Augmented Reality by Hand Gesture Recognized Commands and Movements

MEHDI VALINEJAD, *Student*. MOHAMMADYOUSEF SADRIALAMDARI, *Student*

***Abstract*—This project proposal explores the intersection of Human-Computer Interaction (HCI) and Augmented Reality (AR), focusing on the integration of Hand Gesture Recognition (HGR) and Marker-Based AR. Hand gesture recognition enhances user interactions with computing systems, providing a natural interface. The synergy with AR relies on effective HCI principles, offering immersive and user-friendly experiences.**

Planned methods include Landmark Detection with MediaPipe, ArUco markers for AR, and Homography Finding for perspective transformations. The expected results involve dynamic user experiences, controlling virtual elements through real-world hand movements.

The paper details a Python-based application design, incorporating landmark detection and hand gesture recognition, extended to AR using ArUco markers. The convergence of these technologies addresses limitations of traditional input methods, providing natural, efficient, and hands-free interaction within augmented environments.

In conclusion, this work signifies a step forward in interactive experiences, enriching HCI paradigms and paving the way for innovative applications that bridge the virtual and physical worlds. The fusion of landmark detection, ArUco markers, and hand gesture recognition promises novel possibilities in shaping the future of interactive technologies and user experiences.

***Index Terms*—Computer Vision, HCI, Hand Gesture Recognition, Augmented Reality, ArUaco Markers, Landmark, MediaPipe.**

I. Introduction

THE Human-Computer Interaction (HCI) represents a nuanced and interdisciplinary domain that intricately investigates the symbiotic relationship between humans and computer systems, with a primary focus on optimizing the design and utilization of technological interfaces. At its core, HCI delves into the intricacies of how individuals interact with computing devices, aiming to refine and elevate these interactions to levels of heightened effectiveness, efficiency, and overall user satisfaction.

Hand gesture recognition plays a significant role in Human-Computer Interaction (HCI) by providing a natural and intuitive means for users to interact with computers and devices. The relationship between hand gesture recognition and HCI lies in the application of gesture-based interfaces to enhance the way users communicate with and control technology.

Augmented Reality (AR) is a technology that overlays digital information and virtual objects on the real-world environment, creating an augmented view for users. Unlike virtual reality, which immerses users in a completely computer-generated environment, augmented reality enhances the real-world environment by adding computer-generated perceptual information.

The relationship between Human-Computer Interaction (HCI) and Augmented Reality (AR) is profound, as AR technologies heavily rely on effective HCI principles to deliver immersive and user-friendly experiences. Some of their interconnection aspects can be noted as User Interface and User Experience Design, Motion and Gesture Interaction, Spatial Awareness, Usability in 3D space, Feedback and user guidance, Context-Aware Computing, Accessibility in AR.

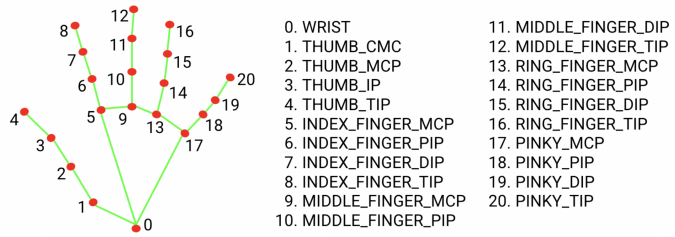
In recent years, the integration of cutting-edge technologies has propelled Human-Computer Interaction (HCI) to new heights, revolutionizing the way users engage with digital content. One captivating synergy within this realm lies at the intersection of Hand Gesture Recognition (HGR) and Marker-Based Augmented Reality (AR). This amalgamation presents a promising frontier for creating intuitive and immersive interactive experiences, where users can seamlessly navigate and manipulate augmented environments through natural hand gestures.

Hand gesture recognition, a pivotal facet of HCI, enables users to communicate with computing systems using natural and instinctive movements. Leveraging advanced computer vision algorithms, these systems interpret and respond to the intricacies of hand gestures, introducing an unparalleled level of intuitiveness to human-computer interactions. Concurrently, Marker-Based Augmented Reality employs identifiable markers, such as QR codes or images, as triggers to overlay digital content onto the physical world. This technology augments reality by blending virtual elements with the user's immediate environment.

The convergence of hand gesture recognition and marker-based AR holds great promise for advancing the field of HCI. By integrating gesture recognition into AR interfaces, users can interact with digital content in a manner that mirrors real-world actions, fostering a more intuitive and engaging experience. This marriage of technologies addresses the limitations of traditional input methods, offering a more natural, efficient, and hands-free mode of interaction within augmented environments.

This paper explores the synergy between hand gesture recognition and marker-based AR, examining the implications for user interface design, interactive applications, and the overall user experience. We delve into the technical intricacies of combining these technologies, exploring how gesture-driven interactions can enhance the usability and accessibility of marker-based AR systems. Moreover, we investigate the potential impact on diverse application domains, from gaming and education to industrial training and collaborative workspaces.

As we navigate through the intricate landscape of this technological convergence, we seek to unravel the syner-



The list of detected point of the hand by landmark hand detector.

Fig. 1. Landmark hand gesture detected joints

gies and challenges that arise when marrying hand gesture recognition with marker-based AR. By doing so, we aim to contribute to the ongoing discourse in HCI, providing insights that inform the design and development of more natural and immersive interfaces for augmented reality environments. The journey into this fusion promises to unlock novel possibilities, shaping the future landscape of interactive technologies and user experiences.

II. Methods and Tools

1. Landmark Detection

Landmark detection, also known as keypoint detection, refers to the process of identifying and locating specific points or features within an image or a visual scene. These points are often distinctive and can be used as references or landmarks (Fig.1) for various computer vision tasks. Landmark detection [1] is a crucial step in understanding and analyzing images, enabling applications in fields such as facial recognition, pose estimation, object tracking, and augmented reality [2].

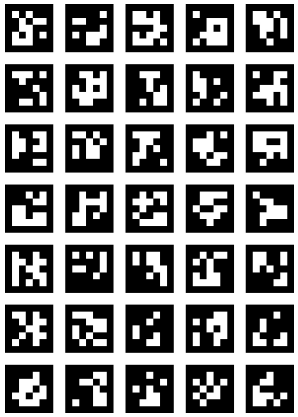
MediaPipe utilizes a specific model called the MediaPipe Hands for hand tracking, including hand landmark detection. The algorithm for hand landmark detection in MediaPipe Hands is based on a convolutional neural network (CNN) architecture.

The hand landmark model in MediaPipe can detect and track 21 3D landmarks on each hand, including key points such as the tips of the fingers, the base of the palm, and points along the hand contours. This enables precise tracking and recognition of hand gestures in real-time video or image streams.

2. ArUco markers

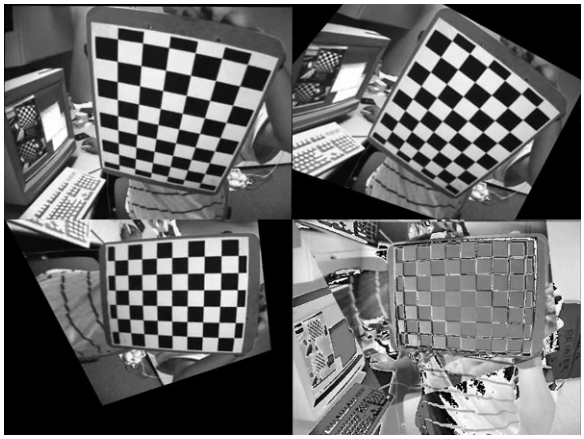
ArUco markers are a type of augmented reality marker designed for computer vision applications, particularly for tasks such as camera calibration and augmented reality (AR) systems. These markers are black-and-white square patterns that are easy to detect and identify in images or video frames. using markers we are able to detect critical points in an video stream frame, generally, markers are used for Visual Pattern Recognition, Unique identification, camera calibration and etc.

OpenCV (Open Source Computer Vision Library) provides a dedicated module for ArUco marker detection (Fig.2), and the implementation is based on the ArUco library. The ArUco module in OpenCV is specifically designed for easy integration of ArUco markers into computer vision applications.



Sample chart of ArUco Markers.

Fig. 2. ArUco Markers



Perspective sample of open cv homography detection.

Fig. 3. Homography

3. Homography Finding

To estimate the transformation matrix (homography matrix) that relates corresponding points in two images of the same scene. OpenCV offers builtin homography detection (Fig.3) methods that can be used for perspective point matching between real image and the way of residing in the the parent image visible area.

III. Expected Results

The expected result of landmark-based hand gesture detection combined with augmented reality (AR) can be a dynamic and interactive user experience where real-world hand movements are used to control and interact with virtual elements. Here's an overview of the potential outcomes:

Real-Time Hand Gesture Detection

Accurate and real-time detection of hand gestures using landmarks provides a natural and intuitive way for users to convey commands or interact with digital content.

Precise Tracking of Hand Poses

Landmark detection allows for precise tracking of the positions and orientations of key points on the hand, enabling accurate recognition of various hand poses and movements.

Virtual Object Manipulation

Users can interact with virtual objects overlaid on the real-world environment using their hand gestures. This could include grabbing, moving, rotating, or resizing virtual elements.

Gesture-Based Controls

Hand gestures can serve as a gesture-based control interface for AR applications. For example, a specific gesture might trigger an event, change the color or view, or initiate a specific action.

Immersive AR Experiences

Landmark-based hand gesture detection enhances the immersion of AR experiences by allowing users to engage with the virtual content in a more natural and hands-on manner.

Natural Interaction in AR Painting

In AR painting game, users can control characters or perform in-game actions using hand gestures detected by landmarks. This adds a layer of physicality and engagement to the painting experience.

Overall Expectation

The expected result is an interactive and immersive AR experience where users can use their hands and gestures to manipulate digital content, enhancing the naturalness and engagement of human-computer interactions in augmented reality.

IV. Method of implentation

Gesture Detection

The formalized design and development of a Python-based application represent a conscientious endeavor to harness real-time webcam video streams and exploit the functionalities of the MediaPipe library, incorporating a landmark hand gesture detection model. The objective of this application is to solicit user commands through discerning Victory and Thumb Up hand gestures. The framework further accommodates the tracking of the index finger tip's spatial coordinates on the screen, leveraging these movements to effectuate a drawing mechanism, metaphorically akin to the operation of a pen with the index finger tip serving as the point of contact.

This application is planned to stablish a robust command-handling mechanism within the detector module, facilitating the interpretation of user gestures and triggering specific actions accordingly. In particular, the application is going to be configured to capture the screen, inclusive of the ongoing drawing, upon the issuance of a specific command.

The intricacies of the application's design are rooted in its capacity to engage with the webcam's live video feed, an attribute that epitomizes real-time interaction. The utilization of the MediaPipe library augments this capability by seamlessly integrating landmark detection models, offering precise tracking of salient points on the user's hand, particularly focusing on the Victory and Thumb Up gestures.

Moreover, the nuanced tracking of the index finger tip is planned to emerge as a pivotal feature, enabling users to articulate their creative expressions by translating the finger's movements into a dynamic drawing interface. The application's gesture recognition module is adept at discerning predefined hand poses, such as Victory and Thumb Up, transforming them into actionable commands within the application's framework.

The comprehensive command-handling module is instrumental in orchestrating specific responses to detected gestures. In this context, the capability to capture screenshots encapsulating the current screen, inclusive of any ongoing drawings, underscores the versatility of this application as a tool for creative expression and documentation.

This formalized application, is going to blend Python programming language with advanced computer vision techniques to exemplify a paradigm shift in the realm of interactive systems. It not only underscores the convergence of webcam video stream analysis and landmark hand gesture detection but also signifies a synthesis of practical utility and creative expression. As technology continues to evolve, this formalized approach provides a testament to the adaptability and innovation inherent in applications that harmoniously integrate real-time video analysis and human-computer interaction.

Augmented reality

Augmenting the aforementioned application's capabilities, the incorporation of ArUco markers introduces an augmented reality (AR) dimension to the framework. This strategic integration enriches the user experience by seamlessly aligning the screenshots, captured through hand gesture commands, within the context of discernible ArUco markers. The symbiosis of these markers and the application's responsive mechanisms supports dynamic movement and perspective transformations, facilitated by rigorous homography calculations.

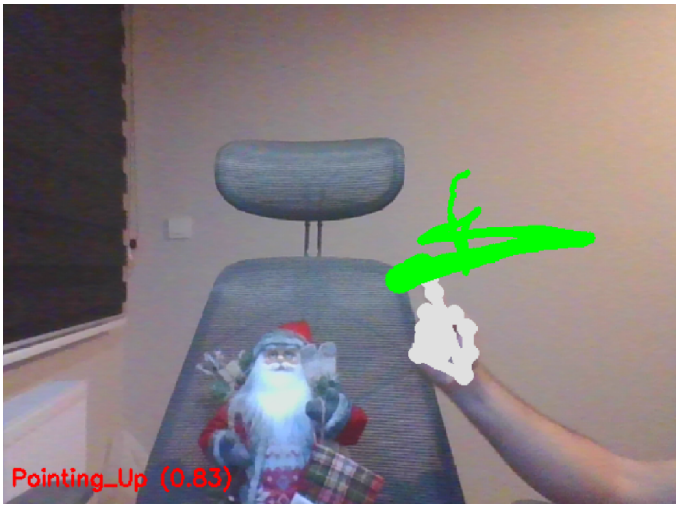
The introduction of ArUco markers serves as a pivotal element in bridging the virtual and physical realms within the AR paradigm. These markers, strategically placed within the physical environment, become reference points that anchor the AR content. The screenshots obtained through hand gesture commands seamlessly inhabit this augmented space, aligning with the markers and adapting to changes in movement and perspective.

The application, thus, leverages homography calculations to ensure a coherent and immersive AR experience. Homography plays a critical role in mapping the two-dimensional screen space to the three-dimensional world defined by the ArUco markers. This computational process enables the seamless integration of digital content, in this case, the captured screenshots, into the physical environment captured by the webcam.

The synergistic interplay of ArUco markers, homography calculations, and hand gesture commands unveils a sophisticated framework capable of not only recognizing and responding to user intent but also dynamically situating augmented reality content within the physical context. This intersection of computer vision, augmented reality, and human-computer interaction defines a technologically advanced and versatile application that transcends conventional boundaries.

Open Palm detected gesture by landmarks model.

Fig. 4. Gesture Detection



Poining up movement detection by application logic.

Fig. 5. Movement Detection

As technology perpetuates its evolution, this formalized integration of ArUco markers into the application underscores the capacity to propel user interactions beyond mere screen-based interfaces. It exemplifies a forward-thinking approach, wherein augmented reality becomes an integral facet of user engagement, fostering a dynamic and immersive interactive experience.

V. Acheived Results

Gesture Detection

Utilizing the pre-trained gesture recognition model within the OpenCV framework, (Fig.4), specifically designed for the recognition of the ”Thumb Up” and ”Victory” gestures, the application facilitates distinct commands, namely ”Take Screenshot” and ”Goodbye and Close Gently” respectively. When the ”Thumb Up” gesture is detected, the application captures the current frame of the webcam, appending the hand drawing to generate a screenshot. This image is then stored in memory for subsequent utilization by the Augmented Reality (AR) module.

Conversely, the ”Victory” gesture triggers the execution of the ”Goodbye and Close Gently” command. In response to this gesture, the application initiates a process wherein the screen is gradually faded to gray. The program then awaits a predefined count of frames before effecting the closure of the application. Furthermore, the application employs the tip of the index finger to mimic the functionality of a pen (Fig.5), capturing its movement for artistic purposes. The sequence of points representing the finger-tip pixels is stored in a two-dimensional array, forming the basis of the final painting. To optimize memory usage and computational efficiency, a dequeuing functionality has been implemented to retain only a specified number of recent coordinates of the finger tip. This sequence of points is depicted in a vivid green color, contributing to the creation of the visual masterpiece.

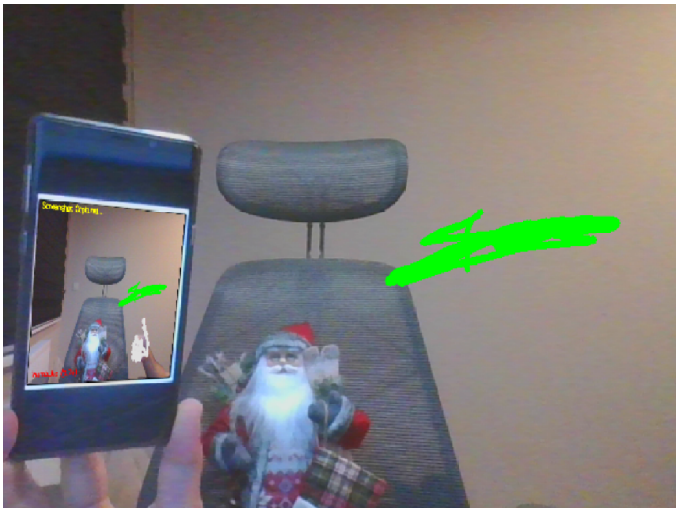
Augmented Reality

The acquired screenshot, resultant from the hand gesture and movement recognition module, is considered a virtual object within the computational framework. This virtual object is subjected to the requirement of aligning itself within the confines of four ArUco markers. Simultaneously, the Augmented Reality (AR) module undertakes the task of scrutinizing the input from the webcam to ascertain a correspondence between the identified markers and the captured screenshot. This matching process is accomplished through the application of OpenCV Marker detector methods.

During each frame analysis, the AR module diligently evaluates the webcam input, discerning the presence of four ArUco markers. Subsequently, the application’s underlying code engages in a sophisticated computation to calculate and seamlessly position the screenshot within the spatial bounds defined by the four markers. It is noteworthy that this alignment procedure is executed with meticulous consideration for perspective, ensuring a visually coherent and realistic integration of the virtual object into the augmented environment. This sophisticated synchronization mechanism enhances the immersive quality of the augmented reality experience, exhibiting a commitment to precision and realism within the framework of the application.

VI. Conclusion

The current instantiation of the application serves as a proof of concept, strategically designed to demonstrate the formidable synergy between Augmented Reality (AR) and Gesture Detection technologies. It is imperative to underscore that the existing algorithm, while illustrative of the conceptual integration, necessitates a comprehensive reevaluation to attain the level of refinement requisite for deployment in a mature, production-ready product. As a prototypical implementation, the application serves as a foundational framework, warranting iterative enhancements and optimizations (Fig.4).



Result of the application.

Fig. 6. Result

Critical to the advancement of this prototype is the recognition of performance bottlenecks inherent in the current implementation. Addressing these concerns demands a meticulous low-level revision, ideally executed in a more expeditious programming language. This imperative revision aims not only to rectify the performance issues plaguing the application but also to fortify its overall efficiency and responsiveness.

The envisioned trajectory of development involves an incisive reassessment of the algorithmic underpinnings, ensuring the alignment of the application with industry standards and user expectations. Concurrently, the migration to a swifter programming language constitutes a strategic imperative, poised to elevate the application's computational speed and responsiveness.

In essence, the present iteration stands as a foundational testament to the amalgamation of AR and Gesture Detection technologies, but its ultimate viability as a mature product necessitates a conscientious recalibration, both algorithmically and in terms of implementation, to fulfill the stringent requisites of a sophisticated and performant application.

In this research, we have explored the synergies of landmark detection, ArUco markers, and hand gesture recognition in the context of webcam video streams, leveraging the capabilities of the OpenCV library. The fusion of these technologies has paved the way for a multifaceted approach to enhancing Human-Computer Interaction (HCI), providing users with a more intuitive and immersive experience in real-time video environments.

The integration of landmark detection serves as the backbone of our approach, enabling precise tracking and recognition of key points on the human hand and face. This information is then seamlessly combined with the distinctive properties of ArUco markers, which act as visual anchors for accurate pose estimation and perspective transformations within the webcam video stream. Together, these components form a robust foundation for spatial awareness and object tracking in dynamic scenes.

Adding another layer of sophistication, hand gesture recognition brings a natural and gestural interface to our HCI paradigm. The ability to interpret and respond to hand movements in real time enhances the user's ability to interact with digital content in an instinctive and hands-on manner. Gestures such as Thumb Up/Down, Like/Unlike, Victory Sign can be translated into meaningful commands, expanding the range of possibilities for interactive applications within the webcam video stream.

Through our exploration, we have witnessed the potential applications of this fusion across diverse domains. In educational settings, users can manipulate 3D models or engage with virtual simulations through hand gestures and markers. In gaming environments, the combination of landmark detection and hand gesture recognition provides an immersive and interactive gaming experience. Moreover, the integration of ArUco markers facilitates accurate registration of virtual objects within the physical space captured by the webcam.

The significance of our approach extends to accessibility considerations, as gestural interfaces offer alternative interaction methods for individuals with physical disabilities.

Looking ahead, the fusion of landmark detection, ArUco markers, and hand gesture recognition in webcam video streams opens avenues for further research and development. Improvements in algorithmic efficiency, robustness, and the exploration of deep learning techniques may contribute to even more seamless and accurate interactions. Additionally, the integration of haptic feedback or voice commands could further enhance the multimodal nature of our HCI system.

In conclusion, our work signifies a step forward in the evolution of interactive experiences within webcam video streams. By combining landmark detection, ArUco markers, and hand gesture recognition with OpenCV, we have created a versatile and accessible example that empowers users to interact with digital content in a natural and engaging manner. This fusion not only enriches HCI paradigms but also paves the way for innovative applications that bridge the gap between the virtual and physical worlds. As technology continues to evolve, so too will the opportunities to redefine how we engage with and perceive the digital landscape.

REFERENCES

[1] G. Sánchez-Brizuela, A. Cisnal, E. de la Fuente-Lopez, J.-C. Fraile, and J. Pérez-Turiel, "Lightweight real time hand segmentation leveraging mediapipe landmark detection," *Virtual Reality*, 2023.

[2] F. Lin, B. Price, and T. Martinez, "Ego2hands: A dataset for egocentric two-hand segmentation and detection," 2021.

Mehdi Valinejad Received the B.S. degree in industrial engineering from the Azad University (South Tehran Branch) in 2012, and is currently working Master's. degree at the University of Bahcesehir at Istanbul.

Mohammadyousef Sadrialamdari Received the B.S. degree in Electrical-Electronics engineering from the Istanbul University - Cerrahpasa in 2023, and is currently working Master's. degree at the University of Bahcesehir at Istanbul.