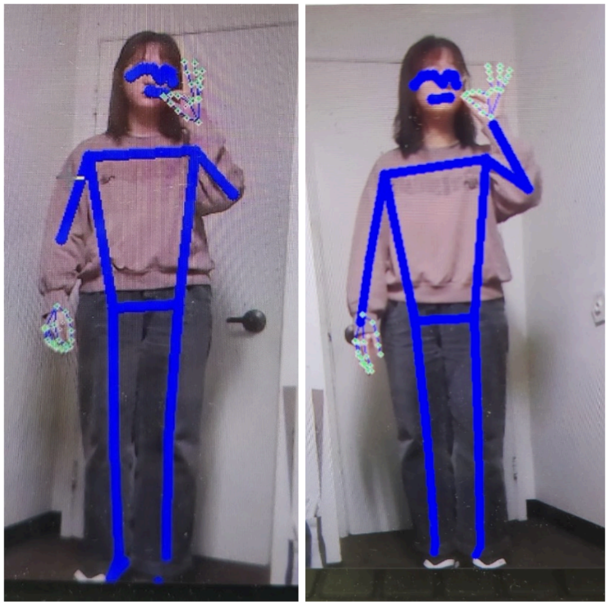

Figure 3. *Full body with the blue skeleton incomplete overlay (left) and improved overlay (right)*

The challenge is how to merge the hand tracking with finger motion with the full body motion. Figure 3 shows the full body tracking results. It is common that the overlaid skeleton is not complete. This is due to the distance between the user and the camera. After adjusting the distance, the tracking results are improved with the full skeletal overlay. The distance sensing can be a separate modality for improving gesture tracking accuracy. In another case, we tested the gesture tracking in a huge exhibition hall at a hotel. The crowd in front of the camera exceeds the tracking capacity. Again, after adjusting the distance between the subject and the camera, and adding the poster board as a background, we managed to track the single person's gesture in real time. Our results show that the method is able to track the user's gesture in multiple resolutions, including the full body (low-resolution) and the hands (high- resolution) and the important factor of the distance.

Extreme Gesture Tracking

In this study explore extreme gesture capture and interaction technology for emergency response training, which includes extreme gaits such as crawling through the floor and real-time gesture control of the user's movement in the training scenarios. Existing motion tracking systems heavily rely on wearable inertial motion units (IMUs) or head-mounted displays (HMDs), which adds both cost and physical burden to the user. We examine a new multimodal gesture tracking interface to support the real-time operation and training without the need for wearable equipment. We tested the duo-modality of gesture tracking with LiDAR and webcam separately for comparison. Figure 4 shows the extreme pose tracking of crawling, kneeing, and standing with LiDAR and Webcam.

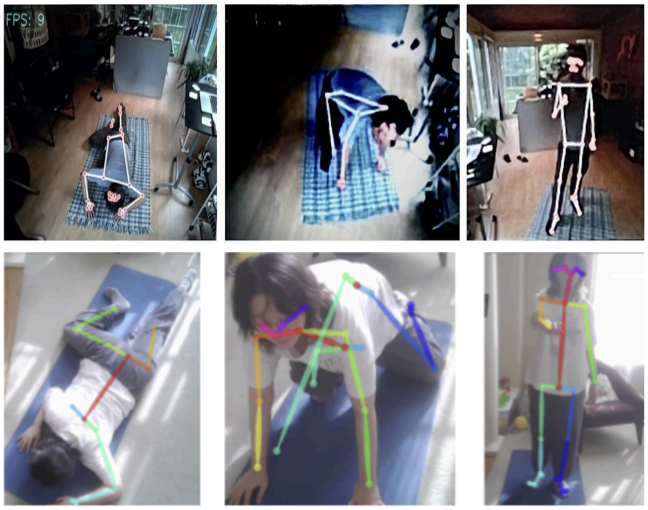

***Figure 4.** Extreme pose tracking of crawling, kneeing, and standing with LiDAR (up row) and webcam (bottom row)*

We measured gesture detection latency, key point requirements, and accuracy with LiDAR and webcam for complex gesture recognition. Our experiments show that the usability of the tracking interfaces can be tested in a simulated environment and measured with instruments objectively. We focused on the *detection latency* as an indicator of whether our gestures detection algorithm is suitable to operate in applications which require real-time response. The latency is defined as the time frame from the image being captured, overlaying skeleton key-points on the target's body, to the decision of the gesture classification. Here we consider four types of gestures and their corresponding detection latency and algorithm accuracy when using lidar or webcam: *standing still, walking forward, walking backwards*, and *head rotation angle* in horizontal direction representing the direction the user is facing.

The experiment was conducted indoors under natural lighting conditions provided by sunlight through the window. To set up the experiment, the user was standing at least 1.5 meters away from the sensor to ensure the sensor captured the whole body. During the experiment, the devices operated on battery power rather than being connected to a power outlet. The algorithm works under circumstances with dynamic gesture change, therefore, the experiment was taken in real time and the user moved their body to perform four movements each for around 10 seconds. That is — *walking forward and backward, standing still, and head turning*. For a lidar sensor, within a 7 second timeframe the sensor typically could capture and output up to 110 detection results. Using lidar and webcam, we collected two sets of data and summed up every latency data value characterized by different gestures in each set, and took the average value as the resulting latency. The formula used is shown in equation (1). To evaluate accuracy, we labeled gestures "*standing still*", "*walking forward*", and "*walking backward*" as separate classes. For each class *c*, we recorded the number of true positives (number of times class *c* was correctly predicted) and the number of false negatives (number of times class *c* was missed). The resulting accuracy was then calculated using the formula provided in equation (2). The experimental data is shown in Table 2.

$$L = \frac{\sum_{k=1}^{n}\left(latency_k\right)}{n} \tag{1}$$

$$A_c = \frac{TP_c}{TP_c + FN_c} \tag{2}$$

where L is the overall latency; A_c is detection accuracy by class; TP_c is the number of times class c was correctly predicted; FN_c is the number of times class *c* was missed.

When using a lidar camera the latency can be shortened at least 4 times then using a webcam for detection. The average human reaction time to a visual stimulus is approximately 180–200 ms; therefore, the latency introduced by the lidar camera can be considered acceptable. The accuracy can be affected by several reasons. The most common issue is lighting, which may affect stability of the skeleton overlay. Accuracy may also vary due to differences in walking habits; for example, distinguishing walking forward or backward becomes more difficult if the user's body naturally swings back and forth. In addition, detection smoothness/latency can further affect accuracy. As shown in the "algorithm accuracy" column, "standing still" and "walking forward" accuracy dropped by ~10% when we switched from lidar camera to webcam. In particular, when detecting "walking backward" using a webcam, the accuracy dropped to 42.53%. This was due to the detection lag, which made it difficult to distinguish changes in shoulder length associated with walking forward versus walking backward. However, the computational power of the device (i.e. the laptop which was used to run the program) must also be considered, as it directly affects system performance. For instance, we later observed that when the device was plugged in, the webcam (which was from the laptop) side had latency reduced to approximately 150 ms. Additionally, by using a more lightweight model for human skeleton detection could further improve latency.

Table 2 summarizes the detection latency and accuracy of LiDAR and webcam for gestures such as standing, walking forward, walking backward, and head rotation.

We compared the number of keypoint requirements for detecting the gesture types between the LiDAR and webcam. For example, for LiDAR: walking forward (2), walking backward (2), and head rotation (3); for webcam: walking forward (2), walking backward (4), and head rotation (3). The keypoint requirements for the LiDAR and Webcam are similar. LiDAR requires relatively fewer keypoints.

The results show that LiDAR has better accuracy than webcam. LiDAR also has much lower latency for tracking because the LiDAR device processes the raw data with the on-board processor before being transmitted to the host computer.

Table 2. Detection Latency and Classification Accuracy

Gesture	Sensor	Latency (ms)	Accuracy
Standing Still	Lidar	60.0	96.67%
Walking Forward	Lidar	60.7	93.33%
Walking Backward	Lidar	61.2	72.41%
Head Rotation	Lidar	60.4	_____*
Standing Still	Webcam	274.3	85.71%
Walking Forward	Webcam	291.1	81.94%
Walking Backward	Webcam	295.4	42.53%
Head Rotation	Webcam	301.8	_____*

Note: The head rotation is detected by relative value (i.e. how much the user is turning with respect to the "front" of their face reflects on the rotation value), therefore an absolute accuracy metric is not directly applicable in this case.

Multimodal Gesture Tracking

Gesture tracking is important to humanitarian assistance and disaster relief (HADR) missions. It is challenging to detect humans and gestures from a drone video in distance and in altitude due to low-resolution in human-figures, dynamic pose transition, and figure-to-figure and figure-to- background interaction. Conventional pose estimation methods often fail to register proper gaits from drone footages. In this study, we explore how to improve the detection of humans in distance and altitude where the target sizes are unknown [1]. The objective is to overlay the estimated human filter on the frames for human detection before gesture tracking. To accomplish this, we incorporate drone flight data to estimate the onboard camera's pitch angle and the distance to the target. Figure 5 shows a diagram of the algorithm.

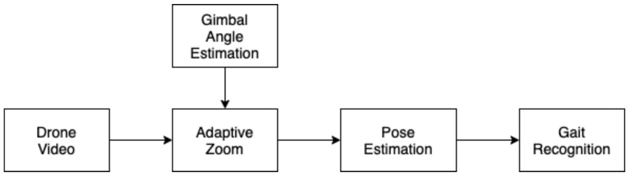

Figure 5. *Drone video fused with gimbal angle data for gait recognition in altitude and distance*

The spacing and orientation of keypoints is inconsistent in drone detections, because of the numerous possible camera and person locations. Calculating relative keypoint positions using background information can help make the data usable for gait prediction. For each keypoint we calculate the horizontal distance from the mean horizontal location and the vertical distance from the lowest keypoint in terms of the estimated torso length. To estimate the torso size of each person, we use the camera and landscape models shown earlier in human size estimation. However, if the person's torso angle is near parallel to the camera's line of view, the size will be underestimated. To partially solve this problem, an average torso size estimate of 45 cm is used when the calculated size is below a lower bound of 30 cm. These lengths were chosen from anthropometric data [19].

The next feature we extracted from the keypoint data is the velocity of the human detection. To calculate the person's velocity, we estimate vertical and horizontal distance scales in meters per pixel at the detection's location. Next, we measure the human's change in position by comparing the current average location of their keypoints with their previous positions. With these calculated values and the frames per second of the video, we calculate the person's velocity by dividing the estimated distance traveled by time.

To increase our gait recognition model's ability to differentiate between standing, walking, and running, we estimated movement frequency for each detection. Walking and running are cyclic movements, so the actions have a corresponding frequency. The discrete time fourier transform (DTFT) is an equation to transform a sequence of measurements from the time domain to the frequency domain [26]. To calculate a detection's movement frequency we apply the DTFT to a keypoint's relative positions and take the maximum frequency between horizontal and vertical motions.

The multinomial logistic regression method predicts the probability that a set of features falls into a certain category [27]. To classify gaits we use this model because of its explainability.

To further identify the deciding features for classification we utilize the LASSO technique. This method penalizes large coefficients in the loss function of the regression model. The geometric properties of this operator select for a sparse solution, which helps to identify important features [28].

We ran adaptive pose estimation on drone video and compared the results to standard pose detection to evaluate the approach's effectiveness. The testing set included 150 frames from three videos with different perspectives. The camera height, angle, and distance have the same range as in the gait recognition training dataset. The results show an increased correct detection percentage and a decreased missed detection percentage when compared to the method without predictive zoom. However, there was a small increase in the false detection rate. We defined 6 gaits in our experiments. That includes walking, running, standing, lying, sitting, and bend-over. To assess the effectiveness of our gait recognition model, we estimated the activity for 120 randomly selected samples not used in training. This testing dataset contained 20 examples for each class, and produced the results shown in Table 3 and 4.

Figure 6. *Gesture tracking results before the drone data fusion (left) and after the drone data fusion (right)*

Table 3. Confusion matrix of pose detection accuracy

150 frames	Detected Pose (before / after)	No Pose (before / after)
Pose	0.51 / **0.87**	N/A
No Pose	0.02 / **0.12**	0.47 / **0.81**

Table 4. Confusion matrix of recognition accuracy

	Stand	Walk	Run	Sit	Lie	Bend
Stand	1.00					
Walk		1.00				
Run		0.35	0.65			
Sit		0.05		0.95		
Lie					1.00	
Bend						0.90

Our results show that the pose estimation accuracy increases 36% and the recognition rate for six gaits reaches 91.7% on average. Figure 6 shows the results of the activity recognition before the drone data fusion and after the drone data fusion.

Conclusion

Human gestures in the real-world are complex, ranging from sign language, to full body motion, to extremely dynamic poses such as crawling and dancing. This study examines a set of multimodal sensory fusion methods to support the real-time operation and training without the need for wearable equipment.

We articulate the gesture tracking sensor modality based on the gesture tracking types, accuracy, detection latency, distance, key point requirements, and accuracy with LiDAR, webcam and inertial measurement unit (IMU) for complex gesture recognition.

We explore the due-use multimodality with LiDAR and webcam sensors for sign language and extreme gesture tracking. The results show that LiDAR has better accuracy than webcam. LiDAR also has much lower latency for tracking because the LiDAR device processes the raw data with the on-board processor before being transmitted to the host computer.

We applied the multimodal sensor fusion to applications of gait detection and tracking in a high altitude, sign language detection, and background noise removal in a crewed space. Our experiments show that the usability of the multimodal interfaces can be tested in a simulated environment and measured with instruments objectively. Our results show that the pose estimation accuracy increases 36% and the recognition rate for six gaits reaches 91.7% on average. Figure 6 shows the results of the activity recognition before the drone data fusion and after the drone data fusion.

Author Biography

Hsu-Wei Chen is a student at University of California, San Diego in the Master Degree Program of Computer Engineering. She is also Research Assistant of Visual Intel Studio (VIS) in the area of gesture recognition and human-computer interaction interfaces.

Dr. Yang Cai is Director of Visual Intel Studio (VIS) in the research area of Visual Computing and Extended Reality. He is the author of the books “Instinctive Computing” (Springer-London, 2016) and “Ambient Diagnostics” (CRC Press and Taylor-Francis, 2014) and editor of four volumes of Lecture Notes in Artificial Intelligence (Springer-Verlag).

Acknowledgements

The authors would like to thank the discussions with experts Dr. Mel Siegel, Scott Ledgerwood, Neta Ezer, Nick Molino, Justin Viverito, and Dennis Fortner. Thanks for the software development of the research assistants Justin Kiefel, Xiangyu Wu. The authors are grateful to the support from NIST PSCR and PSIAP and Northrop Grumman Corporation.

References

[1] Y. Cai., J. Kiefel, X. Wu, “Gait Recognition from Drone Videos”. In: Ahram T.Z., Falcão C.S. (eds) Advances in Usability, User Experience, Wearable and Assistive Technology. AHFE 2021. Lecture Notes in Networks and Systems, vol. 275. Springer, Cham, 2021
[2] B. Bhanu. and J. Han, “Human recognition at a distance in video, Advances” in Pattern Recognition book series, Springer, 2010
[3] Y. Cai, “Multimodal Extended Reality for Laparoscopic Surgery Training”, in Training, Education, and Learning Sciences, AHFE, 2024
[4] A. Poltorak, C. Hooper, and Y. Cai, “Holograms for Minimally Invasive Surgery Training and Planning”, in Healthcare and Medical Devices, AHFE, 2024
[5] Y. Cai,., “Episodic Memory with Interactive 3D Sequential Graph,” In Artificial Intelligence and Social Computing, AHFE, 2024
[6] Y. Cai, J. Park,. and L. Hope, “Articulated Spatial Audio for MIS Training”, in Human Factors in Virtual Environments and Game Design, AHFE, 2024
[7] Y. Cai, and T. Perez, “Haptic Perception with Artificial Tissues, in Digital Human Modeling and Applied Optimization, AHFE (2024)
[8] Y. Cai, and M. Siegel, “Field-of-View Studies for Stereo Interactions,” in Human-Centered Metaverse and Digital Environments, AHFE, 2024
[9] Y. Cai, “Rediscovering Symbolism in Extended Reality,” in Human-Centered Metaverse and Digital Environments, AHFE, 2024
[10] Y. Cai, Y. and M. Siegel, “Tangible extended reality with sensor fusion, “ proceedings of ERVR Conference, IS&T Electronic Imaging Conference at San Francisco, January 17-18, 2023
[11] Y. Cai, Y. and P. Freeman, “Real-Time Image Processing for Low-Vision Users,” presented at the Human Vision and Electronic Imaging Conference post-proceedings of the IS&T Electronic Imaging, 2023
[12] Y. Cai, “Wearable Spectrum Imaging and Telemetry at Edge,” In proceedings of the Imaging and Multimedia Analytics at the Edge Conference, IS&T Electronic Imaging, 2023
[13] Y. Cai, “ Mobile incident commanding dashboard (MIC-D),” proceedings of Mobile Devices and Multimedia: Enabling Technologies, Algorithms, and Applications, IS&T Electronic Imaging, San Francisco, January 17-18, 2023
[14] Y. Cai, “Learn on the Fly, in the volume “Advances in Human Factors in Robots, Drones and Unmanned Systems, 2021; 1210:56-63, published online 2020 May 31. DOI: 10.1007/978-3-030-51758-8_8 Published online by NIH, National Library of Medicine (2020) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7325289/
[15] S. Stokes, “Engineering Meets Extreme Reality,” CMU News,

Feb.17, 2022 https://engineering.cmu.edu/news-events/news/2022/02/17-extreme-reality.html
[16] NIST PSCR, Extreme Reality Telemetry for First Responders, Feb 23, 2023, https://www.nist.gov/news-events/news/2023/02/extreme-reality-telemetry-first-responders
[17] J. Thompson, A. Jain, Y. LeCun, and C. Bregler. “Joint training of a convolutional network and a graphical model for human pose estimation,” in NIPS, 2014.
[18] Wikipedia. Pinhole Camera Model, captured January 19, 2021 https://en.wikipedia.org/wiki/Pinhole_camera_model
[19] J. McConville, “Anthropometric Relationships of Body and Body Segments Moments of Inertia Anthropology Research Project, Inc. A/F Aerospace Medical Res 12-80
[20] CsvView Website link: https://datfile.net /index.html
[21] G. Olmschenk, H. Tang and Z. Zhu, “Pitch and Roll Camera Orientation from a Single 2D Image Using Convolutional Neural Networks," 2017 14th Conference on Computer and Robot Vision (CRV), Edmonton, AB, 2017, pp. 261-268, doi: 10.1109/CRV.2017.53.
[22] A. Kendall and C. Roberto, "Geometric loss functions for camera pose regression with deep learning." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017
[23] A. Kendall, M. Grimes and R. Cipolla, "PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization." Proceedings of the International Conference on Computer Vision (ICCV), 2015.
[24] E. Malis, M.l Vargas. “Deeper understanding of the homography decomposition for vision-based control. [Research Report] RR-6303, INRIA. 2007, pp.90.
[25] Z. Cao, G. Hidalgo, T. Simon, S. Wei, and Y.r Sheikh “OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields.” arXiv preprint arXiv:1812.08008, 2018.
[26] A.V. Oppenheim,, R.W. Schafer, and J.R. Buck, “Discrete-time signal processing” Upper Saddle River, N.J.: Prentice Hall. ISBN 0-13-754920-2, 1999.
[27] J. Engel,. "Polytomous logistic regression". Statistica Neerlandica . 42 (4): 233–252. doi:10.1111/j.1467-9574.1988.tb01238.x, 1988.
[28] R. Tibshirani, "Regression Shrinkage and Selection via the lasso," Journal of the Royal Statistical Society. Series B (methodological). Wiley. 58 (1): 267–88, 1996.
[29] J. Friedman, T. Hastie, and R. Tibshirani, “Regularization Paths for Generalized Linear Models via Coordinate Descent.” Journal of Statistical Software, 2010.
[30] Evercoast website. evercoast.com

JOIN US AT THE NEXT EI!

Imaging across applications . . . Where industry and academia meet!

• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

ADVANCING IMAGING SCIENCE
75
1947 IS&T 2022
imaging.org