Gesture-to-Image Generation

DLP Project Report

M.Mahad Munir (21k-3388)
Asad Noor Khan (21k-4678)
FAST-NUCES, Karachi

May 6, 2025

1. Objective

The objective of this project is to build an AI-driven system capable of converting real-time hand gestures into visually meaningful images. This is achieved by combining real-time gesture recognition with cutting-edge generative AI, enabling novel human-computer interaction using natural hand movements.

2. Problem Statement

Traditional methods of controlling digital interfaces rely on touch or voice input, which can be limiting in certain environments. There is a need for an intuitive and contactless interaction mechanism that maps hand gestures to semantically meaningful content. The problem is to design a pipeline that recognizes user hand gestures through a webcam, classifies them in real-time, and generates high-resolution images using AI based on predefined textual mappings.

3. Methodology

3.1 Hand Gesture Detection using MediaPipe Hands

MediaPipe Hands is a lightweight and efficient real-time hand tracking pipeline developed by Google. It detects 21 3D hand landmarks using a multi-stage machine learning architecture:

- **Palm Detection:** A single-shot detector model identifies palm regions instead of full hands for improved robustness and speed.
- **Hand Landmark Model:** A regression model takes the palm ROI and predicts 21 keypoint landmarks in 3D space.
- **Output:** 21 keypoints per hand with coordinates (x, y, z) and handedness classification.

Pipeline Architecture:

```
Input Image → Palm Detector (CNN) → ROI Cropper → Hand Landmark Model →
3D Keypoints
```

3.2 Gesture Classification

The extracted 3D landmarks are normalized and converted into feature vectors. A shallow Multi-Layer Perceptron (MLP) classifier trained on labeled gesture data then predicts the performed gesture.

- **Input:** 63-dimensional vector (21 landmarks × 3 coordinates)
- **Model:** Two-layer MLP with ReLU activations and Softmax output
- **Training:** Custom dataset of common gestures like thumbs-up, OK sign, victory, fist, etc.

3.3 Gesture-to-Text Mapping

Each recognized gesture is mapped to a semantically rich text prompt. For instance:

Gesture	Mapped Prompt
Thumbs-up	"a peaceful forest landscape at sunrise, digital art"
Victory	"a futuristic cyberpunk city skyline at night, neon lights"
Fist	"a mighty dragon breathing fire on a mountain peak, fantasy art"
Open Palm	"a surreal cosmic landscape with colorful nebulae and planets"
Pointing Up	"a majestic castle floating in the clouds, dreamlike atmosphere"

3.4 Image Generation using Stable Diffusion

Stable Diffusion is a latent text-to-image diffusion model that generates images from natural language descriptions.

Architecture Components:

- **VAE (Variational Autoencoder):** Encodes images into low-dimensional latent space and reconstructs them.
- **U-Net Denoiser:** Learns to reverse noise at each timestep conditioned on the prompt.
- **CLIP Text Encoder:** Converts textual prompt into embeddings for the U-Net.

Generation Pipeline:

```
Text Prompt → CLIP Encoder → U-Net (Latent Diffusion) → VAE Decoder →
Image
```

Training Overview

While the project uses pre-trained weights, the training regime was studied extensively:

- Trained on LAION-5B (5B image-text pairs)
- Uses denoising score matching loss (MSE)

- Requires multi-GPU (e.g., A100) for training from scratch

3.5 System Integration

- **Frontend:** React.js interface with webcam feed and image display.
- **Backend:** FastAPI server for handling gesture classification and image generation requests.
- **API Routes:**
  - `POST /gesture`: returns classified gesture
  - `POST /generate-image`: returns image for corresponding prompt

This multi-stage pipeline brings together real-time CV, classification, text abstraction, and generative AI into a unified, interactive system.

4. Results

The system successfully maps user hand gestures to high-quality AI-generated images in real-time. The following screenshots demonstrate the result of different gestures:

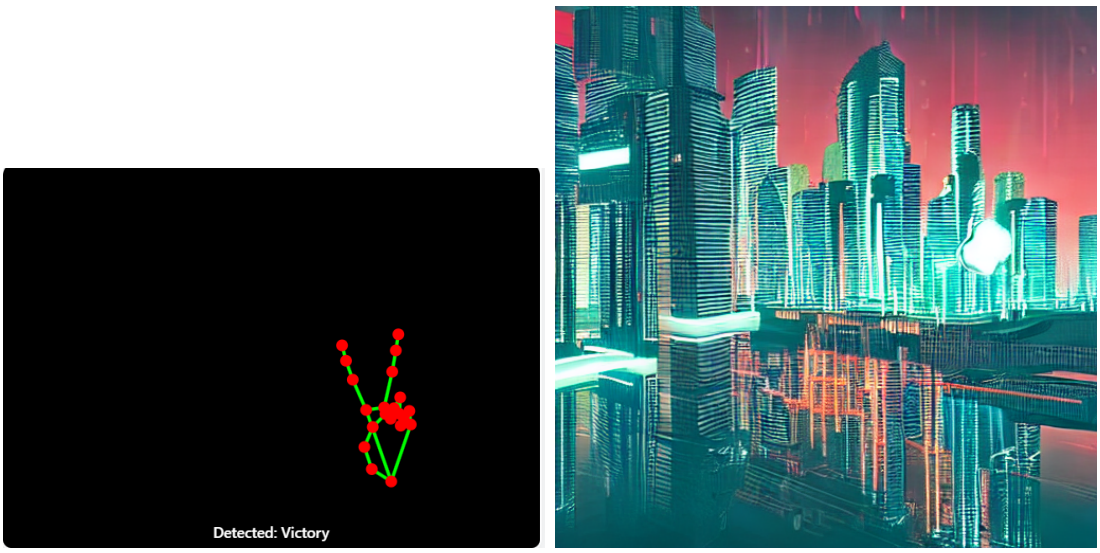

Figure 1: Victory Gesture → Generated Forest Image

Figure 2: Fist Gesture → Generated Ocean Storm Image

The project demonstrates real-time responsiveness and semantic coherence between gesture and generated image, fulfilling the original objective with effective implementation of advanced AI components.

5. References

[R1] Google AI Blog. "Real-time Hand Tracking with MediaPipe." 2020. `https://ai.googleblog.com/2019/08/on-device-real-time-hand-tracking-with.html`

[R2] Rombach, Robin, et al. "High-Resolution Image Synthesis with Latent Diffusion Models." CVPR 2022.

[R3] OpenAI. "CLIP: Learning Transferable Visual Models From Natural Language Supervision." 2021.

[R4] GitHub - CompVis/stable-diffusion: `https://github.com/CompVis/stable-diffusion`

[R5] MediaPipe Framework: `https://github.com/google/mediapipe`