# Real Time Cursor Control Using Hand Gestures

Dr. Deepali Newaskar[1], Ketan Kaiwart[2], Prasad Khanapure[3], Pratik Manjarekar[4]
[1, 2, 3, 4]*Computer Engineering, RMD Sinhgad School of Engineering, Pune, India.*

***Abstract*—"Real-Time Cursor Control Using Hand Gestures" is important because it establishes contactless connection between human–computer interface. It provides an environment in which physical touch cannot be taken as mandatory stuff and it is more effective project for pandemic conditions such as COVID-19. This paper shows important concepts for real-time cursor control using hand gestures by using Webcam since computer vision methods, such as color-based tracking that is achieved till 96% accuracy percentage in basic environment but for complex background accuracy level is low. Various frameworks like media pipe and CNNs are used for gesture recognition from which accuracy level goes up to 98% but still it gets affected by lighting and background fluctuations. To overcome these limits, Vision Transformer (ViTs) comes as an impactful alternative for improvement purpose in real time control system. It is a better and precise object tracking technology compare to old school technologies such as CNNs as it better for global context understanding. Actually, it has higher computational demand and also high data requirements despite that in human interaction ViTs shows significant advancement.**

***Index Terms*—Virtual Mouse, Gesture Recognition, Human-Computer Interaction (HCI), Computer Vision, OpenCV, CNN, Vision Transformer (ViT), Media Pipe.**

## I. INTRODUCTION

In this model we introducing the Vision Transformer based Cursor movement control that we are creating using some frameworks such as Media Pipe, Autopy and Vision based frameworks. Basic concepts such as machine learning and deep learning are used for this model and also for high accuracy we inserted Vision transformers for accuracy percentage.

Concurrently, deep learning, particularly Convolutional Neural Networks (CNNs), became the standard for complex pattern recognition. This models, trained on large sets of images, have demonstrated high accuracy in linking specific hand gestures (e.g., 'L' gesture or 'Paper' gesture) to different mouse functions.

However, traditional CNNs and other deep learning models can encounter performance difficulties with changing illumination and complicated backgrounds. This review surveys these existing methods and identifies a crucial research gap: the limited application of the Transformer architecture to this problem. Originally designed for Natural Language Processing, the Vision Transformer (ViT) has recently demonstrated state-of-the-art results and superior performance over comparable CNNs in image recognition tasks when sufficient data is available.

The ViT architecture, which processes images by converting them into a sequence of flattened patches and using a self-attention mechanism, is uniquely suited to overcome the limitations of current systems by capturing global contextual information and long-range dependencies. This paper posits that the ViT architecture, which has demonstrated high prediction accuracy (e.g., 93% for top-1 accuracy in one implementation) and reduced inference time through parallel processing, is poised to achieve a new state-of-the-art in the accuracy, speed, and robustness of gesture-based cursor control.

## II. REVIEW

Manav Ranawat et al. (2021) mentioned in the review paper "Hnad Gesture Recognition Based Virtual Mouse Events", it tells us that a system can operates just like virtual mouse does by using hand gestures in place of physical contact with hardware. It Operates with a Webcam to detect the hand movements and use of the tools like OpenCV and PyAutoGUI involves data managing and data processing for various operations. This Method controls conversion of the picture into HSV color domain for better understanding of human skin tones and better results under complex backgrounds even under fluctuation conditions. Webcam uses Haar Cascade Classifier to

detect the hand visible in the Webcam and also remove or hide the face/object (if visible) in the webcam by which the complete focus should be on hand itself, it also detects the hand outline for proper cauterization of hand. A CNN architecture uses to detect various hand movements. mentioned model in paper is learned on six gestures for which each gesture created using 1200 images during the training environment and it got the accuracy of 98.47%. all gestures were intact with a related mouse method for example, an 'L' gesture was guided pointer where to move, the paper gesture guided to execute a left-button click and the Rock gesture was guided to open a Notepad. The model also performed very well despite all different skin tones and background complications showing models practicality and good performance in actual real-world scenarios. In short, the paper talks about how vision-based movement platform can replace physical hardware for human-computer interactions.[1]

Gayatri Jagnade et al. (2023) mentioned in paper "Hand Gesture-based Virtual Mouse using Open CV" tells us that touchless function can be perform on the system using hand gestures for human-computer interaction through detecting hand movements, through eyes and also through facial gestures and even through objects that was colored items. It uses MediaPipe for Real-time hand movement detection of hand and facial expressions, OpenCV for visual process, and PyAutoGUI as a cursor manager and performing mouse control functions. In this setup, the index finger guides the cursor, while thumb and middle finger tell the system to trigger left and right clicks, yellow color indicates to scroll down the bar or also by opening the mouth. The System operates the volume of system by separation distance between fingertip and the interface of keyboard for entry. This model uses pyttsx3 library to keep on the operations such as scrolling, eye related gestures and using virtual keyboard and right-click movements. In short, Paper tells us that to make computer easier to use for the people who might feel difficult to use of keyboard and mouse; to provide a contactless system during the dangerous environment such as COVID-19 this model is very useful from safety purpose. [2]

Kalaiwani P.C.D et al. (2023) mentioned in the paper "Hand Gestures for Mouse Movements with Hand and Computer Interaction model"tells us that human can manage the computer interaction without any physical touch with the hardware system of computer it presents a vision-based virtual mouse controller that can do the cursor movements like we does with actual system using physical components just by captured hand gestures by Webcam. The model uses MediaPipe framework for recognizing hand gestures quickly, OpenCV for image processing and Autopy for performing mouse operations such as moving and clicking.There is step by step processing that is done by model such as firsly involves hand detection, focus on exact finger, Hand outlines and their coordinates, and clicks. The Webcam firstly focuses on right-hand movements as it is easy to render and more precise. This model is actually less expensive as well as the performance of the model is much on higher side but model needs to do verifications for multiperson engagement in the system, when number of persons occur on the Webcam some fluctuations problems arise, but This models state that human-computer interactions can be done without physical contact and can be done using touchless interface.[3]

Apoorva and colleagues et al. (2024) mentioned in the paper "Gesture-Controlled Virtual Mouse and Finger Air Writing" Demonstrate a framework that connects two methods which are movement based virtual control device with an air writing capability for human-computer interactions. The model uses Mediapipe for detecting hand positions and PyAutoGUI for guiding the cursor, executing clicks, and changes in sound levels. Different gestures like finger placements and to seperate the digits designed to start functions like left-right clicks and volume management. The air writing thing performs by CNN Architechture to identify characters and digits recognized in the air, a Tkinter based portal is used for to generate forecast to users, store and alter operational modes. This model uses colored tips of fingers by which webcam will get easy to focus on hand gestures without any fluctuations. In short, the paper tells that system can work smoothly and without any fluctuations but from This model the accuracy level cannot be easy to detect. Author proposes that without any physical contact with computer hardware we can connect with it by touchless interface. [4]

Balakrishnan N. et al. (2025), in their paper “Virtual Mouse Using Machine Learning,” describe a system that replaces a physical mouse with a virtual one. It uses a webcam or laptop camera to detect hand and finger movements. The system is made using Python,

with OpenCV for image processing and PyAutoGUI to perform computer actions like moving the cursor, clicking, and scrolling. The gesture recognition part works using MediaPipe and the Single Shot Detection (SSD) algorithm, which can find the hand and track 21 key points on it to understand different gestures. To make the system faster, hand detection happens only once at the start or when the hand goes out of view. The system is designed in 4 statges: Data Domain (hand gesture capturing and results), Data Conversion Domain (turning hand gestures into mouse actions), Operation Domain (detection, recognition and verification of hand gestures) and Interface Domain (for showing results). Specific hand postures and movements are allocated to functions: for instance, cursor manipulation is triggered when the forefinger is raised, a secondary selection when the thumb and index finger are raised near each other, and moving a standard shape (ID 4 and ID 8) to the left or right of the display manages sound level up/down. The deployment of the SSD-based virtual mouse yielded detection exactness of around 90% and processed pictures at 30-40 frames per second, confirming its viability for immediate touch-free engagement. Efficiency challenges arose with fluctuating light and intricate settings, however. Upcoming advancements encompass surmounting the present 25 cm range identification constraint by integrating an adaptive magnification feature, boosting hardware capacity, and facilitating Android gadgets. [5]

Prof. Monali Shetty et al. (2020) mentioned in paper "Virtual Mouse Using Object Tracking" it proposes a reduced-price virtual mouse setup employing a standard laptop webcam and distinct objects (like lids or strips) on the user's digits for monitoring. The apparatus is built using Python and the OpenCV visual processing package, depending on an HSV hue identification method to separate and follow the color-coded items instantly. The process involves capturing the webcam picture in RGB format, transforming it to HSV, and then enforcing a limit to generate a stencil for item recognition, utilizing smoothing techniques like Median Blur to remove disturbances from illumination. Cursor motion is attained by moving the objects apart from one another, while selecting, double-selecting, and pulling are accomplished by holding the objects near each other for specific frame counts. The setup demonstrated superb precision (96%) against a plain backdrop, though efficiency markedly declined with a complex backdrop. [6]

Elizabeth Rani. G et al. (2023) mentioned in paper "Cursor Movement Based on Object Detection Using Vision Transformers" it suggests a virtual mouse structure that employs Vision Transformers (ViT) for object recognition to allow real-time cursor operation, claiming its excellence over conventional deep learning frameworks such as CNNs and RNNs. ViT, a neural network design modified from NLP designs, evaluates a picture by transforming it into an ordered set of flattened segments and employing a self-attention method to grasp broad contextual data and extended relationships, which yields exact object identification and following. The system uses a model called Detection Transformer (DeTR), which can find objects in one go. It controls the cursor in two steps first, it detects and tracks the hand, and then it recognizes the hand's movement to move the cursor. Distinct hand motions are linked to mouse functions: an index digit for shifting the cursor and choosing elements, and a shut hand for pulling. The advantages of adopting ViT encompass enhanced precision, superior management of long-term links in video feeds, and diminished processing duration via simultaneous evaluation of patches. Nevertheless, the algorithm confronts a drawback owing to the considerable memory needs of ViT. [7]

Sudarshan S et al. (2025) mentioned in paper "Object Detection using Vision Transformer and Deep Learning for Computer Vision Applications" it describes the architecture and benefits of the Vision Transformer (ViT) model for image recognition and object detection. The Vision Transformer (ViT) is a type of neural network that uses the Transformer model which was originally made for language tasks and uses it for image recognition. Unlike normal image models, ViT doesn't use convolution layers. Instead, it divides an image into small square patches, flattens them into a sequence, and then uses a self-attention mechanism to understand how different parts of the image relate to each other. This helps the model see the whole image context and long-distance relationships between regions. Compared to regular Convolutional Neural Networks (CNNs), ViT offers several benefits: it's more accurate, easier to scale, works with different image sizes, and can understand global features better. In this proposed paper, a pre-trained ViT model from Hugging Face was used, which had already been trained on a large dataset

called ImageNet-21k. The workflow includes preparing the data (resizing, normalizing, and converting it into tensors), selecting and loading the model, testing it, and then deploying it for use. The prediction accuracy obtained for the proposed implementation was 93% for top-1 accuracy. The paper discusses that while ViT excels at modeling long-range dependencies, its main limitations include high computational and memory requirements and a lack of interpretability in its attention mechanisms. The authors suggest a broad range of applications for ViT, including video analysis, semantic segmentation, medical image analysis, and robotics/autonomous driving systems. [8]

Dosovitskiy et al. (2021) mentioned in paper "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale" the Vision Transformer (ViT) is presented, a framework that applies the conventional Transformer design initially created for natural language processing straight to series of picture segments for image categorization goals. In contrast to prior methods that merge self-attention with convolutional neural networks (CNNs), ViT discards convolutions completely, viewing pictures as ordered arrangements of set-sized segments (e.g., 16x16 pixels) and encoding them as input signals for the Transformer. The researchers show that while ViT performs worse than CNNs when trained on smaller collections due to an absence of image-specific inherent tendencies, it attains top performance when pre-trained on extremely large datasets (like ImageNet-21k and JFT-300M) and subsequently adjusted for smaller assessments. ViT surpasses analogous CNNs in precision and processing speed on sets such as ImageNet, CIFAR-100, and the VTAB collection, with the top model achieving 88.55% top-1 accuracy on ImageNet. The research also examines model growth, combined designs, and preliminary unsupervised learning, concluding that extensive data and model magnitude are vital for ViT's achievement, and indicating that Transformers can equal or exceed CNNs for visual assignments when adequate information exists. [9]

Berroukham et al. (2023) mentioned in paper "Vision Transformers: A Review of Architecture, Applications, and Future Directions" Provides study of vision transformers an evolution from original Transformer architecture in NLP to their computer vision applications. The limitations of CNN such as restricted receptive field and limited capacity. It explains how vision transformer addresses these challenges using self-attention mechanism and positional encoding. The Document Service key vision transformer design and their variations include Deit, Vit-Hybrid, Lite Transformer and Pyramid vision transformer. It compares Their performance, speed and suitability for different application and also cover major application areas such as image classification, object detection, semantic segmentation, anomaly detection and practical application in healthcare imaging and autonomous vehicles. The paper review examines the benefits of vision transformer, such as improve accuracy, interpretability. [10]

Sundaram et al. (2025) mentioned in paper "Comparative Study of Vision Transformer (ViT), Swin Transformer, and Transformer in Transformer (TNT) on Image Classification" Shows the Comparison of Vision Transformer (ViT), Swin Transformer, and Transformer in Transformer (TNT) for visual tasks. The analysis explores each model's structure: ViT views images as streams of distinct sections for broad context, Swin Transformer employs a graded organization with sliding apertures for effective local and global feature pulling, and TNT uses twin-level tokenizing to grasp subtle specifics within segments. All systems were educated with equivalent setup parameters and data enhancement methods to guarantee impartiality. Outcomes indicate that TNT attains the top correctness (89.7%) and resilience, succeeded by Swin Transformer (87.9%) and ViT (85.2%), with Swin Transformer presenting the optimal compromise between correctness and processing speed. The paper finishes that while TNT surpasses in outcome, it requires greater computation, and Swin Transformer is superior for instant usages. Restrictions involve employing low resolution pictures and a modest set, pointing toward additional exploration on bigger, high-resolution collections for a fuller appraisal. [11]

Verma et al. (2025) mentioned in paper "Patch Attention Excitation Based Vision Transformer for Small-Sized Datasets" PAEViT is presented, a new Vision Transformer structure intended to boost image classification results on limited datasets without needing prior training on vast datasets. PAEViT improves upon the standard ViT by integrating a Patch Attention Excitation unit, which adjusts attention values based on connections between patches and

selectively multiplies attention to concentrate on image areas pertinent to the objective. This model is low-overhead and modifies its attention process using a trainable excitation element, proving useful for datasets with few instances per category. Tests on CIFAR-10, CIFAR-100, and Tiny-ImageNet show that PAEViT reliably surpasses the base ViT and other leading small-data transformers, attaining up to 2–5% better precision based on the specific dataset and setup. Breakdown analyses verify the significance of the excitation unit and ideal setting choices, while visual representations reveal better feature pinpointing. The researchers finish that PAEViT supplies a flexible and effective answer for niche computer vision problems where extensive pretraining cannot be used. [12]

Gong et al. (2024) mentioned in paper "Multi-Scale Hybrid Attention Integrated with Vision Transformers for Enhanced Image Segmentation" proposes a novel Multi-Scale Hybrid Vision Transformer (MSH-ViT) architecture that integrates multi-scale convolutional attention with Vision Transformers (ViTs) to improve image segmentation, particularly for medical images such as retinal fundus scans. The model addresses the limitation of ViTs in capturing local details by combining convolutional layers of various kernel sizes (to extract fine and coarse features) with the global self-attention mechanism of ViTs, resulting in a hybrid attention map that leverages both local and global information. The architecture includes features like multi-scale convolutional transformer encoder and decoder with skip connection for a specific segmentation. Experimental results on three benchmark datasets is up to 0.960 for optic disc (OD) and 0.975 for optic cup (OC) segmentation. The study shows model's superior performance but it has some limitations such as computational requirements potential, generalizability because of dataset diversity and future need for validation in real world. The author advice for future work to focus on optimizing model efficiency, expand its applications.[13]

Table 1: Comparative Study of Existing Gesture-Based Systems

| Reference | Method/Model Used | Key Features | Accuracy/Performance | Limitations |
|---|---|---|---|---|
| [1] | CNN + OpenCV + Haar Cascade | Virtual mouse using hand gestures and HSV skin detection | 98.47% accuracy | Sensitive to lighting and complex backgrounds |
| [2] | MediaPipe + OpenCV + PyAutoGUI | Real-time hand tracking and cursor control | Real-time capable | Performance affected by background |
| [3] | MediaPipe + Autopy | Touchless cursor movement and clicking | Fast interaction in real-time | Multi-person detection issues |
| [4] | CNN + MediaPipe | Gesture control with air writing capability | Smooth interaction | Accuracy not clearly reported |
| [5] | MediaPipe + SSD | Virtual mouse with 21 hand landmarks | 90% detection accuracy, 30–40 FPS | Limited robustness under varying illumination |
| [6] | Object Tracking + OpenCV | Color object-based virtual mouse | 96% accuracy in simple backgrounds | Poor performance in complex backgrounds |
| [7] | Vision Transformer (ViT) + DeTR | Object detection-based cursor control | Improved object tracking performance | High memory requirements |
| [8] | Vision Transformer (ViT) | Transformer-based object detection | 93% accuracy | High computational cost |
| [9] | Vision Transformer (ViT) | Patch-based self-attention architecture | 88.55% ImageNet accuracy | Requires large-scale datasets |
| [10] | ViT Review Study | Review of transformer architectures | Improved global feature learning | Computational complexity |
| [11] | ViT vs Swin vs TNT | Comparative transformer analysis | TNT: 89.7%, Swin: 87.9%, ViT: 85.2% | Higher computation for TNT |
| [12] | PAEViT | Lightweight transformer for small datasets | 2–5% improvement over baseline ViT | Limited evaluation domains |
| [13] | Multi-Scale Hybrid ViT | Hybrid attention for segmentation | Dice score up to 0.975 | High computational requirements |

The Table 1 shows that traditional computer vision & CNN based gesture recognized system achieve effective cursor control under the good or controlled conditions. But this are having some limitations when actual real-time conditions are there. In ViT-based improved global contextual modelling & robustness. But it also has some challenges like computational.

## III. CONCLUSION

The study concludes that vision transformer development in gesture based virtual mouse control system set a bar for accuracy, flexibility, reliability in HCI. In previous computer vision techniques, the major issues that are address successfully by using Vit with self-attention mechanism. As compare Vision transformer with CNN-based & landmark-tracking system the better accuracy we get by Vision transformer. The Vit's is more user-friendly, contactless & approachable for computing. It is updated version for vision-based models. At the end, Vit's based virtual mouse system shows smooth, human friendly computer interface, future research can further develop this framework.

## REFERENCES


[1] M. Ranawat, R. Shankarmani, M. Rajadhyaksha, and N. Lakhani, "Hand Gesture Recognition Based Virtual Mouse Events," in 2021 2nd International Conference for Emerging Technology (INCET), Belgaum, India, May 21–23, 2021, pp. 1–4, doi: 10.1109/INCET51464.2021.9456388.

[2] G. Jagnade, M. Ikar, N. Chaudhari, and M. Chaware, "Hand Gesture-Based Virtual Mouse Using OpenCV," in 2023 International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT), 2023, pp. 820–825, doi: 10.1109/IDCIoT56793.2023.10053488.

[3] P. C. D. Kalaivaani, R. Kumaresan, A. J. Manow Ranjith, and S. Nandha Kumar, "Hand Gestures for Mouse Movements with Hand and Computer Interaction Model," in 2023 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, Jan. 23–25, 2023, doi: 10.1109/ICCCI56745.2023.10128273.

[4] Apoorva, R. Rani, A. Shankar, G. Jaiswal, A. Bondia, and A. Sharma, "Gesture-Controlled Virtual Mouse and Finger Air Writing," in 2024 14th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2024, pp. 370–375, doi: 10.1109/CONFLUENCE60223.2024.10463446.

[5] N. Balakrishnan, T. Krishnamoorthi, K. Aakash, and A. S. Sathishkumar, "Virtual Mouse Using Machine Learning," in 2025 3rd International Conference on Communication, Security, and Artificial Intelligence (ICCSAI), 2025, pp. 1831–1835, doi: 10.1109/ICCSAI64074.2025.11064183.

[6] M. Shetty, C. A. Daniel, M. K. Bhatkar, and O. P. Lopes, "Virtual Mouse Using Object Tracking," in 2020 5th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 2020, pp. 548–553, doi: 10.1109/ICCES48766.2020.9137854.

[7] G. E. Rani, M. Sakthimohan, S. S. Surya, S. Kalaiselvi, T. Bernice, and V. Gunasekran, "Cursor Movement Based on Object Detection Using Vision Transformers," in 2023 2nd International Conference on Vision Towards Emerging Trends in Communication and Networking Technologies (VITECON), 2023, doi: 10.1109/VITECON58111.2023.10157042.

[8] S. Sudarshan, A. Kavitha, Mohana, K. Nataraj, G. Ambika, and B. S. Sudhangowda, "Object Detection Using Vision Transformer and Deep Learning for Computer Vision Applications," in Proceedings of the 7th International Conference on Intelligent Sustainable Systems (ICISS-2025), 2025, pp. 1524–1529, doi: 10.1109/ICISS63372.2025.11076364.

[9] A. Dosovitskiy et al., "An Image Is Worth 16×16 Words: Transformers for Image Recognition at Scale," arXiv preprint arXiv:2010.11929v2 [cs.CV], 2021.

[10] A. Berroukham, K. Housni, and M. Lahraichi, "Vision Transformers: A Review of Architecture, Applications, and Future Directions," in 2023 7th IEEE Congress on Information Science and Technology (CiSt), 2023, pp. 205–210, doi: 10.1109/CiSt56084.2023.10410015.

[11] D. M. Sundaram, Y. Kasukurthi, and S. Jayachandran, "Comparative Study of Vision

Transformer (ViT), Swin Transformer, and Transformer in Transformer (TNT) on Image Classification," in 2025 5th International Conference on Soft Computing for Security Applications (ICSCSA), 2025, pp. 1469–1473, doi: 10.1109/ICSCSA66339.2025.11170958.

[12] A. Verma, S. R. Dubey, and S. K. Singh, "Patch Attention Excitation Based Vision Transformer for Small-Sized Datasets," in ICASSP 2025 – 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2025, doi: 10.1109/ICASSP49660.2025.10888545.

[13] Z. Gong and M. Chanmean, "Multi-Scale Hybrid Attention Integrated with Vision Transformers for Enhanced Image Segmentation," in 2024 2nd International Conference on Algorithm, Image Processing and Machine Vision (AIPMV), 2024, doi: 10.1109/AIPMV62663.2024.10691911.