**Department of Electrical, Computer, and Software Engineering**

**Part IV Research Project**

Literature Review and Statement of Research Intent

Project Number:

Project #65: Guide and logistics robot system: taking elevator with interaction

Finn Tracey

Cale Ying

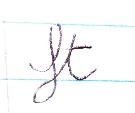
Ho Seok Ahn

JongYoon Lim

08/08/2023

**Declaration of Originality**

This report is my own unaided work and was not copied from nor written in collaboration with any other person.



Name: Finn Tracey

**ABSTRACT:** Human robot interaction is a topic that has been theorized for a long time, and recent advances in technology brings us closer to making it a reality. Our goal is to implement a guide robot with functionality beyond tis primary directive, decreasing the interactivity boundary between humans and robots.

# Project Scope

To achieve the final goal of an interactable guide robot, we first need to establish a scope in which we will undertake our research within. This is to ensure that we do not stray too far from the original intent, or waste time in areas that de not advance the progress of our project. So far, we have established that our project scope is:

* Use the Choregraphe simulator to experiment the capabilities of the Pepper robot in an ideal environment and interact with the NAOqi operating system.
* Implement real time human tracking and motion mimicking into the Pepper robot.
* Implement face emotion recognition and emotion mimicking into the Pepper robot.
* Implement an interactive chatbot into the Pepper robot.

The scope of the project will fall within the region of the stated topics. For example, gaining an understanding of how the robot can recognize human joints to estimate a human pose, or how the robot can track certain facial features to convey emotions effectively.

# Research Intent

This project aims to advance the inclusion of robot guides in our modern society by proving both its usefulness in guiding people around buildings, as well as its potential for interactivity beyond its primary functionality. The current issue that our project answers is the outdated implementation of current guidance systems i.e: maps. As buildings tend to be more complex with several levels, so too does the task of traversing the building. The addition of robot guides could prove to solve this issue. We believe that our research can assist the advancement of current robot guide systems as a proof of concept and prove that its functionality should not be overlooked as superfluous to current solutions. We will achieve our results by first undertaking research about each function (Motion mimicking, facial expression detection, chatbot interactivity, guidance system), and once we feel that sufficient research has been achieved, move to research the next area while also working to implement the functionality of the previous area.

# Literature Review

Throughout the course of the year, we will undertake research on many topics related to our objectives i.e., motion mimicking requires research about Human Pose Estimation.

## System Diagram

So far, we have undertaken many hours of research on motion mimicking, and we have a system diagram demonstrating step-by-step on how we will achieve this backed by this research:

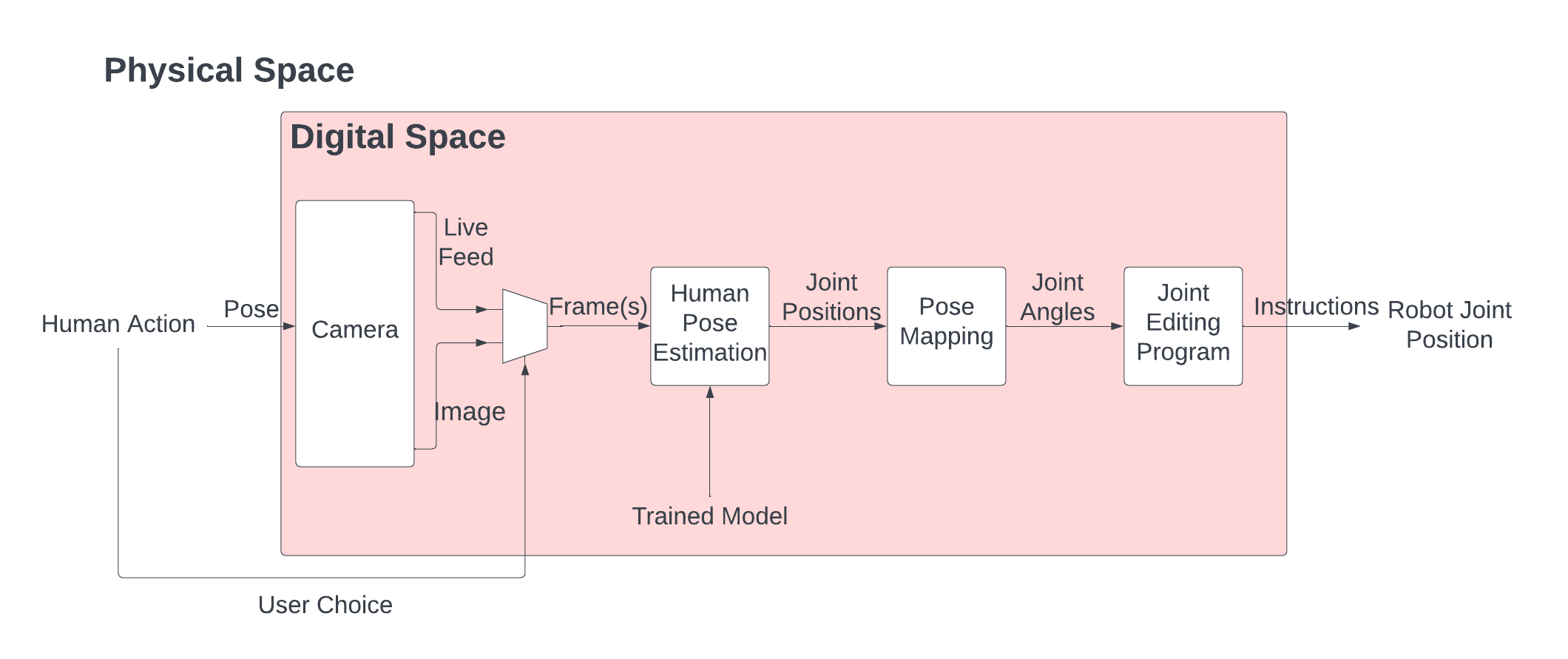


Fig. 1. System Diagram for Human Motion Mimicking.

In this system diagram, the system boundary within ‘Physical Space’ is ‘Digital Space’. These boundaries are self-explanatory, where the digital space contains all the processing, and the physical space is what the user will witness/experience. However, the two inter-space transitional input/outputs; ‘Pose’ and ‘Instructions’, are ambiguous in these boundaries and do not entirely conform to this definition.

## Human Action

We plan for the robot to have a range of different functions, from a chatbot to face emotion recognition, so we will require the user to select the motion mimic program, as well as whether the user would like the robot to track them frame-by-frame in real time, or whether they would like the robot to react to a single frame i.e., an image.

## Camera

The Pepper robot is equipped with three different cameras: Two 2D RGB HD cameras located on the mouth and forehead, and one 3D camera located behind the eyes for obstacle detection and depth perception. We plan to use one of the 2D cameras to track the user, however, we are open to switching to both 2D cameras or the 3D camera if they improve our current model without impacting it’s performance. The 2D camera angles can be found in [1]

### Images

One of the two possible outputs of the camera subsection is ‘Images’. The image of the users pose is taken through one of the cameras using the ALPhotoCaptureProxy method. This frame is processed in the ‘Human Pose Estimation’ subsection.

### Live Feed

The other output of the camera subsection is the ‘Live Feed’. This is a video stream where the video is broken down into individual frames and processed similarly to ‘Images’. The video of the users pose is taken through one of the cameras using the ALVidoeRecorderProxy method. These frames are individually processed in the ‘Human Pose Estimation’ subsection.

## Human Pose Estimation (HPE)

Human Pose Estimation, henceforth referred to as HPE, is the subsection we are most familiar with as this is where the bulk of our research efforts have gone into. Among all our research, common themes and methods have been observed. Recognizing these common occurrences and synthesizing ideas from multiple papers is key to fitting HPE to our purpose.

### Datasets

Every approach to HPE uses some variation of a Convolutional Neural Network, and these networks are trained using some dataset to create a working model. There are three popular datasets that many of our studied models employ with individual skeleton models/keypoints included in the training split of the datasets:

1. COCO

COCO (Common Objects in COntext) is an object detecting dataset containing many images of various objects. This dataset is separated into different subsets. Keypoint Detection dataset is used as it provides all the necessary information for joint detection and doesn’t include excess data that could potentially hinder performance. More information is available in [2]

1. MPII

MPII (Max-Planck Institute for Informatics) is a dataset that focuses more on collecting data from people in their natural environments as well as performing activities. This gives a wider variety compared to other datasets, as it provides data on basic human poses, as well as humans in extremely varying positions. More information is available in [3].

1. LSPe

LSPe (Leeds Sports Pose extended) is a dataset focused on data surrounding humans in an athletic state, resulting in wildly different poses. This aims to capture poses from one extreme to the other, without considering idle or relaxed positions. More information is available in [4].

### Multi-Person Joint Detection

Many popular methods and implementations tend to focus on where modern research is heading towards, which is currently multi-person joint detection. There are options to determine whether your model detects a single person or multiple people described in [5]. With multi-person HPE comes two different pipeline models:

* Top-Down pipeline: “Localize the humans in the image or video and then estimate the parts followed by calculating the pose.”
* Bottom-Up pipeline: “Estimate the human body parts in the image followed by calculating the pose.”

### Research Papers

This section contains annotations of the relevant research papers that will help us understand more about the topic of HPE.

#### Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields

This paper [6] was researched as one of our favored approaches to HPE, OpenPose, uses Part Affinity Fields in its method as well as modifying its branching CNN architecture. The Part Affinity Field (PAF) method is a bottom-up, nonparametric representation that aims to learn to associate body parts with individuals in images. The method described in this paper revolves around stage (b) and (c), where a neural network predicts the confidence maps and a set of vector fields of part affinities. The confidence map aims to locate the body parts, while the part affinity vector fields aim to encode a degree of association to each body part.

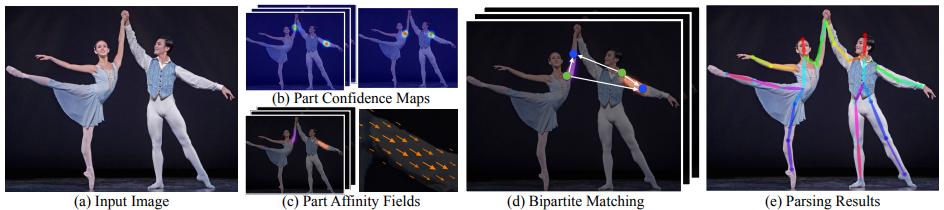


Fig. 2a. Overall Pipeline of HPE [6]

Diagram

Description automatically generated

Fig. 2b. Architecture of Multi-Stage CNN [6]

#### The Progress of Human Pose Estimation: A Survey and Taxonomy of Models Applied in 2D Human Pose Estimation

This paper [7] describes the current state of HPE and existing solutions. The main process of HPE is broken down into two steps: localizing human body joints and grouping those joints into valid human pose configurations.

#### DeepPose

DeepPose [8] is an approach to HPE that we found early in our research and attempted to implement ourselves to no avail, possibly because of inexperience and lack of research. This approach may be revisited if issues arise with current approaches. DeepPose is a more basic solution to HPE compared to other popular approaches, where it produces outputs for one single human and works in 2D. This approach to HPE treats the problem as regression.

#### OpenPose

As stated in **3.4.2.1**, this approach to HPE [9] utilizes PAFs and confidence maps and employs the same pipeline method shown in figure 6a and a modified version of their multi-stage CNN as shown in figure 8. OpenPose utilizes the bottom-up approach to body localization, as the complexity of the person detector rises with more people present in the image. It is stated that “…the top-down approach faces early commitment issues…”, which proves to be an issue if the human detector begins with multiple people close together while there is no recourse for recovery if failure occurs. For single-body pose estimation, the top-down approach is no issue, however, OpenPose is aimed towards multiple bodies, and really benefits from the robustness of the bottom-up approach.

Diagram

Description automatically generated

Fig. 3. Architecture of OpenPose CNN [9]

### Joint Positions

The product of any HPE model aim to have an output of points and skeleton models on an image. We can use these models to calculate the joint positions relative to the size of the image. Some implementations can output joint positions in other file types i.e: OpenPose can output .json files for each frame. This can be carried into the pose mapping subsection.

## Pose Mapping

Pose mapping includes taking the output of an HPE model and transforming the information into a modular form. This allows us to bridge the gap between the HPE program and joint adjustment program. There are two different proposed methods to pose mapping:

### Joint Angles

Our current method involves angles between joints. Angles and distances are very basic measurements and can allow any sort of information to work in almost any form of problem, provided the translation from information to angles is possible. In our case, it is indeed possible if keypoints on a body can be linked to other keypoints.

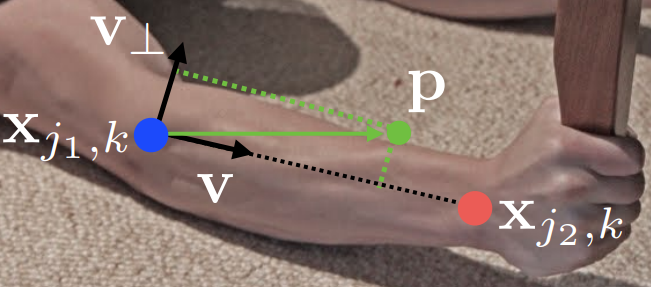


Fig. 4. Angles and Distances Between Keypoints [6]

### Classification

A theorized method where research is sparse is image classification. This involves creating a neural network model that predicts the pose based on a dataset of other poses. This is the less favored method.

## Joint Editing Program

From this point, we have refined, modular information regarding the pose of a person which can finally be used to move the robot itself. Using ALMotionProxy method, we can finely adjust the angle of each joint of the robot as well as the speed of which the joints can move at. Using the joint angle method can potentially replicate a model with relatively high accuracy and is a modular method itself as we could add more parameters to increase accuracy even further i.e: third dimension, degree of precision (joint detection precision down to the fingers), etc.

## Robot Joint Positions

The final subsection of our system diagram is the real robot joint positions and is our goal for this part of the project. This is the most crucial step in terms of safety as this is the robot interacting with ‘Physical Space’, so multiple safety plans must be put in place to prevent the robot from damaging itself and/or harming the user. Some theorized general safety methods include:

* A monitor in the joint adjustment code observes the relative positions between the robot’s limbs that are under threat of collision.
* An emergency routine allows the user to instantly stop the robot in case of an immediate threat.

As we continue the project, we will be sure to add more systems to improve the overall safety of the robot and any user that interacts with it.

# Discussions

At this point in time, we have essentially completed the HPE section, with all the other subsections still in progress. With OpenPose, we have achieved a consistent frame rate through extensive testing, and .json outputs of joint positions. BlazePose has skeleton models output in 3D. The Choregraphe simulator has movement through the ALMotionProxy method.

# Conclusions

Our project aims to implement a guide robot while also improving human-robot interaction. One method to address this problem is a motion mimicking program, where we have a system with extensive research, particularly on human pose estimation, to achieve this. Throughout this literature review is proof that this concept is possible.

# Acknowledgements

I would like to thank Ho Seok Ahn and JongYoon Lim for supervising our research and assisting us in the progress of our project. We acknowledge the efforts of the authors of OpenPose, DeepPose, The Progress of Human Pose Estimation, and Pose Estimation using Part Affinity Fields, as these papers have expanded our knowledge in this field. This project is supported by The University of Auckland.

# References

1. “Pepper – 2D Cameras”. Aldebaran Documentation. http://doc.aldebaran.com/2-0/family/juliette\_technical/video\_juliette.html.
2. ‌“COCO 2020 Keypoint Detection Task”. COCO Dataset. https://cocodataset.org/#keypoints-2020
3. MPII Human Pose Dataset. http://human-pose.mpi-inf.mpg.de/
4. “LSPe - Leeds Sports Pose Extended”. dbcollection. https://dbcollection.readthedocs.io/en/latest/datasets/leeds\_sports\_pose\_extended.html
5. “A Comprehensive Guide to Human Pose Estimation”. V7labs https://www.v7labs.com/blog/human-pose-estimation-guide
6. Z. Cao, et al. “Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields”. The Robotics Institute, Carnegie Mellon University. 14 April 2017. Accessed: 16 April 2023. [Online]. Available: https://arxiv.org/pdf/1611.08050.pdf
7. T. L. Munea, et al. “The Progress of Human Pose Estimation: A Survey and Taxonomy of Models Applied in 2D Human Pose Estimation” in IEEE Access, vol. 8, pp. 133330-133348, 2020, doi: 10.1109/ACCESS.2020.3010248. Date Accessed: 9 April 2023.
8. A. Toshev C. Szegedy. “DeepPose: Human Pose Estimation via Deep Neural Networks” Google, 1600 Amphitheatre Pkwy, Mountain View, CA 94043. 20 August 2014. Date Accessed: 7 March 2023. [Online]. Available: https://arxiv.org/pdf/1312.4659.pdf
9. Z. Cao, et al. “OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields” 30 May 2019. Date Accessed: 6 April 2023. [Online]. Available: https://arxiv.org/pdf/1812.08008.pdf