Hand Pose Recognition using Deep Learning

Dr. Eko Mulyanto Yuniarno

November 2024

Contents

Chapter 1

Introduction

Hand pose are one of the most common communication methods in daily human life. Interaction modalities of user interfaces play a dominant role in the relationship between people and computer technology. The way users interact with interfaces has undergone a significant transformation, with most of the population now using touch devices. In today's technological advancements, human-computer interfaces (HCI) are highly appealing, mainly because they aim to enhance human lifestyles. Alongside these developments, technologies that facilitate interaction with these devices are also required. Initially, interactions were conducted using a mouse and keyboard with computers. However, with the rise of ubiquitous computing, gestures have become more prevalent—for example, smartphone interactions often involve hand gestures. The human body provides a wide range of poses that can be used as computer input. Images captured by cameras exhibit a vast amount of variation. This occurs because images are spatial data where each pixel represents a color at a specific coordinate. For pose classification, many images are needed for training to account for variations in scale, position, and hand orientation within the images. The desired outcome is a feature invariant to scale, position, and orientation, ensuring accurate and robust classification.

1.1 Importance of Hand Pose Recognition

Hand pose recognition is pivotal in advancing human-computer interaction (HCI), enabling intuitive communication between humans and machines. It bridges the gap between natural human gestures and machine understanding, opening new avenues in virtual reality, robotics, and accessibility technologies. For instance, hand gesture recognition is essential for sign language translation, empowering people with hearing impairments to communicate

seamlessly with others. Moreover, the ability to accurately recognize hand poses enhances precision in applications such as gaming, remote robotic control, and augmented reality, where fine motor movements dictate user experience. As technology continues to evolve, the importance of hand pose recognition grows, particularly in creating inclusive, adaptive, and user-friendly systems.

1.2 Challenges in Hand Pose Recognition

Despite its significance, hand pose recognition presents numerous challenges, primarily due to the variability and complexity of human hands. Hands have intricate structures with multiple degrees of freedom, leading to a vast range of poses that can be difficult to capture and interpret accurately. Variations in lighting, occlusions caused by overlapping fingers, and differences in hand shapes across individuals add further complexity. Real-time processing demands also pose a challenge, requiring high computational efficiency without sacrificing accuracy. Additionally, datasets used for training hand pose recognition models often need more diversity, limiting their generalizability in real-world scenarios. Addressing these challenges requires robust algorithms capable of handling variations and noise while maintaining computational efficiency.

1.3 Hand Pose Recognition Implementation in Medical

Hand pose recognition is revolutionizing patient care and therapy methods in the medical field. Surgeons can utilize gesture-based systems to interact with medical imaging during procedures, eliminating the need for physical contact with equipment and ensuring sterility in the operating room. In rehabilitation, hand pose recognition monitors and guides patients recovering from injuries or surgeries involving fine motor skills. For example, it enables therapists to track real-time progress and customize exercises based on a patient's movements. Moreover, hand pose recognition aids in developing assistive devices for individuals with motor impairments, enhancing their ability to perform daily tasks independently. This technology's application in healthcare improves the precision of treatments and fosters patient engagement and empowerment.

Chapter 2

Fundamentals of Hand Pose Recogntion with MediaPipe

2.1 Understanding Pose Recognition

Pose recognition is analyzing and interpreting objects' physical positions and movements, particularly the human body or hands, to identify specific gestures or actions. In the context of hand pose recognition, the objective is to track and understand the intricate movements and orientations of the hand in a given space. This involves detecting the hand's position, identifying key landmarks, and classifying the overall hand configuration. Pose recognition systems combine advanced computer vision techniques with machine learning models to make predictions, offering applications in sign language interpretation, virtual reality, and human-computer interaction.

2.2 MediaPipe

MediaPipe is a cross-platform framework by Google that offers efficient pipelines for machine learning and computer vision tasks, such as pose estimation, hand tracking, and facial landmark detection. It supports mobile devices, web, and desktop environments with pre-trained models.

2.3 Real-Time Hand Tracking

Real-time hand tracking is the ability to detect and monitor hand movements instantaneously. This process relies on high-speed algorithms that process visual data from cameras or sensors to pinpoint the location and orientation of

a hand in every frame of a video stream. A significant focus of real-time hand tracking is minimizing latency to ensure the system's responsiveness. Techniques like region-of-interest optimization, GPU acceleration, and lightweight neural networks contribute to achieving this. Real-time tracking is critical for applications such as augmented reality (AR), gaming, and touchless interfaces where seamless interaction is essential.

2.4 Landmark Detection

Landmark detection is a core element of hand pose recognition. It involves identifying key points or nodes on a hand, such as joints, knuckles, or fingertips, collectively defining its structure. These landmarks are typically represented in 2D or 3D coordinates, forming the foundation for analyzing hand gestures and poses. Advanced models like Mediapipe Hand and OpenPose leverage machine learning to perform this task efficiently. Landmark detection is critical for understanding hand dynamics, as it provides the data needed for gesture classification, tracking movements, and reconstructing hand positions in 3D space. The hand landmark can be seen in Figure 4.1.

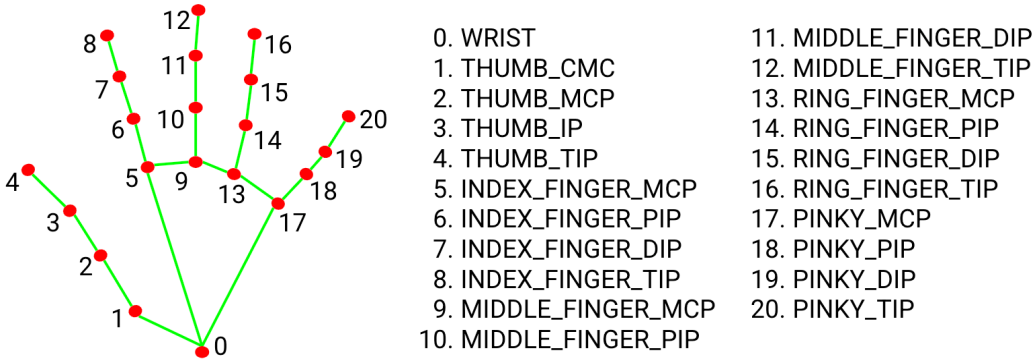

Figure 2.1: Example of Hand Landmark obtained from Mediapipe

2.5 Model architecture

The model architecture for hand pose recognition typically combines convolutional neural networks (CNNs) for feature extraction with regression or classification layers for predicting hand poses. State-of-the-art models often utilize encoder-decoder frameworks to process input images and generate

heatmaps or landmark coordinates. Mediapipe, for instance, uses a pipeline that first detects the hand region and then applies a specialized model to predict precise landmarks. Lightweight and efficient architectures are crucial to ensure real-time performance without compromising accuracy. These models are trained on large datasets to generalize across different hand shapes, orientations, and lighting conditions.

2.6 Pose Estimation in 3D

Pose estimation in 3D extends the capabilities of 2D models by incorporating depth information to reconstruct hand poses in a three-dimensional space. This process often integrates data from depth sensors or multi-view camera setups to triangulate landmark positions. 3D pose estimation provides more accurate representations of hand gestures, especially in scenarios involving rotations or occlusions. It is beneficial in applications like robotics, where precise hand positioning is required, or virtual reality, where realistic interactions with virtual objects depend on spatial accuracy. Each landmark consists of the following:

1. x and y: Landmark coordinates normalized to $[0.0, 1.0]$ by the image width and height respectively.

2. z: Represents the landmark depth with the depth at the midpoint of hips being the origin, and the smaller the value the closer the landmark is to the camera. The magnitude of z uses roughly the same scale as x.

3. visibility: A value in $[0.0, 1.0]$ indicating the likelihood of the landmark being visible (present and not occluded) in the image.

2.7 Multi-Hand Support

Multi-hand support is the capability of a hand pose recognition system to track and analyze multiple hands within the same frame simultaneously. This requires the model to distinguish between hands, assign unique identifiers, and accurately detect the landmarks for each hand. Challenges in multi-hand support include managing occlusions (where one hand partially covers the other) and handling varying hand orientations. Robust systems, like Mediapipe, employ efficient region segmentation and tracking algorithms to maintain consistent performance in multi-hand scenarios. Multi-hand support is essential for collaborative applications like multi-user gaming or shared virtual workspaces.

Chapter 3

Hand Pose Recogntion using Deep Learning

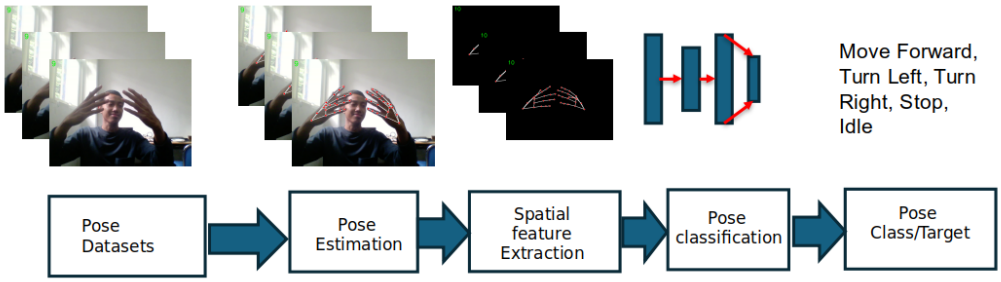

Figure 3.1: Hand Recognition Pipeline

3.1 Dataset Acquisition

This code provides a systematic way to acquire a dataset for hand pose recognition using Mediapipe and OpenCV. It captures hand landmarks and saves the corresponding frames from a webcam feed, enabling the creation of a labeled dataset for tasks like training gesture recognition models.

The process begins with the initialization and setup of necessary components. Mediapipe's Hands model is used for detecting and tracking hand landmarks, while OpenCV handles video capture and frame display. Command-line arguments allow customization of parameters such as the dataset path, label name, frame saving rate, maximum frames to save, and a delay before data collection starts. The script ensures that a directory for the dataset

is created, organizing the captured data based on the provided label name. The Figure 3.2 shows the sample of data acquisition.

Figure 3.2: Sample of Data Acquisition Process

Hand detection and landmark extraction are key steps in the process. The Mediapipe Hands model detects 21 landmarks on each hand, including knuckles and fingertips, and supports tracking up to two hands simultaneously. For every frame captured, the script converts it to RGB format and processes it using Mediapipe. Landmark coordinates are normalized relative to specific reference points, such as the wrist and a finger joint, making the data invariant to variations in hand size and position. The normalized landmarks for both hands are then concatenated into a single array, representing the frame's hand pose data.

3.2 Pose Estimation

The 'create_dataset' function is a crucial part of the pose estimation process. It facilitates the acquisition of hand pose data by capturing frames from a webcam feed, extracting hand landmarks using Mediapipe's 'Hands' solution, and organizing the data into labeled directories. This function serves as the foundation for building datasets tailored to pose estimation tasks, enabling further applications like gesture recognition and motion analysis.

Pose estimation involves detecting and interpreting key hand landmarks in real-time. The function initializes Mediapipe's 'Hands' module, which is capable of identifying 21 specific landmarks on each hand, such as fingertips, joints, and the wrist. These landmarks represent the structural configuration of the hand and are essential for understanding its pose. By converting the captured video frames to RGB format and processing them through Mediapipe, the function extracts these landmarks with high precision.

To normalize the landmark data, the function computes positions relative to key reference points—namely, the wrist and a finger joint. This normalization step ensures that the data remains invariant to hand size and camera

perspective, enhancing its generalizability for machine learning models. The landmark coordinates for both hands are flattened into a single array, representing the spatial configuration of the hands in each frame.

Beyond pose estimation, the 'create_dataset' function also incorporates real-time feedback and organization. Frames are captured and saved at user-defined intervals, allowing control over the frame rate and total number of frames collected. Each frame is stored in a labeled directory, organized by the name of the gesture or pose being recorded. The overlay of the detected landmarks on the video feed, along with a live counter of saved frames, enhances the usability of the function by providing visual confirmation and progress tracking.

This function exemplifies the integration of pose estimation techniques with practical data collection workflows. It not only extracts meaningful pose information but also structures the data in a format ready for subsequent analysis and model training. This makes it a powerful tool for researchers and developers working on applications in gesture recognition, human-computer interaction, and beyond.

3.2.1 Import required Python libraries

```
import cv2
import mediapipe as mp
import numpy as np
from datetime import datetime
import os
import time
import argparse
```

3.2.2 MediaPipe Initialization

```
mp_hands = mp.solutions.hands
hands = mp_hands.Hands(static_image_mode=False,
max_num_hands=2,
min_detection_confidence=0.5,
min_tracking_confidence=0.5)
```

This part initializes the Mediapipe Hands solution, which is responsible for detecting hand landmarks in video frames. The configuration allows real-time processing of up to two hands per frame, with parameters controlling the confidence thresholds for detection and tracking. This setup ensures accurate and stable landmark recognition during data collection.

3.2.3 Webcam Access and Directory Setup

```
cap = cv2.VideoCapture(0)
save_dir = os.path.join(direktori_path, nama_label)
if not os.path.exists(save_dir):
    os.makedirs(save_dir)
```

The webcam is opened using OpenCV's VideoCapture, which serves as the data source for real-time frame capture. The save_dir is created dynamically based on the dataset path and label name, ensuring an organized directory structure for storing the collected data.

3.2.4 Landmark Extraction

```
def ekstraksi_fitur(frame):
frame_rgb = cv2.cvtColor(frame, cv2.COLOR_BGR2RGB)
results = hands.process(frame_rgb)
height, width, _ = frame.shape
...
# Normalization and concatenation of landmarks
return np.concatenate((left_hand_landmarks.flatten(),
   right_hand_landmarks.flatten()))
```

This function extracts and processes hand landmarks from each frame:

1. Conversion to RGB: Mediapipe requires frames in RGB format for processing.

2. Landmark Processing: If hand landmarks are detected, their coordinates are normalized relative to reference points (e.g., the wrist and a finger joint) to ensure invariance to hand size and position.

3. Concatenation: The landmarks for both hands are flattened and concatenated into a single array, representing the full pose information for the frame.

3.3 Geometric Feature Extraction

3.3.1 Function Definition and Initialization

```
def create_dataset(direktori_path, nama_label, frameratesave
   =2, maxframe=100, start_delay=5):

    # Initialize Mediapipe Hands and Drawing
```

```
    mp_hands = mp.solutions.hands
    mp_drawing = mp.solutions.drawing_utils
    hands = mp_hands.Hands(static_image_mode=False,
    max_num_hands=2,
    min_detection_confidence=0.5,
    min_tracking_confidence=0.5)
```

The create_dataset function begins with its definition, which includes several customizable parameters: the dataset directory (direktori_path), the label for the dataset (nama_label), the frame saving rate (frameratesave), the maximum number of frames to save (maxframe), and the delay before starting the capture (start_delay). Inside the function, the Mediapipe Hands model is initialized. This model is set up to dynamically process video streams (static_image_mode=False) and detect up to two hands simultaneously. Detection and tracking confidence thresholds are set to 0.5, balancing accuracy and performance. The DrawingUtils module is also initialized for later use in visualizing detected landmarks.

3.3.2 Webcam Initialization and Directory Setup

```
        # Open webcam
    cap = cv2.VideoCapture(0)
    if not cap.isOpened():
         print("Failed to open webcam.")
         return

    # Create directory path
    save_dir = os.path.join(direktori_path, nama_label)
    if not os.path.exists(save_dir):
         os.makedirs(save_dir)
         print(f"Directory created: {save_dir}")
```

Next, the function initializes the webcam using OpenCV's VideoCapture. If the webcam cannot be opened, an error message is printed, and the function exits gracefully. The function then prepares the directory for storing the dataset. If the directory does not already exist, it is created using os.makedirs. The dataset is organized by the label provided (nama_label), ensuring clear structure and traceability for the saved frames.

3.3.3 Countdown Timer and Frame Processing

Before data collection begins, the function sets up a countdown timer `start_delay` using `time.time()`. The `ekstraksi_fitur` function processes each frame. It converts the frame to RGB (as required by Mediapipe) and detects hand

landmarks using the `hands.process` method. If landmarks are detected, they are normalized relative to specific reference points (e.g., the wrist and a finger joint) to account for hand size and positioning. The normalized coordinates for the left and right hands are stored in separate arrays and concatenated for a unified representation.

```
1
2  time_before = datetime.now()
3  start_time = time.time()
4  counter = 0
5
6  def ekstraksi_fitur(frame):
7      # Convert to RGB
8      frame_rgb = cv2.cvtColor(frame, cv2.COLOR_BGR2RGB)
9
10     # Process frame with Mediapipe
11     results = hands.process(frame_rgb)
12
13     # Frame dimensions
14     height, width, _ = frame.shape
15
16     # Initialize numpy arrays for left and right hands
17     left_hand_landmarks = np.zeros((21, 2))
18     right_hand_landmarks = np.zeros((21, 2))
19
20     # Helper function to process landmarks
21     def process_landmarks(hand_landmarks, width, height):
22         landmarks = [(lm.x * width, lm.y * height) for lm in
               hand_landmarks.landmark]
23         landmark_0 = np.array(landmarks[0])
24         landmark_5 = np.array(landmarks[5])
25         normalized_landmarks = [
26             ((x - landmark_0[0]) / (landmark_5[0] -
                   landmark_0[0] + 1e-6),
27             (y - landmark_0[1]) / (landmark_5[1] - landmark_0
                   [1] + 1e-6))
28             for x, y in landmarks
29         ]
30     return np.array(normalized_landmarks)
31
32     # If hands are detected
33     if results.multi_hand_landmarks and results.
           multi_handedness:
34         for hand_landmarks, hand_handedness in zip(results.
               multi_hand_landmarks, results.multi_handedness):
35         # Identify hand as left or right
36         handedness = hand_handedness.classification[0].label
```

```
            processed_landmarks = process_landmarks(
                hand_landmarks, width, height)
            if handedness == 'Left':
                left_hand_landmarks = processed_landmarks
            elif handedness == 'Right':
                right_hand_landmarks = processed_landmarks

            # Draw landmarks on the frame
            mp_drawing.draw_landmarks(frame, hand_landmarks,
                mp_hands.HAND_CONNECTIONS)

            # Concatenate left and right hand landmarks
        concatenated_landmarks = np.concatenate((
            left_hand_landmarks.flatten(), right_hand_landmarks.
            flatten()))
        return concatenated_landmarks
```

3.3.4 Frame Capture and Saving

```
while cap.isOpened():
    ret, frame = cap.read()
    if not ret:
        print("Failed to read frame from webcam")
        break

    # Countdown before saving starts
    elapsed_time = time.time() - start_time
    if elapsed_time < start_delay:
        countdown_text = f"Starting in {int(start_delay -
            elapsed_time)} seconds"
        cv2.putText(frame, countdown_text, (10, 50), cv2.
            FONT_HERSHEY_SIMPLEX, 1, (0, 0, 255), 2)
        cv2.imshow('Hand Landmark Detection', frame)
        if cv2.waitKey(1) & 0xFF == ord('q'):
            break
        continue

    # Check framerate and save frame
    time_now = datetime.now()
    if (time_now - time_before).total_seconds() > 1 /
        frameratesave:
        # Generate file name
        file_name = datetime.now().strftime("%Y%m%d%H%M%S%f")
            [:-3] + ".jpg"
        file_path = os.path.join(save_dir, file_name)

        # Save frame
```

```
25          cv2.imwrite(file_path, frame)
26          counter += 1
27
28          # Update the previous time
29          time_before = time_now
30
31      # Display counter on frame
32      cv2.putText(frame, f"Files saved: {counter}", (10, 50),
            cv2.FONT_HERSHEY_SIMPLEX, 1, (0, 255, 0), 2)
33
34      # Process landmarks and draw them on the frame
35      concatenated_landmarks = ekstraksi_fitur(frame)
36      print("Concatenated Landmarks:", concatenated_landmarks)
37
38      # Display the frame with landmarks
39      cv2.imshow('Hand Landmark Detection', frame)
40
41      # Stop saving after maxframe is reached
42      if counter >= maxframe:
43          print(f"Reached maximum of {maxframe} frames. Exiting
                ...")
44          break
45
46      # Handle key press
47      key = cv2.waitKey(1) & 0xFF
48      if key == ord('q'):  # Exit on 'q'
49          break
```

The webcam feed is processed frame by frame. During the countdown period, a message informs the user of the time remaining before data collection begins. Once the countdown ends, frames are saved at intervals defined by the `frameratesave` parameter. Each frame is assigned a timestamp-based filename and stored in the appropriate directory. The landmarks are processed in real-time, printed to the console, and visualized on the webcam feed using Mediapipe's drawing utilities.

3.3.5 Cleanup

```
1       # Release resources
2       cap.release()
3       cv2.destroyAllWindows()
4       hands.close()
```

After reaching the specified frame limit or upon user interruption, the function releases the webcam and closes all OpenCV windows. The Mediapipe Hands object is also closed, ensuring that all resources are properly deallo-

cated. This function effectively combines pose estimation, real-time feedback, and dataset organization, making it a comprehensive tool for collecting hand pose data.

3.4 Model Architecture

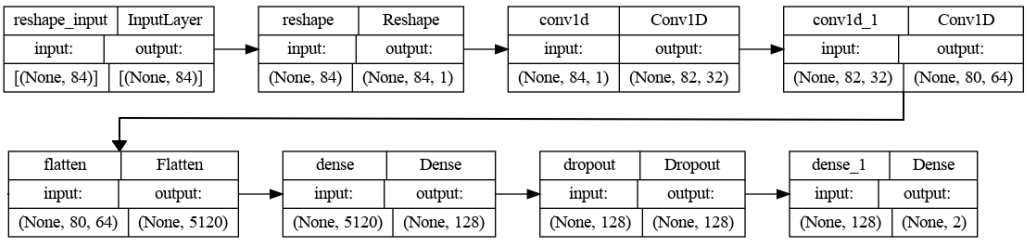

Figure 3.3: Model Architecture

This model represents a simple feed-forward neural network designed for tasks such as classification, particularly for hand pose recognition. The architecture utilizes Convolutional Neural Network (CNN) model implemented in TensorFlow, specifically designed for tasks involving 1D data, such as time-series analysis, signal processing, or structured data (like hand pose landmarks).

```
1    Reshape((input_size, 1), input_shape=(input_size,))
```

This layer reshapes the 1D input vector of size `(input_size,)` into a 2D array with dimensions `(input_size, 1)`. This transformation is necessary because Conv1D layers require a 2D input format where one dimension represents the sequence length `(input_size)`, and the other represents the number of channels (here, a single channel).

The reshaped input is then passed to the first Conv1D layer:

```
1    Conv1D(filters=32, kernel_size=3, activation='relu')
```

This layer applies 32 convolutional filters with a kernel size of 3 to extract local patterns from the data. The `relu` activation introduces non-linearity, enabling the model to learn complex features. The next layer is another Conv1D layer:

```
1    Conv1D(filters=64, kernel_size=3, activation='relu')
```

This layer increases the number of filters to 64, allowing it to capture more complex patterns and higher-level features from the output of the previous

layer. Both convolutional layers reduce the dimensions of the input while retaining its essential features.

After the convolutional layers, a Flatten layer:

```
Flatten()
```

is used to convert the multi-dimensional output from the Conv1D layers into a 1D vector. For example, if the output from the second Conv1D layer is of shape (80, 64), the Flatten layer transforms it into a vector of size (80 * 64 = 5120). This step prepares the data for the subsequent fully connected layers.

The flattened output is fed into a Dense layer:

```
Dense(128, activation='relu')
```

which consists of 128 neurons. Each neuron applies a linear transformation followed by a `relu` activation function to learn combinations of features extracted by the convolutional layers. This dense layer reduces the feature space while focusing on the most significant patterns.

To prevent overfitting, the model includes a Dropout layer:

```
Dropout(0.5)
```

which randomly deactivates 50% of the neurons during training. This regularization technique ensures that the model does not rely too heavily on specific neurons, improving its ability to generalize to unseen data.

Finally, the model ends with an output Dense layer:

```
Dense(num_classes, activation='softmax')
```

This layer has a number of neurons equal to the number of target classes (`num_classes`). The `softmax` activation function converts the raw outputs into probabilities for each class, ensuring that the sum of all probabilities equals 1. This makes it suitable for multi-class classification tasks.

Overall, the model starts with reshaping the input, extracts meaningful features using convolutional layers, flattens the extracted features, and uses dense layers for classification, with dropout for regularization. The final softmax layer outputs class probabilities, making it effective for tasks like hand pose classification or time-series analysis.

3.5 Pose Classification

3.6 Pose Inference

Chapter 4

Fundamentals of Hand Pose Recogntion