

Real-Time 3D Virtual Object Manipulation Using MediaPipe Hands for Web Based Human Computer Interaction

Yenny Rahmawati^{1)*}, Danu Andrean²⁾, Kustanto³⁾

¹⁾ Prodi Rekayasa Komputer, Fakultas Teknik, Universitas Tiga Serangkai

²⁾ *Software engineer at Gameloft South East Asia*

³⁾ Prodi Rekayasa Komputer, Fakultas Teknik, Universitas Tiga Serangkai

¹⁾ yennyrahmawati@tsu.ac.id, ²⁾ danuandrianid@gmail.com, ³⁾ kustanto@tsu.ac.id

ABSTRACT

Real-time hand tracking is one of the challenges in the field of computer vision, particularly in supporting the development of intuitive and natural Human Computer Interaction (HCI) systems. This research aims to design and implement a system for manipulating three-dimensional (3D) virtual objects based on MediaPipe Hands and WebGL in a web-based environment. The system utilises an RGB webcam as an input device to detect 21 hand landmark points, which are subsequently processed into three main types of gestures: 'grabbing' for object translation, 'rotating' for object rotation, and the 'pinch' gesture for changing the object's scale. The system was evaluated by testing variations in the user's distance from the camera (0.5 m, 1.0 m and 1.5 m), lighting conditions (bright, dim and dark), and measuring system performance based on frame rate (FPS), Detection Confidence and Tracking Confidence. Test results showed that the system was capable of operating in real time at an average of 25–30 FPS, with Detection Confidence of 80–95 per cent and Tracking Confidence of 80–94 per cent under adequate lighting conditions. The best performance was achieved at distances of 0.5–1.0 m, whilst a decrease in accuracy occurred at greater distances, under low-light conditions, and in the presence of occlusion. The contribution of this research is the development of a browser-based 3D object manipulation system that integrates MediaPipe Hands, JavaScript and WebGL, enabling the entire process of hand detection, gesture recognition and object visualisation to be carried out markerlessly without the need for specialised hardware or a GPU accelerator. The research results indicate that the proposed system has the potential to be applied to various Human Computer Interaction applications, virtual simulations, virtual reality, and gesture-based interfaces.

Keywords: *MediaPipe Hands, Hand Tracking, Gesture Recognition, 3D Object Manipulation, Human Computer Interaction, WebGL*

I. PENDAHULUAN

Pemantauan, pengenalan, dan interpretasi gerakan alami manusia tanpa menggunakan perangkat tambahan merupakan salah satu topik yang berkembang pesat dalam bidang *Computer Vision* (CV) (Yadav et al., 2021), (Wang et al., 2021). Teknologi ini berperan penting dalam berbagai aplikasi, seperti *Human Computer Interaction* (HCI), realitas virtual (*Virtual Reality*), robotika, simulasi berbasis komputer, hingga sistem pelatihan interaktif (Sarafianos et al., 2016), (Pareek & Thakkar, 2021). Dalam konteks tersebut, *Human Pose Estimation* (HPE) dan *Human Action Recognition* (HAR) telah banyak dimanfaatkan untuk memahami serta menganalisis gerakan tubuh manusia secara otomatis. Meskipun demikian, pelacakan gerakan yang akurat masih menjadi tantangan, terutama pada bagian tubuh yang memiliki struktur kompleks seperti tangan dan jari. Tangan manusia memiliki banyak *degrees of freedom* sehingga mampu menghasilkan berbagai gerakan halus yang memerlukan metode deteksi dengan akurasi tinggi dan waktu respons yang cepat (Dubey & Dixit, 2023). Pendekatan awal menggunakan perangkat berbasis kontak, seperti data gloves, mampu memberikan hasil yang baik, namun memiliki beberapa keterbatasan, antara lain biaya yang tinggi, ketidaknyamanan pengguna, serta mengganggu gerakan alami tangan (Al-Faris et al., 2020).

Seiring berkembangnya metode *Deep Learning*, pendekatan *vision-based hand tracking* menggunakan kamera RGB menjadi alternatif yang lebih praktis. Salah satu

framework yang banyak digunakan adalah *MediaPipe Hands*, yang mampu mendeteksi 21 titik landmark tangan secara *real-time* dengan kebutuhan komputasi yang relatif rendah (Yasen & Jusoh, 2019). Framework ini telah banyak dimanfaatkan dalam berbagai penelitian maupun aplikasi, seperti pengenalan bahasa isyarat, navigasi antarmuka, pengendalian robot, hingga manipulasi objek virtual. Meskipun demikian, sebagian besar penelitian terdahulu lebih berfokus pada pengembangan algoritma pelacakan tangan atau klasifikasi gesture. Penelitian lain yang mengimplementasikan *MediaPipe Hands* untuk manipulasi objek 3D umumnya menggunakan aplikasi desktop, *game engine*, atau sistem berbasis GPU dengan kebutuhan perangkat keras yang relatif tinggi (Mohamed et al., 2021), (Ahmad et al., 2019). Selain itu, sebagian besar implementasi hanya berfokus pada satu jenis interaksi, misalnya rotasi atau translasi objek, tanpa mengintegrasikan berbagai jenis *gesture* dalam satu sistem yang dapat berjalan secara *real-time* melalui browser. Kondisi tersebut menunjukkan bahwa pemanfaatan *MediaPipe Hands* pada arsitektur berbasis web yang ringan, portabel, dan mampu mendukung manipulasi objek 3D secara interaktif masih relatif terbatas.

Berdasarkan kondisi tersebut, terdapat kesenjangan penelitian (*research gap*) pada pengembangan sistem manipulasi objek 3D yang tidak hanya mampu melakukan pelacakan tangan secara *real-time*, tetapi juga dapat diimplementasikan pada platform berbasis web tanpa memerlukan perangkat keras khusus, sensor tambahan, maupun instalasi aplikasi desktop. Selain itu, masih sedikit penelitian yang mengintegrasikan beberapa gesture utama, seperti **grabbing**, **rotating**, dan **zooming**, ke dalam satu sistem interaksi yang dilengkapi dengan pemantauan performa secara *real-time*, seperti nilai *frame rate* (FPS), tingkat confidence deteksi, dan status pelacakan. Oleh karena itu, penelitian ini mengusulkan sistem manipulasi objek virtual 3D berbasis web menggunakan *MediaPipe Hands* dan WebGL yang memungkinkan pengguna berinteraksi secara langsung dengan objek virtual melalui gesture tangan. Sistem memanfaatkan kamera RGB sebagai perangkat masukan untuk mendeteksi posisi tangan dan menginterpretasikan tiga gesture utama, yaitu *grabbing* untuk translasi objek, *rotating* untuk rotasi objek, dan *pinch gesture* untuk mengubah skala objek. Seluruh proses dilakukan secara *real-time* menggunakan browser sehingga sistem bersifat ringan, portabel, dan mudah diimplementasikan pada berbagai perangkat tanpa memerlukan sensor kedalaman maupun perangkat khusus.

Kontribusi utama penelitian ini adalah: (1) pengembangan sistem manipulasi objek 3D berbasis browser yang mengintegrasikan *MediaPipe Hands* dan WebGL secara *real-time*; (2) implementasi kontrol *multi-gesture* menggunakan dua tangan yang mencakup translasi, rotasi, dan perubahan skala objek; serta (3) penyediaan mekanisme pemantauan performa sistem secara langsung melalui parameter *frame rate*, *tracking confidence*, dan status interaksi sebagai bagian dari evaluasi kinerja sistem. Dengan demikian, penelitian ini diharapkan dapat menjadi alternatif yang ringan dan portabel dalam pengembangan aplikasi *Human Computer Interaction* berbasis gesture tangan.

II. TINJAUAN PUSTAKA

2.1 Pendekatan *Deep Learning* (DL) untuk pelacakan tangan

Pendekatan *Deep Learning* (DL) untuk pelacakan tangan pada kondisi nyata (*in-the-wild*) dapat diklasifikasikan berdasarkan jenis data masukan yang digunakan, yaitu citra RGB, peta kedalaman (*depth map*), atau kombinasi keduanya (RGB-D). Masing-masing pendekatan memiliki karakteristik, kelebihan, dan keterbatasan yang memengaruhi performa sistem dalam berbagai skenario aplikasi. Metode berbasis *depth* dikembangkan untuk memungkinkan rekonstruksi tangan secara tiga dimensi (3D), terutama seiring meningkatnya ketersediaan sensor kedalaman di pasaran. Pada umumnya, pendekatan ini

menggunakan *Convolutional Neural Networks* (CNN) untuk memproses data kedalaman dan mengekstraksi informasi posisi serta struktur tangan. Beberapa penelitian juga mengombinasikan pendekatan ini dengan aturan berbasis kinematika guna meningkatkan akurasi estimasi. Meskipun mampu memberikan hasil yang cukup presisi, metode berbasis *depth* memiliki beberapa kelemahan, seperti konsumsi energi yang tinggi, ukuran perangkat yang relatif besar, keterbatasan dalam mendeteksi objek pada jarak dekat, serta performa yang kurang optimal pada lingkungan luar ruangan akibat gangguan cahaya terhadap teknologi *Time-of-Flight* (ToF) (Mueller et al., 2017).

Pendekatan multimodal RGB-D menggabungkan keunggulan citra RGB dan data kedalaman. Dalam beberapa metode, citra RGB digunakan untuk mendeteksi posisi tangan dalam dua dimensi (2D), kemudian informasi kedalaman dimanfaatkan untuk mengestimasi posisi titik-titik sendi dalam ruang 3D. Alternatif lainnya adalah menggabungkan kedua jenis data tersebut secara langsung dalam satu model untuk menghasilkan estimasi secara simultan. Selain itu, kombinasi RGB-D juga sering dimanfaatkan dalam proses pelatihan model DL, meskipun pada tahap inferensi hanya menggunakan data RGB (Sarafianos et al., 2016).

Pada pendekatan berbasis RGB, ekstraksi *landmark* tangan dalam 2D menjadi bagian penting dalam berbagai metode *Human Pose Estimation* (HPE) modern, seperti *OpenPose* dan *AlphaPose*. Namun, pendekatan ini memiliki keterbatasan karena hanya mampu merepresentasikan gerakan tangan dalam bidang datar (2D). Untuk memperoleh informasi 3D, biasanya diperlukan beberapa kamera untuk melakukan triangulasi geometris. Meskipun efektif, metode ini meningkatkan kompleksitas sistem serta biaya implementasi (Reissner et al., 2019), (Amprimo et al., 2022).

Seiring perkembangan teknologi, beberapa penelitian terbaru berfokus pada estimasi langsung koordinat 3D dari video RGB tunggal (*monocular*). Berbagai arsitektur telah dikembangkan untuk mengatasi keterbatasan tersebut, namun sebagian besar masih menghadapi kendala dalam hal efisiensi komputasi dan implementasi real-time. Beberapa metode bahkan mengklaim performa tinggi, tetapi tidak menyediakan kode sumber untuk validasi lebih lanjut. Selain itu, untuk mencapai akurasi yang tinggi, pendekatan ini umumnya memerlukan perangkat keras dengan performa tinggi seperti GPU, sehingga kurang praktis untuk diterapkan di luar lingkungan penelitian.

Berdasarkan tinjauan tersebut, dapat disimpulkan bahwa pemilihan metode pelacakan tangan sangat bergantung pada kebutuhan aplikasi, khususnya dalam hal keseimbangan antara akurasi, efisiensi, dan kemudahan implementasi. Oleh karena itu, penggunaan pendekatan berbasis RGB yang ringan dan mampu berjalan secara *real-time*, seperti *MediaPipe Hands*, menjadi solusi yang relevan untuk aplikasi interaktif berbasis gestur tanpa memerlukan perangkat tambahan yang kompleks.

2.2 Gesture Recognition dan Gesture to Command Mapping pada Manipulasi Objek 3D

Gesture recognition merupakan salah satu komponen utama dalam sistem *Human Computer Interaction* (HCI) berbasis visi komputer. Teknologi ini memungkinkan gerakan tangan pengguna diterjemahkan menjadi perintah tertentu sehingga interaksi dengan objek virtual dapat dilakukan secara lebih alami dibandingkan penggunaan perangkat konvensional seperti mouse maupun keyboard (Rahmawati et al., 2025).

Dalam lingkungan virtual tiga dimensi, setiap gesture umumnya dipetakan (*gesture to command mapping*) menjadi aksi tertentu pada objek virtual. Beberapa penelitian menerapkan gesture menunjuk (*pointing*) sebagai navigasi, gesture mengepal (*grabbing*) untuk memindahkan objek, gerakan tangan terbuka untuk mengubah orientasi objek, serta gesture mencubit (*pinch*) untuk melakukan perubahan skala (*zooming*) (Chatzis et al.,

2020). Pendekatan ini memungkinkan pengguna berinteraksi secara langsung dengan objek virtual sehingga meningkatkan pengalaman pengguna (*user experience*) dalam berbagai aplikasi seperti simulasi, visualisasi data, pendidikan, maupun *virtual reality*.

Keberhasilan sistem manipulasi objek tidak hanya bergantung pada akurasi pendeteksian gesture, tetapi juga pada mekanisme pemetaan gesture terhadap aksi objek. Proses pemetaan yang sederhana dan intuitif dapat mengurangi beban kognitif pengguna sehingga interaksi menjadi lebih efisien. Oleh karena itu, penelitian ini menerapkan tiga gesture utama, yaitu *grabbing* untuk translasi objek, *rotating* untuk rotasi objek, serta *pinch gesture* untuk mengubah skala objek. Kombinasi ketiga gesture tersebut dipilih karena merupakan interaksi yang paling umum digunakan dalam manipulasi objek tiga dimensi.

2.3 MediaPipe Hands dan WebGL pada Sistem Interaksi Berbasis Browser

MediaPipe Hands merupakan framework berbasis *Deep Learning* yang dikembangkan oleh Google untuk melakukan pelacakan tangan secara *real-time* menggunakan kamera RGB. Framework ini menggunakan dua tahap utama, yaitu *Palm Detection* untuk mendeteksi lokasi tangan dan *Hand Landmark Detection* untuk mengestimasi 21 titik landmark tangan. Arsitektur tersebut memungkinkan proses inferensi dilakukan secara cepat dengan kebutuhan komputasi yang relatif rendah sehingga sesuai untuk aplikasi *real-time* (Madadi et al., 2022).

Di sisi lain, visualisasi objek tiga dimensi pada aplikasi berbasis web umumnya memanfaatkan WebGL, yaitu *Application Programming Interface* (API) berbasis JavaScript yang memungkinkan proses rendering grafis 3D langsung di dalam browser tanpa memerlukan perangkat lunak tambahan (Guo et al., 2022). Dengan memanfaatkan akselerasi GPU yang tersedia pada browser modern, WebGL mampu menghasilkan visualisasi interaktif dengan performa yang baik dan kompatibel pada berbagai sistem operasi.

Integrasi *MediaPipe Hands* dengan WebGL memberikan keuntungan dibandingkan pendekatan desktop konvensional. Seluruh proses deteksi tangan, pengenalan gesture, dan manipulasi objek dapat dilakukan langsung melalui browser menggunakan kamera RGB standar tanpa memerlukan instalasi perangkat lunak khusus maupun sensor tambahan. Karakteristik tersebut menjadikan sistem lebih ringan, mudah diakses, dan memiliki tingkat portabilitas yang tinggi untuk berbagai perangkat.

Pada penelitian ini, *MediaPipe Hands* digunakan sebagai modul utama untuk memperoleh koordinat landmark tangan, sedangkan WebGL dimanfaatkan sebagai media visualisasi dan manipulasi objek virtual tiga dimensi. Integrasi kedua teknologi tersebut memungkinkan pengguna melakukan interaksi secara langsung melalui gesture tangan dengan waktu respons yang mendekati *real-time*.

2.4 Perbandingan dengan Penelitian Terdahulu

Tabel 1 Perbandingan penelitian ini dengan penelitian sebelumnya.

Peneliti	Metode	Platform	Gesture	Objek 3D	Kelebihan	Keterbatasan
<i>OpenPose</i> (Chen et al., 2025)	CNN	Desktop	Ya	Tidak	Akurasi tinggi	Komputasi berat
<i>AlphaPose</i> (Pu et al., 2025)	CNN	Desktop	Ya	Tidak	Multi target	Membutuhkan GPU
<i>MediaPipe Hands</i> (Amprimo et al., 2024)	MediaPipe	Mobile/Web	Ya	Tidak	Ringan dan cepat	Tidak menyediakan manipulasi objek
Penelitian ini	MediaPipe Hands + WebGL	Browser	Grab, Rotate, Zoom	Ya	Ringan, portabel, tanpa instalasi, monitoring FPS dan confidence	Belum menggunakan depth camera

Berdasarkan Tabel 1, sebagian besar penelitian sebelumnya berfokus pada pengembangan algoritma pelacakan tangan atau implementasi manipulasi objek pada platform desktop menggunakan *Unity* maupun framework khusus. Penelitian tersebut umumnya memerlukan proses instalasi aplikasi, sumber daya komputasi yang lebih tinggi, atau perangkat keras tambahan. Sebaliknya, penelitian ini mengintegrasikan *MediaPipe Hands* dengan WebGL pada lingkungan berbasis browser sehingga seluruh proses pelacakan tangan dan manipulasi objek dapat dilakukan secara *real-time* menggunakan kamera RGB tanpa perangkat tambahan. Selain mendukung tiga jenis gesture utama (*grabbing*, *rotating*, dan *zooming*), sistem juga menyediakan pemantauan performa secara langsung melalui nilai FPS, tingkat confidence, dan status pelacakan. Dengan demikian, penelitian ini menawarkan solusi yang lebih ringan, portabel, dan mudah diimplementasikan untuk aplikasi *human computer interaction* berbasis web.

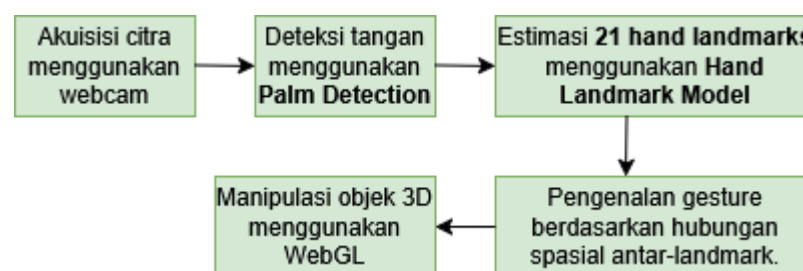
III. METODE PENELITIAN

3.1 Desain Penelitian

Penelitian ini menggunakan metode eksperimental yang bertujuan untuk merancang, mengimplementasikan, dan mengevaluasi sistem manipulasi objek virtual tiga dimensi (3D) berbasis *hand tracking* secara *real-time*. Sistem dikembangkan menggunakan *framework MediaPipe Hands* sebagai modul deteksi tangan dan WebGL sebagai media visualisasi objek 3D pada lingkungan berbasis web. Tahapan penelitian meliputi perancangan arsitektur sistem, akuisisi data, ekstraksi *hand landmarks*, pengenalan gesture, manipulasi objek 3D, dan evaluasi performa sistem.

3.2 Arsitektur Sistem

Arsitektur sistem terdiri atas lima tahapan utama sebagaimana ditunjukkan pada Gambar 1.



Gambar 1. Arsitektur Sistem

Pada tahap pertama, webcam menangkap citra video secara *real-time* dengan resolusi 640×480 piksel. Selanjutnya *MediaPipe Hands* mendeteksi posisi tangan dan menghasilkan koordinat tiga dimensi setiap landmark dalam bentuk.

$$L_i = (x_i, y_i, z_i), \quad i = 1, 2, 3 \dots 21 \quad (1)$$

Dengan x_i dan y_i merupakan koordinat citra yang telah dinormalisasi terhadap ukuran frame, sedangkan z_i menyatakan kedalaman relatif terhadap pergelangan tangan.

3.3 Akuisisi dan Pengolahan Data

Data diperoleh secara langsung melalui webcam RGB dengan resolusi **640×480 piksel** dan kecepatan akuisisi sekitar **30 frame per detik**. *MediaPipe Hands* digunakan untuk mendeteksi maksimal dua tangan secara simultan.

Framework MediaPipe Hands terdiri atas dua model utama:

- *Palm Detection* Model, yang mendeteksi lokasi tangan pada citra.
- *Hand Landmark* Model, yang mengekstraksi 21 titik landmark tangan.

Koordinat landmark yang diperoleh selanjutnya digunakan sebagai dasar dalam proses pengenalan gesture.

3.4 Pengenalan Gesture

Pengenalan gesture dilakukan berdasarkan hubungan geometris antar landmark tangan. Sistem menggunakan tiga jenis gesture utama, yaitu *Grabbing*, *Rotating*, dan *Pinch Zoom*:

a. *Gesture Grabbing*

Gesture *grabbing* digunakan untuk memindahkan posisi objek 3D. Pertama dihitung rata-rata jarak antara empat ujung jari terhadap pusat telapak tangan (landmark 0).

$$D_{avg} = \frac{1}{4} \sum_{i \in \{8,12,16,20\}} \sqrt{(x_i - x_0)^2 + (y_i - y_0)^2} \quad (2)$$

Gesture dianggap sebagai kepalan tangan (*fist*) apabila:

$$D_{avg} < 0,12 \quad (3)$$

Nilai 0,12 diperoleh dari implementasi sistem sebagai nilai ambang (*threshold*) untuk membedakan tangan terbuka dan tangan mengepal.

Selanjutnya dihitung jarak antara ibu jari (landmark 4) dan telunjuk (landmark 8).

$$D_{thumb-index} = \sqrt{(x_4 - x_8)^2 + (y_4 - y_8)^2} \quad (4)$$

Gesture *grabbing* dinyatakan aktif apabila memenuhi dua kondisi.

$$D_{avg} < 0,12 \quad \text{dan} \quad D_{thumb-index} < 0,12 \quad (5)$$

Jika kedua kondisi tersebut terpenuhi maka sistem mengaktifkan mode translasi objek menggunakan posisi telapak tangan sebagai referensi.

b. *Gesture Rotasi*

Gesture rotasi terjadi ketika tangan dalam kondisi terbuka dan mengalami pergerakan. Perubahan posisi titik jari digunakan untuk menghitung nilai rotasi objek pada sumbu tertentu. Perubahan posisi telunjuk dihitung sebagai:

$$\begin{aligned} \Delta x &= x_t - x_{t-1} \\ \Delta y &= y_t - y_{t-1} \end{aligned} \quad (6)$$

Nilai tersebut digunakan untuk menghasilkan sudut rotasi objek:

$$\begin{aligned} R_x &= k_x \Delta x \\ R_y &= k_y \Delta y \end{aligned} \quad (7)$$

dengan,

$$k_x = k_y = 10 \quad (8)$$

Rotasi hanya dilakukan apabila tangan tidak berada dalam kondisi *grabbing*.

c. *Gesture Zoom (Pinch)*

Gesture zoom dideteksi dari jarak antara ibu jari dan jari telunjuk. Jika jarak keduanya berada di bawah ambang tertentu, maka sistem akan melakukan proses pembesaran atau pengecilan objek 3D. Jarak antara ibu jari dan telunjuk dihitung menggunakan:

$$D_{pinch} = \sqrt{(x_4 - x_8)^2 + (y_4 - y_8)^2} \quad (9)$$

Gesture pinch dinyatakan aktif apabila:

$$D_{pinch} < 0,18 \quad (10)$$

Nilai 0,18 merupakan *threshold* yang digunakan pada implementasi sistem. Perubahan skala dihitung berdasarkan perubahan jarak *pinch* antar *frame*.

$$\Delta D = D_t - D_{t-1} \quad (11)$$

Kemudian perubahan skala objek dihitung sebagai:

$$S_{baru} = S_{lama} + 5(\Delta D) \quad (12)$$

Persamaan tersebut menghasilkan efek *zoom in* maupun *zoom out* secara halus (*smooth*)

3.5 Visualisasi Objek 3D

Objek 3D yang digunakan dalam penelitian ini berupa kubus yang divisualisasikan menggunakan teknologi WebGL. Parameter objek yang dikontrol meliputi posisi, rotasi, dan skala. Perubahan parameter dilakukan secara *real-time* berdasarkan gesture yang terdeteksi.

3.6 Implementasi Sistem

Sistem diimplementasikan menggunakan bahasa pemrograman JavaScript dengan dukungan teknologi berbasis web, yaitu HTML, Canvas, dan WebGL. *Framework MediaPipe Hands* digunakan melalui CDN untuk mempermudah integrasi dan meningkatkan portabilitas sistem.

3.7 Evaluasi Sistem

Evaluasi sistem dilakukan secara kualitatif dengan mengamati kinerja sistem dalam merespons gesture tangan pengguna. Parameter evaluasi meliputi:

- Responsivitas sistem terhadap gerakan tangan
- Stabilitas pelacakan *landmark*
- Akurasi pengenalan gesture
- Performa sistem berdasarkan nilai *frame rate* (FPS)

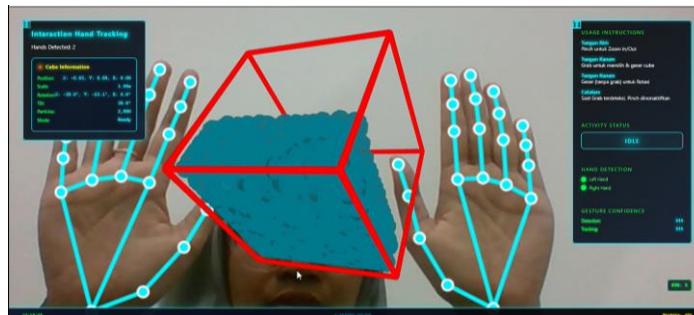
Selain itu, sistem juga menampilkan indikator seperti jumlah deteksi tangan, tingkat kepercayaan (*confidence*), serta status interaksi untuk membantu analisis kinerja secara *real-time*.

IV. HASIL DAN PEMBAHASAN

4.1 Hasil Implementasi Sistem

Sistem *Real-Time Hand Tracking for 3D Object Manipulation Based on MediaPipe Hands and WebGL* berhasil diimplementasikan menggunakan *framework MediaPipe Hands* sebagai modul utama deteksi tangan dan WebGL sebagai media visualisasi objek tiga dimensi berbasis browser. Sistem mampu mendeteksi hingga dua tangan secara simultan dengan menghasilkan 21 titik *landmark* pada masing-masing tangan secara *real-time*.

Antarmuka sistem terdiri atas tiga komponen utama, yaitu tampilan video dari webcam, visualisasi *landmark* tangan, dan objek kubus 3D yang dapat dimanipulasi menggunakan *gesture* tangan. Selain itu, sistem juga menampilkan parameter pemantauan secara *real-time*, meliputi nilai *Frame Rate* (FPS), *Detection Confidence*, *Tracking Confidence*, jumlah tangan yang terdeteksi, serta status interaksi yang sedang berlangsung. Implementasi sistem ditunjukkan pada Gambar 2.



Gambar 2. Dashboard sistem

Pengujian dilakukan menggunakan webcam beresolusi 640×480 piksel pada komputer dengan prosesor Intel Core i5 dan RAM 8 GB. Seluruh proses dijalankan menggunakan browser *Google Chrome* tanpa memerlukan perangkat tambahan sehingga sistem bersifat markerless, portabel, dan mudah diimplementasikan.

4.2 Hasil Pengujian Gesture

Pengujian dilakukan terhadap tiga jenis *gesture* utama, yaitu *grabbing*, *rotating*, dan *zooming*, dengan menggunakan tangan kanan dan kiri sesuai dengan fungsi yang telah dirancang. Pengujian dilakukan sebanyak 20 kali percobaan untuk setiap jenis *gesture*. Pengujian bertujuan mengetahui keberhasilan sistem dalam mengenali *gesture* dan mengeksekusi manipulasi objek. Hasil pengujian *gesture* dapat dilihat pada Tabel 1.

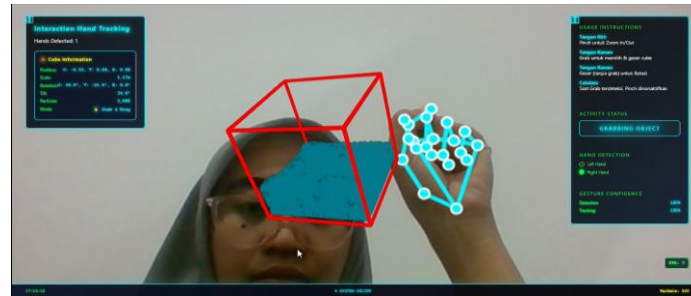
Tabel 1. Hasil Pengujian *Gesture*

<i>Gesture</i>	Jumlah Uji	Berhasil	Gagal	Persentase
<i>Grabbing</i>	20	19	1	96,67%
<i>Rotating</i>	20	18	2	93,33%
<i>Zooming</i>	20	19	1	96,67%

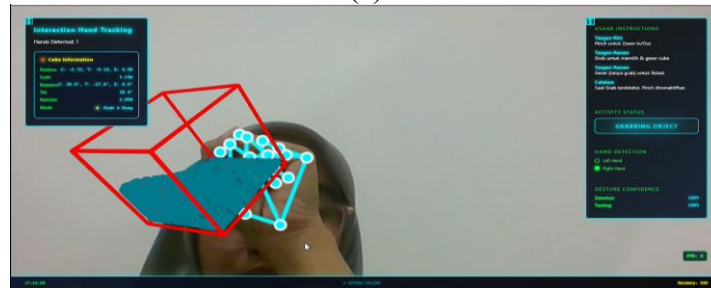
Hasil tersebut menunjukkan bahwa *gesture grabbing* dan *zooming* memiliki tingkat keberhasilan paling tinggi karena menggunakan threshold jarak yang relatif stabil. *Gesture rotating* memiliki tingkat keberhasilan sedikit lebih rendah akibat perubahan posisi tangan yang cepat sehingga menghasilkan perubahan koordinat *landmark* yang lebih fluktuatif.

a. *Gesture Grabbing* (Translasi Objek)

Gesture grabbing dideteksi ketika tangan dalam kondisi mengepal. Berdasarkan hasil pengujian, sistem mampu mengenali *gesture* ini dengan baik dan menggerakkan objek 3D mengikuti pergerakan tangan pengguna. Pergerakan objek bersifat relatif terhadap posisi awal tangan saat *gesture* dimulai. Namun, pada beberapa kondisi tertentu seperti gerakan tangan yang terlalu cepat, terdapat sedikit keterlambatan respons (*delay*) yang memengaruhi kelancaran interaksi. Hasil dari *grabbing* objek dapat dilihat pada Gambar 3, dimana pada Gambar 3a merupakan posisi objek sebelum dipindah, Gambar 3b merupakan gambar objek yang sudah berpindah.



(a)



(b)

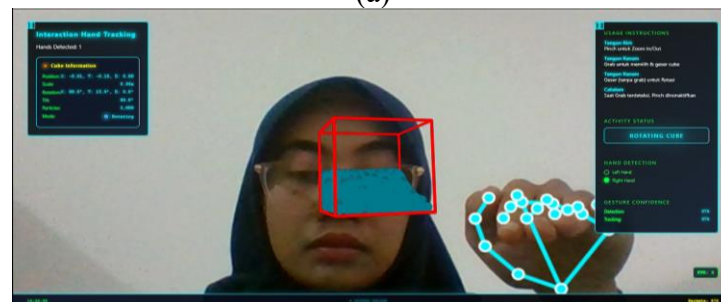
Gambar 3 Posisi objek a) sebelum di pindah dan b) setelah objek dipindah

b. Gesture Rotasi Objek

Gesture rotasi dilakukan menggunakan tangan terbuka dengan mendeteksi perubahan posisi jari. Hasil pengujian menunjukkan bahwa sistem mampu menghasilkan rotasi objek yang respon terhadap pergerakan tangan. Perubahan nilai rotasi dipengaruhi oleh perbedaan posisi koordinat jari (Δx dan Δy), sehingga semakin besar pergerakan tangan, semakin besar pula perubahan sudut rotasi objek. Pada Gambar 4 merupakan hasil dari rotasi objek, dimana objek dapat berputar sesuai dengan pergerakan tangan. Gambar 4a merupakan posisi objek sebelum dirotasi, sedangkan Gambar 4b adalah gambar setelah di rotasi.



(a)

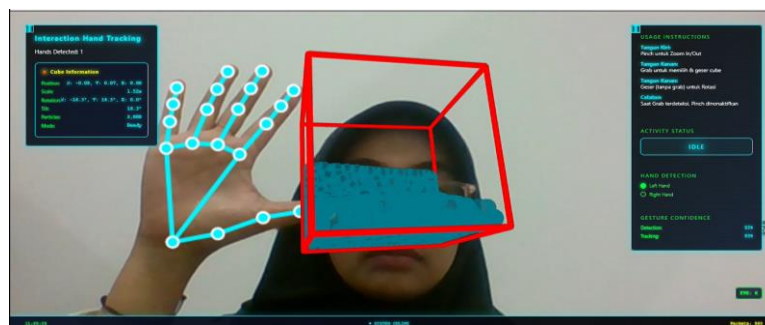


(b)

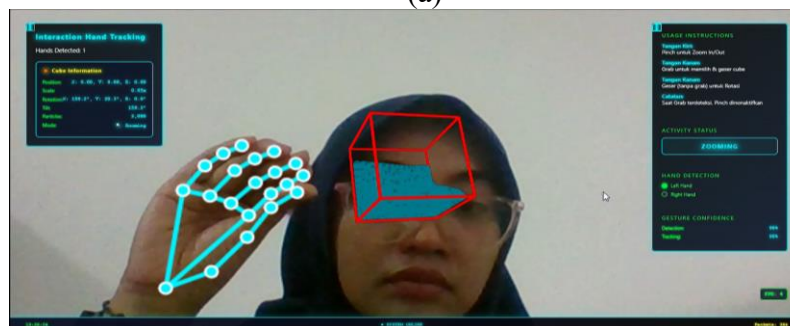
Gambar 4. Rotasi objek a) sebelum dirotasi dan b) setelah objek dirotasi

c. Gesture Zoom (*Pinch Gesture*)

Gesture zoom dideteksi berdasarkan jarak antara ibu jari dan jari telunjuk pada tangan kiri. Ketika jarak kedua titik tersebut berada di bawah ambang tertentu, sistem menginterpretasikannya sebagai perintah untuk memperbesar atau memperkecil objek. Hasil pengujian menunjukkan bahwa sistem mampu mengenali gesture ini secara stabil dan menghasilkan perubahan skala objek yang halus (*smooth*), terutama pada pergerakan dengan perubahan jarak kecil. Gambar 5 merupakan hasil dari gesture zoom, dimana ukuran objek dapat diperbesar dan diperkecil sesuai dengan jarak antara ibu jari dan telunjuk. Gambar 5a merupakan gambar dimana jarak antara ibu jari dan telunjuk berjauhan atau kondisi objek diperbesar. Gambar 5b adalah kondisi dimana jarak ibu jari dan telunjuk berdekatan atau diperkecil.



(a)



(b)

Gambar 5 Zoom objek a) objek diperbesar dan b) objek diperkecil

4.3 Pengaruh Jarak Kamera

Tabel 2 menyajikan hasil pengujian pengaruh jarak kamera terhadap akurasi pengenalan *gesture*. Pengujian dilakukan pada tiga variasi jarak, yaitu 0,5 m, 1,0 m, dan 1,5 m. Variasi jarak tersebut dipilih untuk menganalisis perubahan performa sistem dalam mendeteksi *landmark* tangan serta mengenali *gesture* pada kondisi pengguna berada pada jarak yang berbeda dari kamera.

Tabel 2. Hasil pengaruh jarak terhadap akurasi *gesture*

jarak kamera	<i>grabbing</i>	<i>rotating</i>	<i>zooming</i>
0,5 m	98%	96%	98%
1,0 m	96%	94%	95%
1,5 m	90%	88%	89%

Berdasarkan Tabel 2, sistem menunjukkan performa terbaik pada jarak 0,5–1,0 meter. Ketika jarak pengguna meningkat hingga 1,5 meter, ukuran tangan pada citra menjadi lebih kecil sehingga beberapa *landmark* tidak dapat dideteksi secara konsisten. Kondisi tersebut menyebabkan penurunan akurasi pengenalan *gesture* sekitar 6–8%.

4.4 Pengaruh Kondisi Pencahayaan

Tabel 3 menunjukkan hasil pengujian pengaruh kondisi pencahayaan terhadap kinerja sistem dalam mengenali *gesture* tangan. Pengujian dilakukan pada tiga tingkat pencahayaan, yaitu terang, redup, dan gelap, untuk mengevaluasi pengaruh intensitas cahaya terhadap proses deteksi *landmark* tangan dan akurasi pengenalan *gesture* secara *real-time*.

Tabel 3 Hasil pengaruh intensitas Cahaya

Kondisi	<i>Detection Confidence</i>	<i>Tracking Confidence</i>
Terang	95%	94%
Redup	83%	81%
Gelap	71%	68%

Hasil pengujian menunjukkan bahwa intensitas pencahayaan memberikan pengaruh langsung terhadap kualitas deteksi *landmark*. Pada kondisi terang sistem memperoleh nilai *confidence* rata-rata sebesar 95%, sedangkan pada kondisi redup nilai *confidence* menurun menjadi sekitar 81–83%. Penurunan tersebut disebabkan berkurangnya informasi tekstur tangan yang diterima kamera sehingga proses deteksi *landmark* menjadi kurang stabil. Begitu juga pada kondisi gelap nilai *confidence* turun menjadi 68–71%.

4.5 Performa Sistem

Tabel 4 menunjukkan hasil evaluasi performa sistem berdasarkan beberapa parameter utama, yaitu resolusi kamera, rata-rata *frame rate* (FPS), *detection confidence*, *tracking confidence*, latensi rata-rata, dan jumlah maksimum tangan yang dapat dideteksi secara bersamaan. Pengujian ini bertujuan untuk menilai kemampuan sistem dalam mempertahankan kinerja deteksi dan pelacakan tangan secara stabil pada lingkungan interaktif berbasis web.

Tabel 4. Hasil performa sistem

Parameter	Nilai
Resolusi Kamera	640×480
Rata-rata FPS	28,4 FPS
Detection Confidence	80–95%
Tracking Confidence	80–94%
Latensi rata-rata	63 ms
Maksimum tangan	2 tangan

Sistem memperoleh rata-rata 28,4 FPS, sehingga memenuhi kategori *real-time* (>24 FPS). Nilai tersebut menunjukkan bahwa integrasi *MediaPipe Hands* dan WebGL mampu mempertahankan performa interaktif tanpa memerlukan GPU khusus.

4.6 Perbandingan dengan Penelitian Sebelumnya

Tabel 5 Perbandingan dengan Penelitian sebelumnya

Penelitian	Platform	FPS	Gesture	Browser
<i>OpenPose</i> (Chen et al., 2025)	Desktop	18–22	Ya	Tidak
<i>AlphaPose</i> (Pu et al., 2025)	Desktop	20–25	Ya	Tidak
<i>MediaPipe Hands</i> (Amprimo et al., 2024)	Mobile	>30	Ya	Ya
Penelitian ini	Browser WebGL	25–30	<i>Grab, Rotate, Zoom</i>	Ya

Dibandingkan dengan *OpenPose* maupun *AlphaPose*, sistem yang dikembangkan memiliki keunggulan dari sisi implementasi karena seluruh proses dijalankan langsung melalui browser menggunakan JavaScript dan WebGL. Meskipun nilai FPS sedikit lebih rendah dibandingkan implementasi MediaPipe resmi pada perangkat mobile, sistem tetap mampu mempertahankan performa 25–30 FPS yang memenuhi kebutuhan interaksi *real-time*. Selain itu, penelitian ini tidak hanya memanfaatkan MediaPipe sebagai modul pelacakan tangan, tetapi juga mengintegrasikannya dengan mekanisme *gesture to command* untuk manipulasi objek 3D berbasis web secara langsung.

4.7 Analisa Stabilitas Sistem

Berdasarkan seluruh hasil pengujian, sistem menunjukkan bahwa *MediaPipe Hands* mampu memberikan performa yang stabil pada kondisi pencahayaan yang baik dengan nilai *detection confidence* mencapai 95%. Selain itu, rata-rata *frame rate* sebesar 28,4 FPS menunjukkan bahwa sistem telah memenuhi kebutuhan aplikasi interaktif secara *real-time*. Keterbatasan utama ditemukan ketika pengguna berada pada jarak lebih dari 1,5 meter, kondisi pencahayaan rendah, maupun ketika terjadi *occlusion* antartangan. Pada kondisi tersebut akurasi *gesture* mengalami penurunan karena beberapa *landmark* tidak berhasil dideteksi secara konsisten. Walaupun demikian, sistem tetap mampu mempertahankan manipulasi objek secara interaktif tanpa menggunakan *marker* maupun sensor tambahan sehingga sesuai untuk aplikasi *Human Computer Interaction* berbasis browser.

V. KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan hasil perancangan, implementasi, dan pengujian, penelitian ini berhasil mengembangkan sistem *Real-Time Hand Tracking for 3D Virtual Object Manipulation* berbasis *MediaPipe Hands* dan WebGL yang mampu melakukan manipulasi objek 3D secara interaktif menggunakan *gesture* tangan tanpa memerlukan perangkat tambahan (*markerless*). Sistem mampu mendeteksi hingga dua tangan secara simultan dengan memanfaatkan 21 titik *landmark* pada setiap tangan untuk mengenali tiga jenis *gesture* utama, yaitu *grabbing*, *rotating*, dan *zooming*. Hasil pengujian menunjukkan bahwa sistem dapat bekerja secara *real-time* dengan rata-rata performa 25–30 FPS, tingkat *Detection Confidence* sebesar 80–95%, serta *Tracking Confidence* sebesar 80–94% pada kondisi pencahayaan yang memadai. Pengujian juga menunjukkan bahwa performa terbaik diperoleh pada jarak pengguna 0,5–1,0 m dari kamera, sedangkan penurunan akurasi terjadi pada jarak yang lebih jauh maupun pada kondisi pencahayaan rendah.

Kontribusi utama penelitian ini adalah pengembangan sistem manipulasi objek 3D berbasis arsitektur web yang mengintegrasikan *MediaPipe Hands*, *JavaScript*, dan WebGL dalam satu lingkungan sehingga seluruh proses deteksi tangan, pengenalan *gesture*, dan visualisasi objek dapat dijalankan secara langsung melalui browser tanpa memerlukan instalasi perangkat lunak khusus maupun akselerator GPU. Hasil penelitian menunjukkan bahwa pendekatan tersebut mampu menyediakan antarmuka interaksi manusia-komputer yang ringan, portabel, dan responsif sehingga berpotensi diterapkan pada berbagai aplikasi interaktif berbasis gestur.

5.2 Saran

Berdasarkan hasil penelitian yang telah dilakukan, pengembangan selanjutnya dapat difokuskan pada beberapa aspek. Pertama, sistem dapat dikembangkan dengan mengintegrasikan informasi kedalaman (*depth*) menggunakan kamera RGB-D atau sensor kedalaman untuk meningkatkan akurasi estimasi posisi objek dan *gesture* dalam ruang 3D. Kedua, metode pengenalan *gesture* dapat diperluas dengan menerapkan pendekatan

berbasis *Deep Learning* yang mampu mengenali *gesture* yang lebih kompleks serta lebih adaptif terhadap variasi bentuk tangan, sudut pandang, dan kondisi pencahayaan. Ketiga, optimalisasi algoritma dan proses *rendering* perlu dilakukan untuk meningkatkan *frame rate*, menurunkan latensi, serta mempertahankan performa pada perangkat dengan spesifikasi yang lebih rendah. Selain itu, penelitian selanjutnya disarankan melakukan evaluasi kuantitatif dengan jumlah responden yang lebih banyak, variasi jarak pengguna, kondisi pencahayaan, dan berbagai spesifikasi perangkat untuk memperoleh analisis performa yang lebih komprehensif. Pengembangan lebih lanjut juga dapat diarahkan pada integrasi sistem dengan teknologi *Augmented Reality* (AR), *Virtual Reality* (VR), maupun aplikasi *Human Computer Interaction* (HCI) lainnya yang membutuhkan interaksi alami berbasis gestur tangan.

DAFTAR PUSTAKA

- Ahmad, A., Migniot, C., & Dipanda, A. (2019). Hand pose estimation and tracking in real and virtual interaction: A review. *Image and Vision Computing*, 89, 35–49.
- Al-Faris, M., Chiverton, J., Ndzi, D., & Ahmed, A. I. (2020). A review on computer vision-based methods for human action recognition. *Journal of Imaging*, 6, 46.
- Amprimo, G., Ferraris, G., Masi, G., Pettiti, G., & Priano, L. (2022). GMH-D: Combining Google MediaPipe and RGB-depth cameras for hand motor skills remote assessment. In *2022 IEEE International Conference on Digital Health (ICDH)* (pp. 132–141). IEEE.
- Amprimo, G., Masi, G., Pettiti, G., Olmo, G., Priano, L., & Ferraris, C. (2024). Hand tracking for clinical applications: Validation of the Google MediaPipe Hand (GMH) and the depth-enhanced GMH-D frameworks. *Biomedical Signal Processing and Control*, 96, 106508.
- Chatzis, T., Stergioulas, A., Konstantinidis, D., Dimitropoulos, K., & Daras, P. (2020). A comprehensive study on deep learning-based 3D hand pose estimation methods. *Applied Sciences*, 10(19).
- Chen, S., Fukuda, T., & Yabuki, N. (2025). Development of an occupancy measurement system for micro-zones within open office spaces based on multi-view multi-person 3D pose estimation. *Expert Systems with Applications*, 111, 113037.
- Dubey, S., & Dixit, M. (2023). A comprehensive survey on human pose estimation approaches. *Multimedia Systems*, 29, 167–195.
- Guo, S., Rigall, E., Ju, Y., & Dong, J. (2022). 3D hand pose estimation from monocular RGB with feature interaction module. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(8), 5293–5306.
- Madadi, M., Escalera, S., Baró, X., & González, J. (2022). End-to-end global to local convolutional neural network learning for hand pose recovery in depth data. *IET Computer Vision*, 16, 50–66.
- Mohamed, N., Mustafa, M. B., & Jomhari, N. (2021). A review of the hand gesture recognition system: Current progress and future directions. *IEEE Access*, 9, 157422–157436.
- Mueller, F., Mehta, D., Sotnychenko, O., Sridhar, S., Casas, D., & Theobalt, C. (2017). Real-time hand tracking under occlusion from an egocentric RGB-D sensor. In *2017 IEEE International Conference on Computer Vision (ICCV)* (pp. 1154–1163). IEEE.
- Pareek, P., & Thakkar, A. (2021). A survey on video-based human action recognition: Recent updates, datasets, challenges, and applications. *Artificial Intelligence Review*, 54, 2259–2322.

- Pu, L., Zhao, Y., Hua, Z., Han, M., & Song, H. (2025). Multi-target spraying behavior detection based on an improved YOLOv8n and ST-GCN model with interactive video scenes. *Expert Systems with Applications*, 262, 125668.
- Rahmawati, Y., Woraratpanya, K., Ardiyanto, A., & Nugroho, H. A. (2025). Evaluating the effectiveness of facial actions features for the early detection of driver drowsiness in driving safety monitoring system. *Communications in Science and Technology*, 10(1).
- Reissner, L., Fischer, G., List, R., Giovanoli, P., & Calcagni, M. (2019). Assessment of hand function during activities of daily living using motion tracking cameras: A systematic review. *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine*, 233(8), 764–783.
- Sarafianos, N., Boteanu, B., Ionescu, B., & Kakadiaris, I. A. (2016). 3D human pose estimation: A review of the literature and analysis of covariates. *Computer Vision and Image Understanding*, 152, 1–20.
- Wang, X., Garg, S., Tran, S. N., Bai, Q., & Alty, J. (2021). Hand tremor detection in videos with cluttered background using neural network-based approaches. *Health Information Science and Systems*, 9.
- Yadav, S. K., Tiwari, K., Pandey, H. P., & Akbar, S. A. (2021). A review of multimodal human activity recognition with special emphasis on classification, applications, challenges, and future directions. *Knowledge-Based Systems*, 223.
- Yasen, M., & Jusoh, S. (2019). A systematic review on hand gesture recognition techniques, challenges, and applications. *PeerJ Computer Science*.