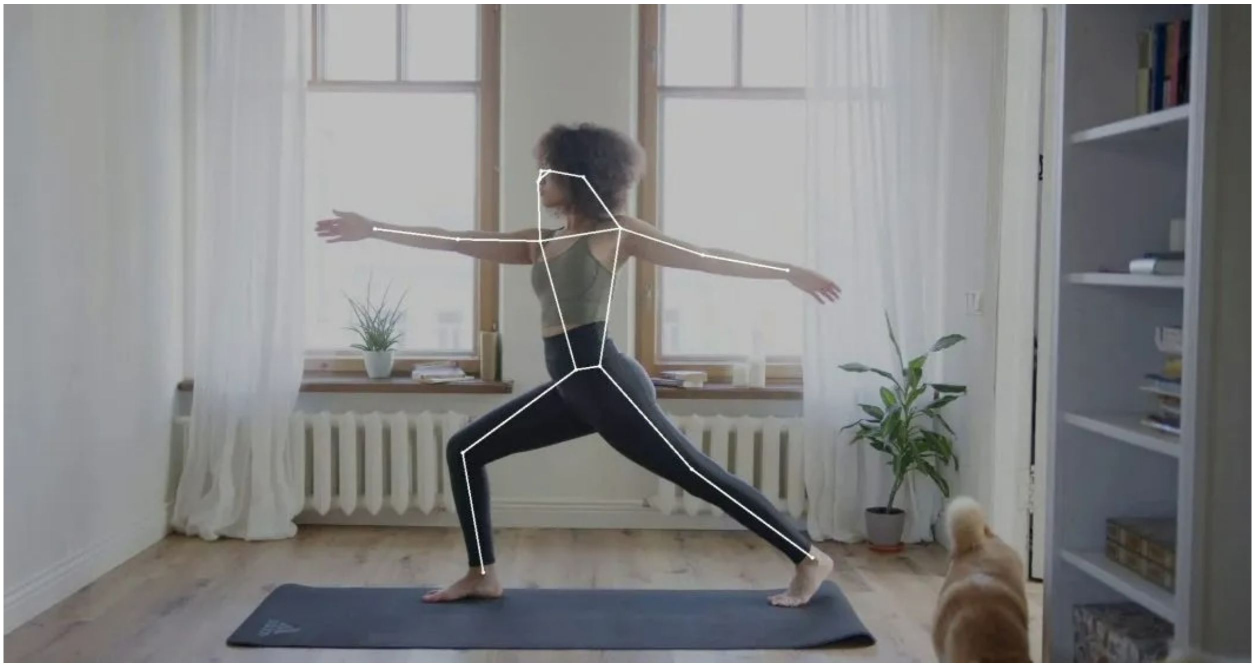
CS3011 - ARTIFICIAL INTELLIGENCE

HUMAN POSE DETECTION



TEAM MEMBERS (GROUP-19) :

20BCS138 - Bhargav Nadipineni

20BCS138 - Pasunuri Aaditya Vardhan

20BCS244 - Vummetla Venkata Sri Datta Charan

20BEC042 - Gurram Sai Eesha

PPT LINK :

https://www.canva.com/design/DAFQmVzFcXI/aQX5oJMK-60xHrhHYimS4w/view?utm_content=DAFQmVzFcXI&utm_campaign=designshare&utm_medium=link&utm_source=publishsharelink

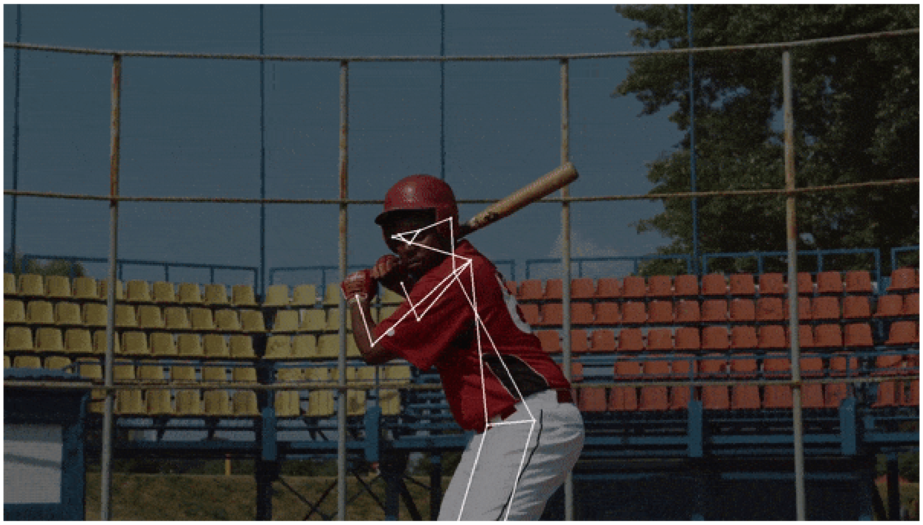
INTRODUCTION :

Human Pose Detection is a computer vision task that includes detecting, associating, and tracking semantic key points. Examples of semantic keypoints are “right shoulders,” “left knees”.

Human pose estimation from video plays a critical role in various applications such as quantifying physical exercises, sign language recognition, and full-body gesture control. For example, it can form the basis for yoga, dance, and fitness applications. It can also enable the overlay of digital content and information on top of the physical world in augmented reality.

Essentially it is a way to capture a set of coordinates for each joint (arm, head, torso, etc.,) which is known as a key point that can describe a pose of a person.

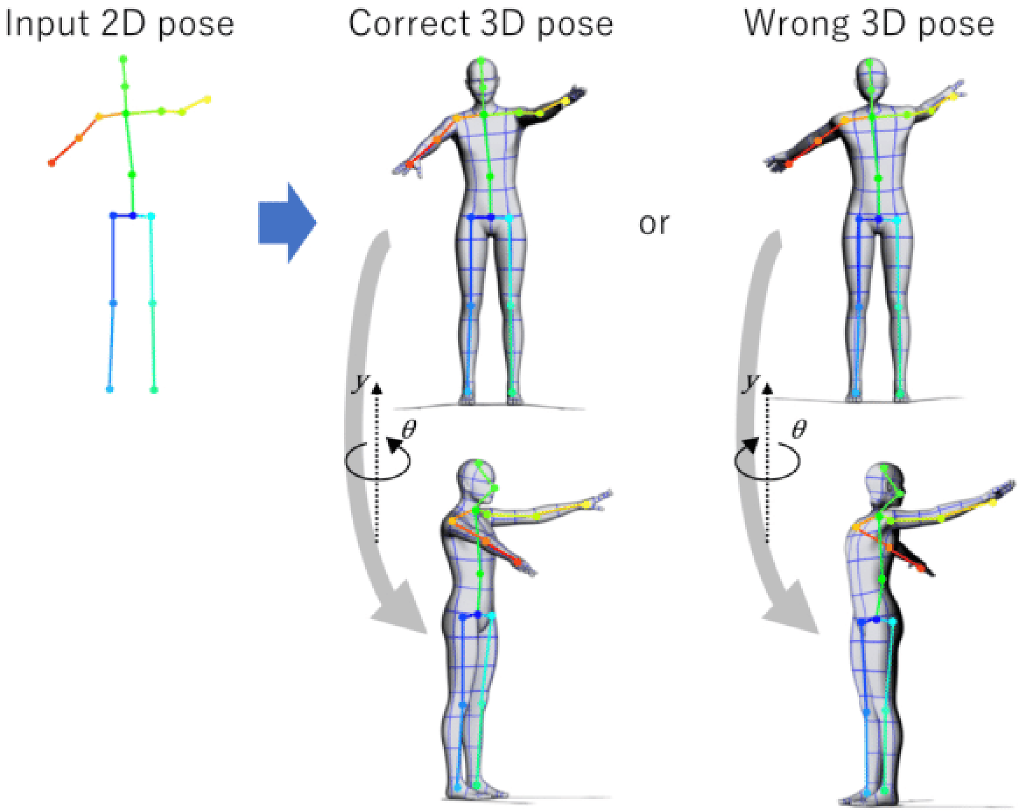
The connection formed between the points has to be significant, which means not all points can form a pair. From the outset, the aim of Human Pose Detection is to form a skeleton-like representation of a human body and then process it further for task-specific applications.



LITERATURE REVIEW :

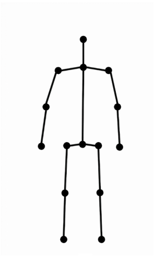
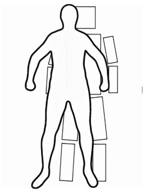
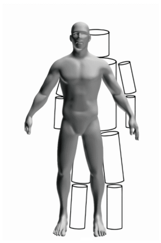
Human pose estimation aims at predicting the poses of human body parts and joints in images or videos. Since pose motions are often driven by some specific human actions, knowing the body pose of a human is critical for action recognition.

2D HUMAN POSE ESTIMATION	3D HUMAN POSE ESTIMATION
2D human pose estimation is used to estimate the 2D position or spatial location of human body key points from visuals such as images and videos. Early computer vision described the human body as a stick figure to obtain global pose structures. However, modern deep learning based approaches have achieved major breakthroughs by improving the performance significantly for both single-person and multi-person pose estimation. Some popular 2D human pose estimation methods include OpenPose, CPN, AlphaPose, and HRNet	3D Human Pose Estimation is used to predict the locations of body joints in 3D space. This field has attracted much interest in recent years since it is used to provide extensive 3D structure information related to the human body. It can be applied to various applications, such as 3D animation industries, virtual or augmented reality, and 3D action prediction. 3D human pose estimation can be performed on monocular images or videos (normal camera feeds).


Input 2D pose
Correct 3D pose
Wrong 3D pose
or
y
θ
y
θ

3D HUMAN BODY MODELING :

In human pose estimation, the location of human body parts is used to build a human body representation (such as a body skeleton pose) from visual input data. Therefore, human body modeling is an important aspect of human pose estimation. It is used to represent features and key points extracted from visual input data. Typically, a model-based approach is used to describe and infer human body poses and render 2D or 3D poses.

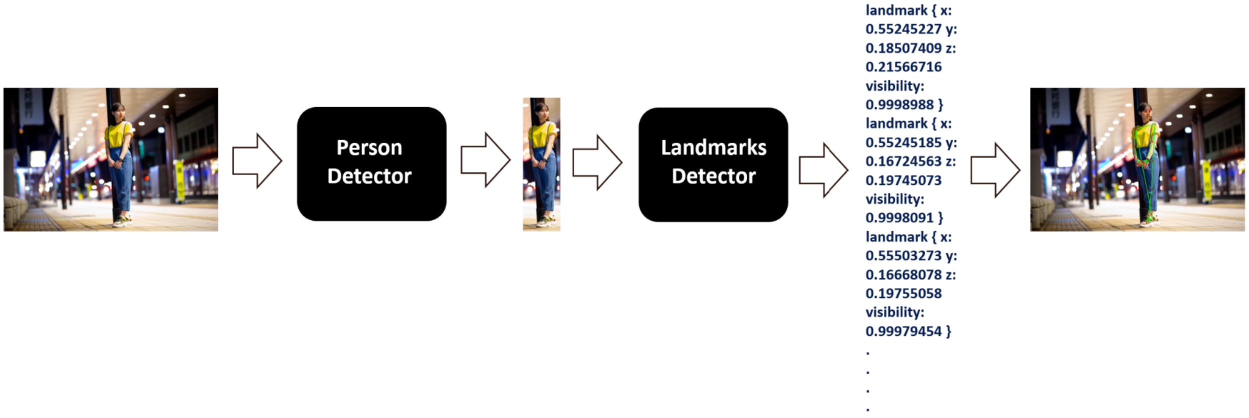
SKELETON - BASED	CONTOUR - BASED	VOLUME - BASED
It is used for 2D pose estimation as well as 3D pose estimation. This flexible and intuitive human body model includes a set of joint positions and limb orientations to represent the human body structure. Therefore, skeleton pose estimation models are used to capture the relations between different body parts. However, kinematic models are limited in representing texture or shape information.	It is used for 2D pose estimation. The planar models are used to represent the appearance and shape of a human body. Usually, body parts are represented by multiple rectangles approximating the human body contours. A popular example is the Active Shape Model (ASM) that is used to capture the full human body graph and the silhouette deformations using principal component analysis.	It is used for 3D pose estimation. There exist multiple popular 3D human body models used for deep learning based 3D human pose estimation for recovering 3D human mesh. For example, GHUM & GHUML(ite), are fully trainable end-to-end deep learning pipelines trained on a high-resolution dataset of full-body scans of over 60’000 human configurations to model statistical and articulated 3D human body shape and pose. It can be used to infer.
 skeleton-based model	 contour-based model	 volume-based model

HOW DOES POSE ESTIMATION WORKS :

Pose estimation operates by finding key points of a person or object. Taking a person, for example, the key points would be joints like the elbow, knees, wrists, etc. There are two types of pose estimation: multi-pose and single pose. **Single pose estimation** is used to estimate the poses of a single object in a given scene, while **multi-pose estimation** is used when detecting poses for multiple objects.

MEDIAPIPE POSE:

MediaPipe Pose is a ML solution for high-fidelity body pose tracking, inferring 33 3D landmarks and background segmentation masks on the whole body from RGB video frames utilizing our BlazePose research that also powers the ML Kit Pose Detection API. Current state-of-the-art approaches rely primarily on powerful desktop environments for inference, whereas our method achieves real-time performance on most modern mobile phones, desktops/laptops, in python and even on the web



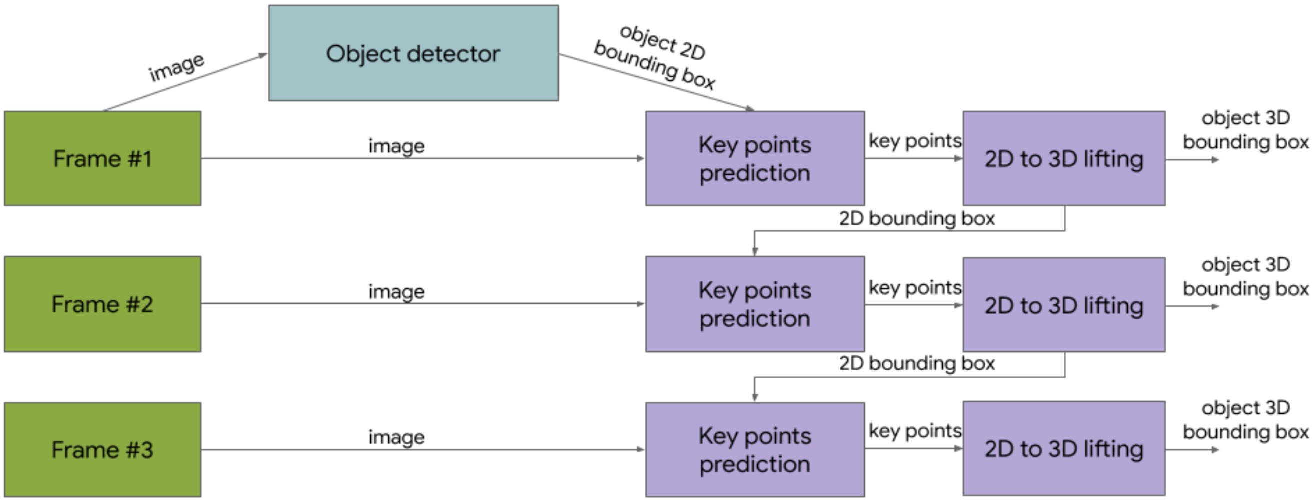
```
landmark { x:
0.55245227 y:
0.18507409 z:
0.21566716
visibility:
0.9998988 }
landmark { x:
0.55245185 y:
0.16724563 z:
0.19745073
visibility:
0.9998091 }
landmark { x:
0.55503273 y:
0.16668078 z:
0.19755058
visibility:
0.99979454 }
.
.
.
.
```

ML PIPELINES FOR 3D OBJECT DETECTION:

There are two ML pipelines to predict the 3D bounding box of an object from a single RGB image: one is a two-stage pipeline and the other is a single-stage pipeline. The two-stage pipeline is 3x faster than the single-stage pipeline with similar or better accuracy. The single stage pipeline is good at detecting multiple objects, whereas the two stage pipeline is good for a single dominant object. Here in out project we implemented 2 staged pipeline

2 - STAGE PIPELINE:

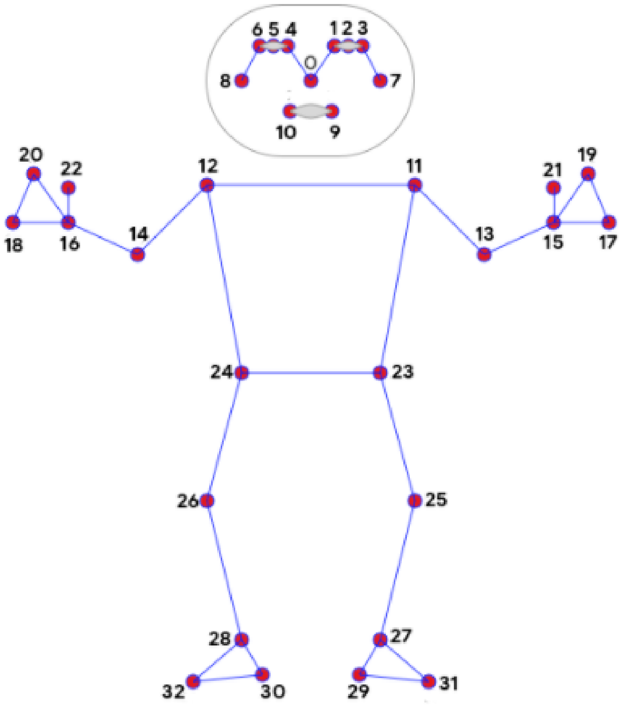
The solution utilizes a two-step detector-tracker ML pipeline, proven to be effective in our MediaPipe Hands and MediaPipe Face Mesh solutions. Using a detector, the pipeline first locates the person/pose region-of-interest (ROI) within the frame. The tracker subsequently predicts the pose landmarks and segmentation mask within the ROI using the ROI-cropped frame as input. Note that for video use cases the detector is invoked only as needed, i.e., for the very first frame and when the tracker could no longer identify body pose presence in the previous frame. For other frames the pipeline simply derives the ROI from the previous frame's pose landmarks.



MODELS :

POSE LANDMARK MODEL :

The landmark model in MediaPipe Pose predicts the location of 33 pose landmarks. Optionally, MediaPipe Pose can predict a full-body segmentation mask represented as a two-class segmentation (human or background).



0. nose
1. left_eye_inner
2. left_eye
3. left_eye_outer
4. right_eye_inner
5. right_eye
6. right_eye_outer
7. left_ear
8. right_ear
9. mouth_left
10. mouth_right
11. left_shoulder
12. right_shoulder
13. left_elbow
14. right_elbow
15. left_wrist
16. right_wrist
17. left_pinky
18. right_pinky
19. left_index
20. right_index
21. left_thumb
22. right_thumb
23. left_hip
24. right_hip
25. left_knee
26. right_knee
27. left_ankle
28. right_ankle
29. left_heel
30. right_heel
31. left_foot_index
32. right_foot_index

OUTPUT

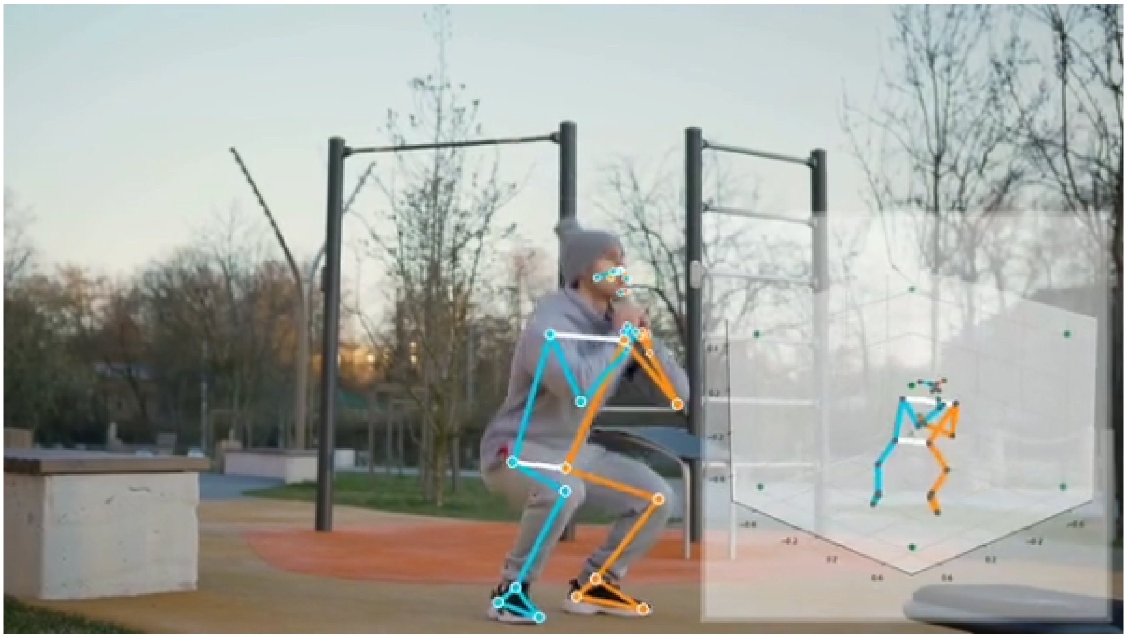
Naming style may differ slightly across platforms/languages.

POSE_LANDMARKS

A list of pose landmarks. Each landmark consists of the following:

- `x` and `y`: Landmark coordinates normalized to `[0.0, 1.0]` by the image width and height respectively.
- `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`.
- `visibility`: A value in `[0.0, 1.0]` indicating the likelihood of the landmark being visible (present and not occluded) in the image.

Fig 5. Example of MediaPipe Pose real-world 3D coordinates.



PROBLEMS FACED / CHALLENGES :

Pose estimation is considered a hard problem in computer vision. Here are a few key challenges faced by current pose estimation algorithms:

1. Differences in human visual appearance in input images due to changes in clothing, skin color, differences in physique between individuals, etc.
2. Differences in lighting conditions, weather, viewing angle, background context, etc.
3. Partial occlusion (objects or other humans obstructing the subject of the analysis)
4. The complexity of the human skeletal structure can make it difficult to identify exact joint coordinates, especially for small points which are barely visible in the image.

High dimensionality of the pose.

1. Loss of three-dimensional information as a result of observing the pose from a two-dimensional image. Obtaining and annotating 3D pose images is complex and expensive.
2. Pose estimation in social scenes is particularly difficult, because of occlusions, interactions between people, and constant change of movements and gestures..

Human pose estimation is a challenging task as the body’s appearance joins changes dynamically due to diverse forms of clothes, arbitrary occlusion, occlusions due to the viewing angle, and background contexts. Pose estimation needs to be robust to challenging real-world variations such as are lighting and weather. Therefore, it is challenging for image processing models to identify the fine-grained joint coordinates. It is especially difficult to track small and barely visible joints.

APPLICATIONS OF POSE DETECTION :

Human Activity Estimation :

A rather obvious application of pose estimation is tracking and measuring human activity and movement. Architectures like DensePose, PoseNet, or OpenPose are often used for activity, gesture, or gait recognition. Examples of human activity tracking via the use of pose estimation include:

Application for detecting sitting gestures
Full body/sign language communication (for example, traffic policemen signals)
Applications to detect if a person has fallen down or is sick
Applications to support the analysis of football, basketball, and sports
Applications to analyze dance techniques (for example, in ballet dances)
Application of posture learning for body works and finesses
Applications in security and surveillance enhancement

Augmented Reality and Virtual Reality :

As of today, pose estimation interfaced with augmented and virtual reality applications gives users a better online experience. For instance, users can virtually learn how to play games like tennis via virtual tutors who are pose represented.

More so, pose estimators can also be interfaced with augmented reality-based applications. For example, The United States Army experiments with augmented reality programs to be used in combat. These programs aim to help soldiers distinguish between enemies and friendly troops, as well as improve night vision.

Action prediction Surveillance Cloth Parsing Online Coaching

Movie and Game AR and VR Healthcare

Training Robots With Human Pose Tracking:

Typical use cases of pose estimators is in the application of making robots learn certain crafts. In place of manually programming robots to follow trajectories, robots can be made to learn actions and movements by following the tutor's posture look or appearance.

Human Motion Tracking for Consoles

Other applications of pose estimation are in-game applications, where human subjects auto-generate and inject poses into the game environment for an interactive gaming experience. For instance, Microsoft's Kinect used 3D pose estimation (using IR sensor data) to track the motion of the human players and to use it to render the actions of the characters virtually into the gaming environment.