mosaic
DATA SCIENCE

http://mosaicdatascience.com/

info@mosaicdatascience.com

Data Science White Paper

Using Computer Vision Techniques for Pose Estimation, Subject Tracking, & Gaze Detection

Mosaic developed a proof-of-concept (POC) system that uses computer vision technologies to infer human subjects' positions, proximity, and gaze directions from video feeds. Future human missions to Mars will pose mental and behavioral challenges for crew members. Travel will require long distances and isolation from Earth, and the crew members face limited communication with mission control, family, and friends.

Machine Learning

Mosaic data scientists collaborate with customers, digging deep into the data to inform design and deployment of custom ML tools that make a difference.

Artificial Intelligence

Mosaic integrates powerful AI tools into clients' existing technology stack to solve complex business challenges

Business Analytics

Mosaic helps corporations of all shapes and sizes take advantage of their data, transforming their decision-making processes.

INTRODUCTION

What is pose estimation?

Pose estimation, subject tracking, and gaze detection are emerging fields of computer vision that predict and track the location of a person or object. The combination of position and orientation is referred to as the **pose** of an object, even though this concept is sometimes used only to describe the orientation.[1]

Deep learning accomplishes this task by learning the visual data points of pose and the orientation of a given person/object; the algorithm will then predict where that object/human is in the real world based on a large set of training images.

Why it matters?

Artificial intelligence, computer vision specifically, can track an object, person, or multiple people in real-world space at a granular level. The ability to do this accurately opens a wide range of potential use cases. The techniques and algorithms involved in this task differ from classic object detection. Object detection is usually coarse-grained, placing bounding boxes on the object. Pose estimation and human tracking takes this further by projecting the precise location of critical points associated with the object/person.

The application of tracking human/object movement automatically, in real-time, can unlock new insights from personal training to monitoring production processes to alerting of safety issues on a factory floor or operational field.

Potential Applications

The applicability of this technology is vast. Early adopters are seeing success in tracking human activity and movement, virtual reality, gaming, risk management, and robotics. Training a model to identify human movement has far-reaching implications in many industries.

A professional sports team can identify successful movements, a virtual trainer can provide tips to maximize training, and a retailer can count crowds measuring foot traffic.

Motion capture in animation & gaming can be more accurate & automated using deep learning approaches. A creator no longer must rely on markers or humans in specialized suits; they can capture in real-time with pose estimation.

Industrial robotics can use these techniques to help machines better orient itself and parts to the environment. Historical approaches to the pose (combination of position & orientation) of the robot have seen several limitations, especially in calibrating where a robot should move, which is brittle under operationally shifting environments. Artificially intelligent systems collect images from the robot, computer vision classifies different objects in these images, and pose estimation provides spatial awareness for optimal movements.

Adoption Challenges

This technology is compelling but also requires a deep understanding of the mechanics behind computer vision algorithms. Pose estimation can be a game-changing technology, but models need to be trained and deployed by a data scientist who knows what they are doing. Mosaic Data Science specializes in computer vision development and was tasked by NASA to monitor the crew health for deep space missions in an R&D effort. In the remainder of the whitepaper, Mosaic will highlight this deep learning deployment.

Figure 1: Example of team relative positioning (blue dots) and gaze angle (gold arrows) as indicators of attitudinal status.[3]

REAL-WORLD APPLICATION | MONITORING NASA SPACE CREWS FOR DEEP SPACE MISSIONS

Mosaic developed a proof-of-concept (POC) system that uses computer vision (CV) technologies to infer human subjects' positions, proximity, and gaze directions from video feeds. Future human missions to Mars will pose mental and behavioral challenges for crew members. Travel will require long distances and isolation from Earth, and the crew members face limited communication with mission control, family, and friends. Accordingly, breakdowns in team cohesion and interpersonal conflict could have critical consequences on the safety of the crew and the success of the mission. Thus, the ability to monitor the crew's behavior is essential to the success of future deep-space missions. Previous NASA studies show that crew members' relative proximity and position within a habitat can be a marker of team cohesion and social dynamics.

The POC tracks people throughout a video, calculates their relative position, and estimates their body positioning and the direction where their face is pointing. This has been implemented through various stages:

1. Identifying and tracking crew members from video
2. Proximity measurement between crew members
3. Pose estimation

CREW MEMBER IDENTIFICATION/ TRACKING

Person detection in videos involves verifying the person's presence in image sequences and possibly locating it precisely for recognition and tracking monitors each person's spatial and temporal changes during a video sequence, including his presence, position, size, shape, etc.

For this, we use deep learning–CNN (Convolutional neural networks) based detection and tracking algorithm. For the detection task, the model uses an already trained neural network, trained on millions of human images, and extracts from a video frame a 128-dim vector for each potential bounding box which captures the key features of the box and is used to determine whether a person is captured within the box. The output of this detection is bounding boxes for each person detected, along with the associated feature vector.

The tracking was then carried out by utilizing the computer vision object tracking model DeepSort-YOLOv3 [BN. Wojke, 2017]. The model first uses the YOLO (you only look once) algorithm that detects objects using a convolutional neural network (CNN). The algorithm tracks the position, velocity, and object's appearance based on the feature vector collected in the detection process. The algorithm then predicts where it is likely to be in the following frames and compares the position of any bounding boxes with those predictions. The detection and tracking output will be a set of bounding boxes, each representing a person detected in the frame. Figure 2 shows a workflow for the object tracking algorithm. Figure 3 shows the output of the detection and tracking algorithm done on the video frames.

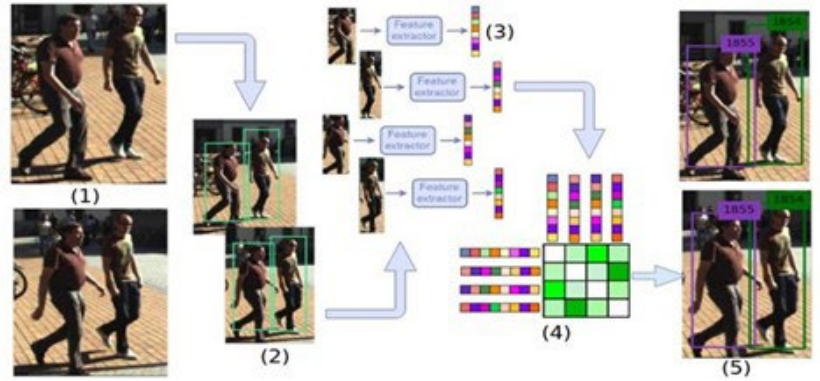


Figure 2: Workflow algorithm for object tracking algorithm. Given the raw frames of a video (1), an object detector is run to obtain the bounding boxes of the ob- jects (2). Then, for every detected object, different features are computed, usually visual and motion ones (3). After that, an affinity computation step calculates the probability of two objects belonging to the same target (4), and finally an association step assigns a numerical ID to each object (5).[4]

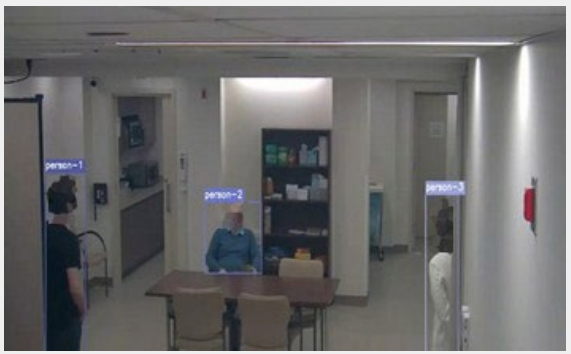


Figure 3: Example of the output bounding boxes (labeled person-1, person-2, person-3) for the detection and tracking done on the video frames. Each bounding box represents a different object (person) identified by the algorithm.

PROXIMITY MEASUREMENT

After person detection and tracking in each frame, we had to measure the proximity between the crew members (i.e., compute the distance between people in an image). For this POC, we used a simple method for proximity measurement. Once we know the bounding box for each person, we can compute the distance between any two people. But the challenge here was to select the right coordinate for representing a person as a bounding box is in the form of a rectangle. For this, we used the bottom center of a rectangle to represent each person to measure the distance. Based on this, we computed the Euclidean distance between each possible pair of centroids.

An approximate distance (in meters) between crew members are shown in Figure 5.

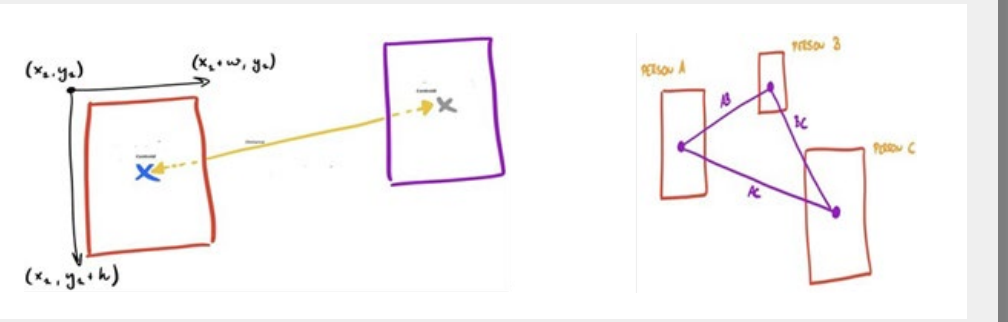


Figure 4a) calculating distances between centroids of the bounding boxes. 4b) calculating different permutations between centroids.

Figure 5: Output of the proximity measurement between identified persons in a video frame.

FURTHER IMPROVEMENTS

In this POC, and for simplicity, Mosaic didn't consider the camera's projection. However, further improvement to our approach is using the universal process (i.e., converting a video into a top view or bird's eye view and then computing the distance between two objects in an image). This task is known as **Camera Calibration**. More details about camera calibration can be found here[2].

POSE ESTIMATION & SUBJECT TRACKING

As the direction of an individual's gaze onto other team members can indicate their relationship's health status, Mosaic also studied the body pose estimation and the gaze direction of each person in the video feeds.

To estimate the pose and facial orientation of subjects in the video, Mosaic used the publicly available AlphaPose model (https://github.com/MVIG-SJTU/AlphaPose). The algorithm first identifies key landmarks on a person's face and body to construct a simple "skeleton" defining that person's posture and positioning. The algorithm then finds a linear transformation that can map "typical" 3D locations of facial features (in real-world coordinates) onto the 2D coordinates of landmarks in imaging plane. This linear transformation tells you the orientation of the world coordinates relative to the camera. So, to estimate the angle of someone's head, Mosaic takes the facial landmarks from AlphaPose (namely the ears, eyes, and nose); given typical spacing of these facial features, Mosaic builds a 3D model of these points and uses a linear transformation to map those points onto a 2D imaging plane to measure landmarks. The 3D world coordinates of the face are then used to estimate the angle of the face.

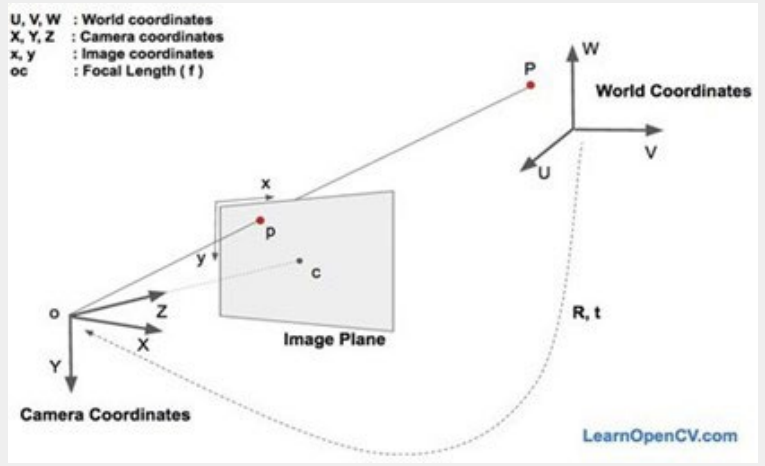


Figure 6 : Pose estimation.[5]

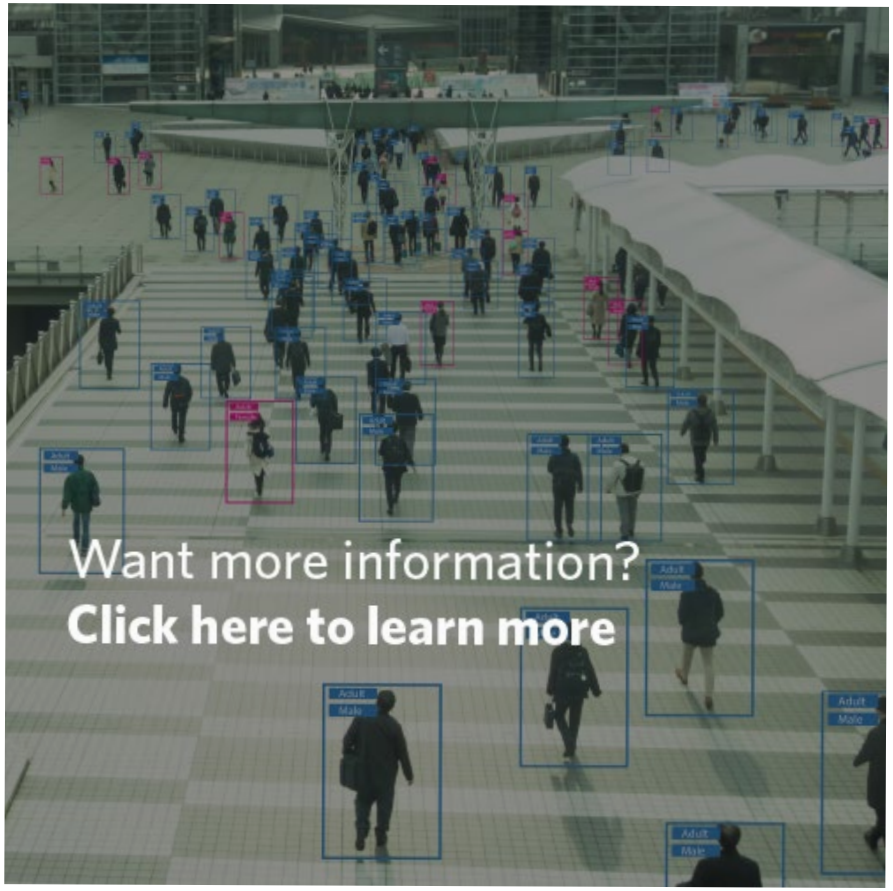
Figure 7 shows the results of body pose estimation and estimation of the direction of each person's face (yellow and teal lines emanating from each face). The yellow line indicates the up and down angle while the teal line indicates the left-right angle.

Figure 7: Output of pose estimation for each person in a video frame.

CONCLUSION

The proof of concept is still under development by Mosaic and awaits a future funding decision from NASA. The concept has shown NASA can monitor the crew's health by analyzing human movement, providing them with many insights they did not previously possess.

After reading this whitepaper, hopefully, you realize the application of pose estimation is worth the technical effort required to deploy it correctly. Artificial intelligence can automate several tasks helping organizations save money, mitigate risk, and provide better experiences.



Endnotes

1. BN. Wojke, A. Bewley and D. Paulus, Simple online and realtime tracking with a deep association metric, 2017 IEEE International Conference on Image Processing, pp.3645-3649, 2017

2. https://developer.ridgerun.com/wiki/index.php?title=Birds_Eye_View/Introduction/Research

3. Image credit: yanalya at Freepik.com

4. Ciaparrone, G.,Sánchez, F.L., Tabik, S., Troiano, L., Tagliaferri, R., & Herrera, F. (2020). Deep Learning in Video Multi-Object Tracking: A Survey. Neurocomputing, 381, 61-88.)

5. https://learnopencv.com/head-pose-estimation-using-opencv-and-dlib/