Hand Gesture Recognition with Hand Landmarks Using MediaPipe

Muhammad Aidan Daffa Junaidi
1906300800
Department of Electrical Engineering, Computer Engineering
University of Indonesia
Depok, Indonesia
Muhammad.aidan@ui.ac.id

***Abstract*— *Gesture* merupakan salah satu komunikasi yang masuk ke dalam komunikasi kinesik, atau komunikasi yang meliputi gerakan tangan dan tubuh, yang artinya hand gesture merupakan bentuk komunikasi yang menggunakan tangan. *Gesture Recognition* bertujuan agar komputer bisa memahami gerakan manusia yang umumnya berasal dari tangan atau wajah. *Hand gesture recognition* dianggap penting, hal itu dikarenakan membuat kompter mengetahui bagaimana cara menafsirkan gerakan tangan. Mediapipe hadir sebagai *framework built-in machine learning* yang memiliki solusi untuk sistem pengenalan gerakan tangan. Dalam penelitian ini, penulis mengembangkan teknologi *hand gesture recognition* yang menggunakan hand landmark dalam implementasinya dengan memamfaatkan framework mediapipe. Penelitian ini berbasis *machine learning* yang artinya aplikasi akan diajari untk menafsirkan *hand gesture* secara *real-time* dari data yang ada.**

***Kata kunci*—Hand Gesture Recognition, Machine Learning, Object Detection, MediaPipe, Real-Time**

I. PENDAHULUAN

Dalam aktivitas manusia sehari - hari, kita berinteraksi dalam dua cara komunikasi. Yang pertama menggunakan komunikasi verbal yang menggunakan kata-kata untuk pengucapan. Selama komunikasi kedua tanpa bahasa. Komunikasi ini menekankan pada gerakan mata, gerak tubuh, dan ekspresi wajah. Bentuk komunikasi kedua ini lebih dikenal dengan komunikasi nonverbal. [1].

Bagi orang dengan pendengaran normal, berinteraksi menggunakan komunikasi verbal dan non-verbal bukanlah sesuatu yang sulit untuk dilakukan. Tapi untuk teman-teman disabilitas yang mempunyai keterbatasan pendengaran (tunarungu), mereka memilih menggunakan bahasa isyarat atau komunikasi nonverbal sebagai media interaksi baik dengan Teman Tuli atau dengan Teman Dengar (yang mempunyai pendengaran normal).

Tunarungu merupakan seseorang yang tidak dapat mendengar atau tuli. [2] Tunarungu memiliki keterbatasan dalam berkomunikasi dengan orang lain. Menurut KBBI Komunikasi merupakan sebuah proses pengiriman dan penerimaan sebuah pesan atau berita antara dua orang atau lebih sehingga pesan yang dimaksud dapat dipahami oleh pengirim dan penerima. [3] Proses komunikasi dapat berjalan dengan baik jika penerima dan pengirim memahami dan menggunakan bahasa yang sama. Jika antara penerima dan pengirim tidak memiliki keduanya maka akan terjadi perbedaan pemahaman.

Proses komunikasi dengan seseorang yang mengalami tunarungu sangat sulit. Hal tersebut disebabkan oleh beberapa faktor, yaitu masih sedikit orang normal yang mengerti bahasa isyarat, penggunaan masker selama pandemi sehingga sulit untuk membaca gerakan mulut. Sehingga diperlukan nya aplikasi real-time untuk menerjemahkan bahasa isyarat sehingga tidak terjadi kesalahpahaman terhadap komunikasi yang sedang berlangsung.

Untuk memenuhi fokus permasalahan terkait komunikasi dengan tunarungu, dipilih lah yayasan atau sekolah, pelajar atau mahasiswa, serta pengajar di segala usia. Berdasarkan latar belakang tersebut, penulis membuat aplikasi real-time yang dapat menunjang komunikasi teman tuli untuk melakukan segala aktivitasnya dengan baik.

Saat ini, banyak framework machine learning atau library untuk pengenalan hand gesture sedang dikembangkan untuk memudahkan siapa saja dalam membangun aplikasi berbasis AI (Artificial Intelligence). Salah satunya adalah MediaPipe. Framework MediaPipe disediakan oleh Google untuk menyelesaikan masalah machine learning seperti Face Recognition, Face Grid, Iris, Hands, Pose, Holistic, Hair Segmentation, Object Detection, Box Tracking, Instant Motion Tracking, Objections, KIFT. Framework MediaPipe membantu pengembang fokus pada algoritma aplikasi dan pengembangan model, mendukung lingkungan aplikasi melalui hasil yang dapat direproduksi di berbagai perangkat dan platform. Ini adalah bagian dari keuntungan menggunakan fitur framework MediaPipe [4].

Pada penelitian ini penulis berfokus pada pengembangan kode hand gesture recognition dengan framework pipeline. Yang pada percobaanya akan membuat aplikasi mengenali beberapa gesture, diantaranya open, close, ok, dan pointer.

II. TINJAUAN PUSTAKA

Hand gesture recognition, merupakan teknologi yang mampu membaca gerak tangan kemudian dirubah menjadi teks dan atau suara. Gesture recognition merupakan topik dalam computer science dan computer vision yang bertujuan agar komputer bisa memahami gerakan manusia yang umumnya

berasal dari tangan atau. [5] Berikut adalah beberapa dasar teori yang perlu diketahui dalam pengembangan aplikasi ini.

A. Object Detection

Object Detection atau deteksi objek merupakan bagian dari Computer Vision. Object Detection mengacu pada kemampuan komputer untuk mendeteksi sejumlah objek pada suatu gambar. Hal ini dapat dilakukan dengan cara mengambil image feature seperti garis, sudut, kontur dan warna dari sebuah gambar.[6] Deteksi objek merupakan bagian dari Object Recognition atau identifikasi objek. Sehingga dapat disimpulkan bahwa untuk deteksi objek pasti harus diidentifikasi terlebih dahulu objek tersebut. Sedangkan pada penelitian ini hanya dilakukan deteksi objek saja tanpa adanya identifikasi objek.

B. Deep Learning

Deep learning adalah sebuah artificial intelligence yang dapat meniru proses kerja otak manusia. Teknologi ini sangat efektif untuk mengolah data mentah dan menciptakan pola untuk keperluan pengambilan keputusan. Deep learning sendiri merupakan bagian dari machine learning yang memiliki jaringan tersendiri. Ia mampu mengenali pola dan informasi tanpa pengawasan dari data yang tidak terstruktur atau tidak berlabel. Beberapa algoritma Deep Learnig diantaranya CNN (Convolutional neural networks), LSTM (Long short term memory network), RNN (Reccurent neural network), SOM (Self organizing maps), dan lainnya.[7]

C. LSTM

Long short-term memory (LSTM) merupakan pengembangan dari arsitektur Recurrent Neural Network (RNN). LSTM memiliki keunggulan dalam mengingat dan menyimpan informasi masa lampau serta mampu mempelajari suatu data yang bersifat sekuensial. Terdapat struktur dasar pada LSTM yaitu input layer, hidden layer, dan output layer.

Terdapat dua fungsi aktivasi yang digunakan pada LSTM yaitu sigmoid dan tanh. Pada LSTM juga terdapat memory cell dan gerbang. Gerbang tersebut tersusun dari tiga gerbang yaitu forget gate, input gate, dan output gate.[8]

D. Mediapipe

MediaPipe adalah framework yang memungkinkan pengembang untuk membangun saluran ML multi-modal (video, audio, seri waktu apa pun). Sebagai kerangka node dan tepi atau landmark, mereka melacak titik-titik kunci di berbagai bagian tubuh. Semua titik koordinat dinormalisasi tiga dimensi. MediaPipe Holistic menggunakan model landmark pose, wajah dan tangan masing-masing untuk menghasilkan total 543 landmark (33 landmark pose, 468 landmark wajah, dan 21 landmark tangan per tangan).[9]

Framework mediapipe sudah bisa mendeteksi bentuk dan gerakan tangan. Ada beberapa model yang dapat membantu dalam hal ini diantaranya adalah :

1. Palm Detection Model

Framework MediaPipe membangun detektor telapak tangan versi awal yang disebut BlazePalm. Mendeteksi tangan adalah tugas yang kompleks. Langkah pertama adalah melatih telapak tangan alih-alih detektor tangan, kemudian menggunakan algoritma penekanan non-maksimum yang dimodelkan pada using square bounding boxes untuk menghindari rasio aspek yang berbeda dan mengurangi jumlah anchors dengan faktor 3 - 5. Selanjutnya, Extraction Encoder Decoder adalah fitur yang digunakan untuk meningkatkan pengenalan konteks adegan untuk objek kecil, yang pada akhirnya mengurangi kehilangan fokus selama training dengan mendukung sejumlah besar anchors yang berasal dari high scale variance. [10] [11]

2. Hand Landmark Model

Dengan ini bisa mendapat lokalisasi titik kunci yang akurat dari 21 titik kunci menggunakan koordinat buku jari tangan 3D yang dilakukan di dalam daerah tangan yang terdeteksi melalui regresi yang akan menghasilkan prediksi koordinat secara langsung yang merupakan model landmark tangan di MediaPipe. [10][11]

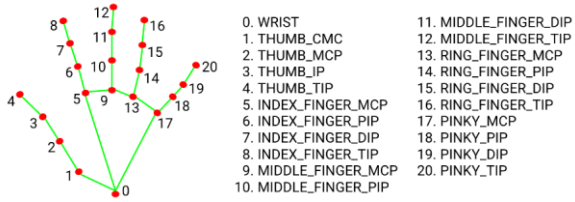

Fig. 1. Hand Landmark dalam MediaPipe

Setiap ruas jari memiliki koordinat yang terdiri dari x, y, z, di mana x dan y ar dinormalisasi menjadi [0.0, 1.0] dengan lebar dan tinggi gambar, dan z adalah Mewakili kedalaman landmark . Kedalaman landmark yang ditemukan di pergelangan tangan menjadi titik utama (wrist dengan index 0). Semakin dekat landmark ke kamera, semakin kecil nilainya.

E. Tensorflow

Tensorflow merupakan kerangka dasar yang digunakan dalam proses machine learning dan sebagai library khusus untuk machine learning yang dikembangkan oleh google.[12]

F. Open CV

OpenCV (Open Source Computer Vision Library) adalah sebuah library perangkat lunak yang ditujukan untuk pengolahan citra dinamis secara real-time. OpenCV (Open Source Computer Vision Library), adalah sebuah library open source yang dikembangkan oleh intel yang fokus untuk menyederhanakan programing terkait citra digital. Di dalam OpenCV sudah mempunyai banyak fitur, antara lain : pengenalan wajah, pelacakan wajah, deteksi wajah, Kalman filtering, dan berbagai jenis metode AI (Artificial Intellegence). Dan menyediakan berbagai algoritma sederhana terkait Computer Vision untuk low level API.[13]

III. Eksperimen & Analisis

A. Spesifikasi Komputer

Spesifikasi komputer yang saya gunakan adalah sebagai berikut :

- Processor : Intel® Core™ i5-9300H CPU @ 2.40 GHz (8 CPUs), ~2.4GHz

- Memory : 8192 MB RAM DDR 4
- System Type : 64-bit
- Internal : SSD Sata 512 GB
- VGA : NVIDIA GeForce GTX 1050

B. Pengumpulan Data

Proses pengumpulan data yang dilakukan yaitu dengan cara mengambil data kordinat dari 21 landmark yang ada pada figure 1. Data yang diambil merupakan data numerik yang telah di normalisasi dari jarak antara tiap landamark ke landmark utama (index 0 = wrist). Dalam percobaan ini terdapat 4 tipe macam data yang artinya aplikasi ini akan bisa mendeteksi 4 macam gesture, yaitu Open, Close, Pointer, Ok.

C. Pengolahan Data

Pengolahan data pada penelitian ini melalui beberapa proses yaitu:

1. Mendeteksi landmarks tangan. Kamera webcam akan merekam tangan menggunakan open cv dan mediapipe untuk mendapatkan model landmark dan menempatkan 21 keypoints pada tangan.
2. Ekstraksi keypoints Keypoints yang telah diperoleh sebelumnya akan diekstraksi dengan menggabungkan nilai keypoints tersebut ke dalam array numpy.
3. Pembuatan dan input data ke file csv. file ini berguna sebagai tempat penyimpanan data hand gesture yang telah direkam oleh webcam dan telah dideteksi menggunakan keypoints mediapipe.
4. Pengumpulan keypoints dilakukan dengan menggunakan webcam dan merekam keypoints dari setiap hand gesture yang telah didefine, yaitu Open, Close, Pointer, Ok. Setiap kosakata memiliki jumlah data yang berbeda - beda. yang jelas semua data tiap label terdapat lebih dari 300. Data bisa ditambah dengan cara merekam keypoints yang ada fiturnya pada kode yang diuji. Keypoint yang telah direkam akan tersimpan otomatis kedalam file yang sudah dibuat sebelumnya.

D. pre-pemrosesan Data

Pada tahap ini setiap kelas hand gesture akan dibuatkan ke dalam label array, kemudian data akan dibagi sebanyak 75% sebagai data training dan sebanyak 25% sebagai data testing.

E. Pemodelan dan Pelatihan

Pada figure 2 kita bisa lihat bagaimana pembangunan model yang akan digunakan pada aplikasi ini. Setelah melewati layer input data akan melewati serangkaian layer sebelum keluar ke layer output, detailnya bisa dilihat pada figure 2.

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dropout (Dropout)            (None, 42)                0
_________________________________________________________________
dense (Dense)                (None, 20)                860
_________________________________________________________________
dropout_1 (Dropout)          (None, 20)                0
_________________________________________________________________
dense_1 (Dense)              (None, 10)                210
_________________________________________________________________
dense_2 (Dense)              (None, 4)                 44
=================================================================
```

Fig. 2. Hand Landmark dalam MediaPipe

Proses training yang dilakukan dengan epochs sebanyak 1000, batch size sebanyak 128. Kemudian saat testing juga menerapkan batch size sebanyak 128.

F. Evaluasi

Model yang telah dilatih akan di evaluasi menggunakan confusion matrix untuk mengukur performa model dalam kemampuan klasifikasi dengan melihat parameter pengukuran performanya seperti akurasi, recall, presisi, dan skor f1 dengan menggunakan fungsi-fungsi pengukuran yang sudah disediakan oleh library sklearn.

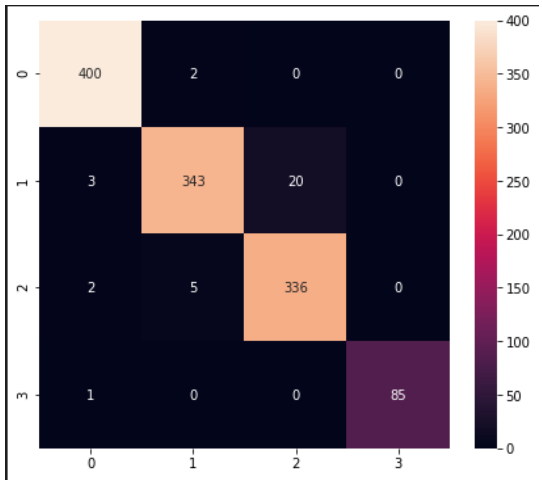

Fig. 3. Hasil Confusion Matrix

TABLE I. CLASSIFICATION REPORT

class	Presicion	Recall	F1-score
0 (open)	0.99	1.0	0.99
1 (close)	0.98	0.94	0.96
2 (pointer)	0.94	0.98	0.96
3 (ok)	1.0	0.99	0.99
Accuracy			0.97

G. Analisis

Dari tabel 1 bisa dilihat hasil dari training yang dilakukan. Menurut tabel tersebut hasil dari training bisa dibilang tinggi dan cukup memuaskan. Alasan kenapa mendapat hasil yang memuaskan yaitu karena dataset yang di gunakan untuk mentraining pada percobaan ini lumayan banyak, dan klasifikasi yang dilakukan baru dibagi menjadi 4 kategori.

IV. PENGUJIAN

Karena hasilnya dari training bisa dibilang sudah bagus, maka langkah selanjutnya adalah pengujian dengan Open CV secara real-time. Pengujian yang akan dilakukan adalah pengujian fungsional. Pengujian Fungsional digunakan untuk melihat sistem sudah berjalan atau belum. Skenario pengujian dilakukan dengan menguji fitur-fitur dari aplikasi yang dirancang, berhasil atau tidaknya aplikasi tersebut sudah bisa mengklasifikasikan hand gesture dengan benar.

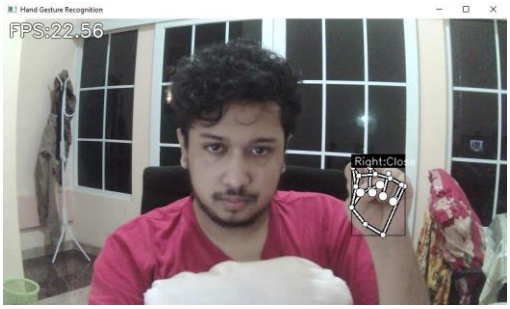

Fig. 4. Pengujian Hand Gesture Close

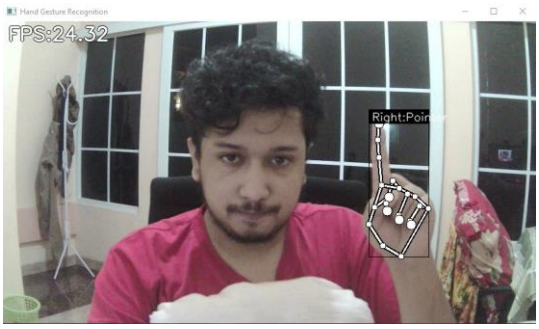

Fig. 5. Pengujian Hand Gesture Pointer

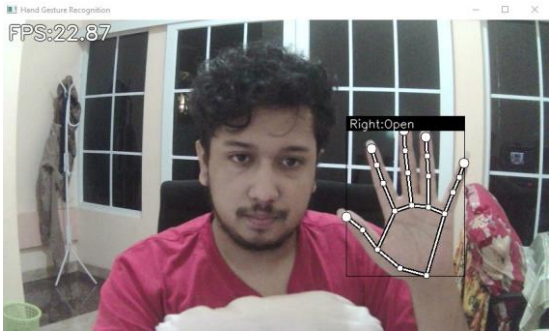

Fig. 6. Pengujian Hand Gesture Open

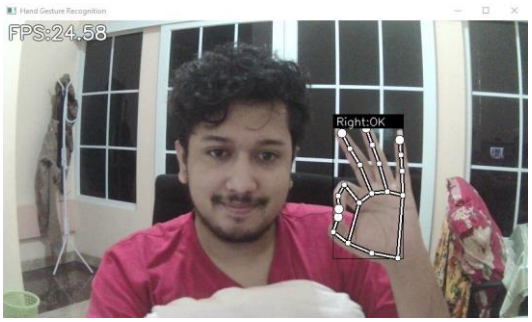

Fig. 7. Pengujian Hand Gesture OK

V. KESIMPULAN

Sistem Hand gesture recognition telah menjadi peran penting dalam membangun interaksi manusia-mesin yang efisien. Implementasi menggunakan Hand gesture recognition menjanjikan cakupan luas dalam industri teknologi. MediaPipe sebagai salah satu framework berbasis machine learning berperan efektif dalam mengembangkan aplikasi ini, dengan hasil menunjukkan kinerja akurasi sebesar 97%. penulis ingin memperluas sistem kami lebih jauh untuk mengembangkan kolaborasi dengan perangkat lain dan bagian tubuh manusia lainnya dan bereksperimen dengan sistem Hand gesture recognition statis dan dinamis.

Untuk pengembangan selanjutnya bisa mengarah pada pengembangan mobile application pada aplikasi ini. Kemudia n karena penguji hanya mencoba 4 macam gesture, akan ada kemungkinan akurasi akan turun jika kita mencoba ratusan gesture. Untuk mencegah hal tersebut bisa menerapkan beberapa hal untuk meninkatkan akurasi, seperti memperbanyak dataset, hyperparameter tunning, model building yang membuat peforma meningkat dan efisien.

REFERENCES

[1] Mulyana, D. (2001). Pengantar Ilmu Komunikasi. Bandung, Remaja.

[2] Arti Kata "tunarungu" Menurut Kamus Besar Bahasa Indonesia | KBBI.co.id .

[3] Arti kata komunikasi - Kamus Besar Bahasa Indonesia (KBBI) Online.

[4] Lugaresi C, Tang J, Nash H, McClanahan C, et al. MediaPipe: A Framework for Building Perception Pipelines. Google Research. 2019. https://arxiv.org/abs/2006.10214.

[5] Raheja, J. L., Singhal, A., & Chaudhary, A. (2015). Android based portable hand sign recognition system. arXiv preprint arXiv:1503.03614.

[6] T. Wenzel, T. W. Chou, S. Brueggert, and J. Denzler, “From corners to rectangles-Directional road sign detection using learned corner representations,” in IEEE Intelligent Vehicles Symposium, Proceedings, 2017.

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[8] Long short term memory. https://mti.binus.ac.id/2019/12/02/long-short-term-memory-lstm/

[9] Mediapipe, Google. https://google.github.io/mediapipe/

[10] M.Panwar, Hand Gesture Recognition Based on Shape Parameters, In International Conferences: Computing Communication and Application (ICCCA), 2012

[11] Marco Maisto, An Accurate Algorithm for Identification of Fingertips Using an RGB-D Camera, IEEE Journal on Emerging and Selected Topics in Circuits and System, 2013. pp. 272-283.

[12] https://dqlab.id/belajar-data-science-pahami-tensflow

[13] Introduction to open cv. https://binus.ac.id/malang/2017/10/introduction-to-open-cv/

[14] Hand gesture recognition using mediapipe, Kazuhito. https://github.com/Kazuhito00/hand-gesture-recognition-using-mediapipe