Hand Gesture Recognition

Samuel Oncken, Steven Claypool

CONCEPT OF OPERATIONS

REVISION - 3
3 October 2022

CONCEPT OF OPERATIONS
FOR
Hand Gesture Recognition

TEAM <72>

APPROVED BY:

Samuel Oncken	10/3/2022
_____	
Project Leader	Date
_____	
Prof. Kalafatis	Date
_____	
T/A	Date

Change Record

Rev.	Date	Originator	Approvals	Description
1	9/15/2022	Samuel Oncken Steven Claypool		Draft Release
2	9/28/2022	Samuel Oncken Steven Claypool		Revision for FSR Release
3	11/28/2022	Samuel Oncken Steven Claypool		Revision for Final Report Release

Table of Contents

List of Figures

1. Executive Summary

To create large training sets for neural networks, our solution is to create a virtual environment in Unity as opposed to using real humans and equipment. The virtual environment takes as input a real hand gesture dataset recorded by a user with the Leap Motion Controller, applies the gestures to hundreds of diverse human models from MakeHuman (imported into Unity), and records images and videos to build a virtual hand gesture training set. This virtual training set requires significantly less time and personnel and additionally will train the neural networks to similar if not improved gesture recognition accuracy when compared to existing real gesture training sets. Using our virtual environment to create synthetic gesture datasets, users can quickly and easily build reliable training sets tailored precisely to their application.

2. Introduction

Building neural network training sets for hand gesture recognition takes significant time and personnel to get diverse and comprehensive data. Thousands of images or videos must be taken manually to build a training set for a new set of hand gestures. To bypass this, we will create a virtual environment that builds training sets from gesture data recorded by one individual that is applied to hundreds of diverse human models and recorded. The neural networks would show similar accuracy in hand gesture recognition with the virtual training sets, potentially replacing the need for "real" training sets. This would notably expedite the creation of new training sets and simplify the entire process.

2.1. Background

Many scholarly articles covering the concept and feasibility of computer vision have been written in the last decade. However, it is no longer an argument that artificial intelligence and machine learning are here to stay. Through our research of the related work done in the field of gesture recognition, we have recorded a wide variety of methods to bring the idea of computer vision into light. In our project, we will not be focusing on the algorithms used in the creation of neural networks such as Convolutional Neural Networks (CNN) [1,2], but instead we will aim to uncover how the construction of a training set can alter the accuracy of a gesture recognition neural network.

Many of the datasets we have found online, summarized by [3], require a handful of real subjects to perform similar actions in front of a motion sensor (Kinect, Wii remote, etc.) which is recording from a set location. By using a number of different human hand models for the same gesture, some variance between gestures is collected which is required given that a gesture recognition system must not only recognize one size or shape of hand for it to function properly. Other datasets like the American Sign Language Lexicon Video Dataset and those derived from it [4] record human subjects using synchronized cameras all recording from different angles. Once again, this is beneficial to the machine learning process because in practice, a system will not always be looking at a human from a single angle.

Our research is aimed at bringing these important considerations together by generating a "virtual" dataset. By virtual, we mean that instead of having a set number of human subjects come into a studio to record movements, we will be recording the movements of one human (under various conditions such as lighting, distance from sensor, etc.) and applying those movements to randomly generated 3-D human models within the Unity game engine environment. After the virtual human model performs the gesture, we will record images and videos from several virtual camera angles to ensure that our training set aids in the recognition of a gesture from all directions. Similar research has been done using a virtual train set of hand gestures[5]. We are looking to expand upon this student's research because creating a virtual training set is a much more cost-effective method of gathering data that still produces high accuracy results, which we are seeking to prove when testing our virtual datasets with state of the art hand gesture recognition neural networks.

A major obstacle in human gesture recognition is that reliable datasets are limited in the number of gestures they contain when compared to the vast number of gestures humans use every day. Our solution makes it fast and easy to create entirely new datasets consisting of application specific gestures, resulting in much more diverse gesture sets available for use and reducing the initial limitation.

2.2. Overview

We will design a virtual environment that can be used to create training sets for hand gesture recognition neural networks. This virtual environment will be implemented through Unity and will be able to map hand gestures recorded through the Leap Motion Controller onto randomly generated human models. We will be recording the gestures from a first-person point of view, mounted specifically on the head or chest of the individual performing the gestures.

The Unity environment will have a script that places virtual cameras at random locations in front of the human model performing the gesture (3rd person point of view) to collect varying angles of each gesture. The final script written for the Unity environment will generate and import a random human model, apply a randomly chosen “animation” from the real hand gesture database to the model, randomly place multiple virtual cameras in front of the model, then record and store the images taken as members of the final train set for the performed gesture. Finally, once the train set is produced, we will apply it to a benchmark, state of the art neural network for hand gesture recognition and analyze the accuracy achieved and compare it to the results using a standard, real gesture datasets. In comparing the recognition accuracy of a real gesture training set versus our synthetic (virtual) dataset, we will recognize whether using a virtual training set is a viable and effective method to train a hand gesture recognition neural network. We plan to create two separate training sets (comprised of different gestures) to train against two separate neural networks for increased confidence of results.

2.3. Referenced Documents and Standards

[1] J. Nagi *et al*., "Max-pooling convolutional neural networks for vision-based hand gesture recognition," *2011 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, 2011, pp. 342-347, doi: 10.1109/ICSIPA.2011.6144164.

[2] J. Yu, M. Qin, and S. Zhou, “Dynamic gesture recognition based on 2D convolutional neural network and feature fusion,” *Sci Rep* 12, 4345, 2022. https://doi.org/10.1038/s41598-022-08133-z

[3] M. Asadi-Aghbolaghi *et al*., "A Survey on Deep Learning Based Approaches for Action and Gesture Recognition in Image Sequences," *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, 2017, pp. 476-483, doi: 10.1109/FG.2017.150.

[4] C. Correia de Amorim, C. Zanchettin, “ASL-Skeleton3D and ASL-Phono: Two Novel Datasets for the American Sign Language,” *arXiv:2201.02065 [cs.CV]*, 2022 https://doi.org/10.48550/arXiv.2201.02065

[5] Z. Helton. *VR Hand Tracking for Robotics Applications.* May 2022, Texas A&M Electrical and Computer Engineering 404 Course Archives.

3. Operating Concept

3.1. Scope

The virtual environment can be used to generate various virtual hand gesture training sets for gesture recognition neural networks. These training sets can be composed of images or videos of the hand gestures. The scope of our project is limited exclusively to hand gesture training sets. With some adjustments to the environment, creating training sets for other body parts or full body gestures is possible as well.

3.2. Operational Description and Constraints

To create a virtual training set, the user must make or find a dataset of gesture recordings and import it to the environment. The environment will animate tens to hundreds of diverse human models according to the imported gesture recordings and take images on each to build the training set for a gesture recognition neural network.

The virtual environment to create gesture recognition training sets uses the Unity game engine and MakeHuman software, both of which are open-source. Users can find gesture datasets online but creating a new gesture dataset would require sensor equipment. In our research, we are using the Leap Motion sensor and tracking software which provides the tracking scripts that allow us to directly map bone transformations to virtual human model hands. We will be recording the transformation information including the position, rotation, and scale of each hand/finger component and exporting the files as animation clips, which are then applied to other (imported) MakeHuman virtual models to be used in data collection. Users of our research can use the same equipment or any other tracking technology with the ability to record rigged component transformations.

3.3. System Description

The system to create virtual training sets is made up of multiple subsystems: Gesture Data Collection, Human Model Generation, Model Animation, and Data Capturing/Collection.

The first subsystem is the Gesture Data Collection subsystem. Bone structure data of the hand of one individual is recorded using the Leap Motion Controller to create a dataset of related hand gestures, such as sign language. Once the gesture animations are recorded and stored properly, they are ready to be used in our complete dataset generation virtual environment.

The Human Model Generation subsystem is where virtual human models are created in MakeHuman and exported as .fbx files to be imported into Unity for future use. This system works in tandem with the Model Animation subsystem that uses the gesture recordings/animation clips to animate the models.

Within the completed dataset generation scene of our virtual environment is the Data Capturing/Collection subsystem that takes images of each model that gets generated, compiling the data into a training set. The subsystem uses a script to take the images or videos at randomized angles to collect comprehensive data of each gesture to ensure the neural networks are trained to acceptable accuracies in recognition.

Shown below is a flowchart depicting the relation of each subsystem.

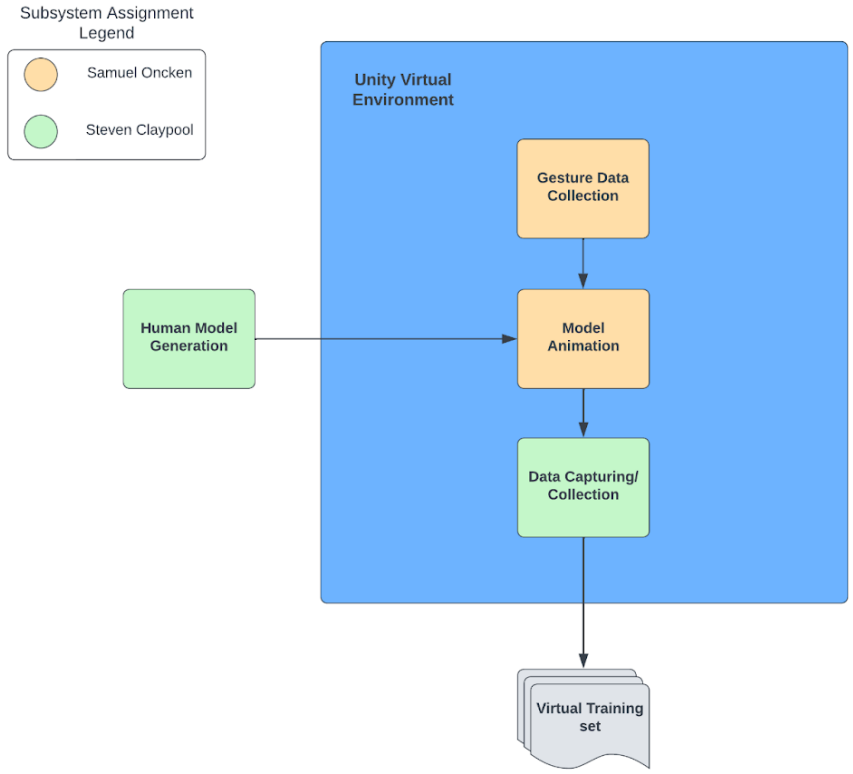

Figure 1: Flowchart of Virtual Hand Gesture Recognition Training Set Generation System

3.4. Modes of Operations

Since our project focuses on the validation of virtual datasets for hand gesture recognition neural networks rather than the creation of an operational tool, it does not necessarily have a mode of operation. However, the process of creating a train set using our virtual environment does contain stages: gesture recording and train set generation. During the gesture recording phase, a user must manually record his/her own hand gestures or find an online hand gesture dataset that they want to use. In the train set generation phase, a user will apply their hand gesture data to our virtual environment and the platform will automatically complete the train set using the imported gesture data.

3.5. Users

The users of our virtual hand gesture train set generation platform will be anyone who wants to build a dataset for their own applications. We will discuss a few scenarios in the coming section to exemplify this statement. Our environment as well as our tested virtual datasets will be available to download through GitHub. Once we validate through this research that using a virtual environment to create a train set results in as accurate (if not better) recognition than using real humans and equipment, users will be free to apply our platform to any human dataset that they wish to build or expand. They will be required to download Unity software

and have some knowledge working within Unity. Users will also be required to be able to create/find a gesture animation dataset to import into our platform as discussed previously, which might require additional software for their chosen sensor (i.e. Leap Motion Controller and Hand Tracking Software).

3.6. Support

Support will be in the form of a written manual which will detail the process of creating a training set using our virtual environment and the Leap Motion Controller. In addition, we will include our gesture animation clips (containing data of the hand gestures recorded with the Leap Motion Controller) and their respective virtual training sets, along with the gesture recognition neural networks we used. A document picturing types of gestures within the datasets will also be included. This can function as a benchmark for the users as they create their own training sets.

4. Scenario(s)

4.1. Software Developers Creating a Game

While creating a game that utilizes sensors such as a VR headset or Kinect, a developer wants to create a gesture dataset of movement controls for a computer vision model to recognize, but he/she does not have the time to find participants or equipment to record the gestures. Instead, the developer will use our virtual train set platform, where he/she will only need one camera to record each gesture. After recording the necessary hand gestures, the developer will apply them to our environment and be able to create an extensive train set with the reliability that a real, time-consuming train set would’ve produced.

4.2. Sign Language Translation

A user wants to create a dataset to train a model to recognize a distinct/less common form of sign language that does not have an existing dataset. He/she will use our virtual hand gesture train set generation platform to do this. First, the user must record the sign language gestures that he/she wants to implement, which requires only his/her movements (no need for large real human data collection). After importing the collected gesture data into our Unity virtual environment, the user will run our train set generation tool to output a large hand gesture dataset, composed of hundreds of unique human models and photographed from many distinct virtual camera angles. The user can then apply the virtual train set into their sign language recognition neural network.

5. Analysis

5.1. Summary of Proposed Improvements

When compared with the traditional method of collecting “real” gesture data, virtually creating a train set will have a variety of requirements/improvements including:

- Requires only one human to perform each gesture. We do not require paid participants, reducing cost and time for multiple parties.

- Utilizes virtual cameras within the Unity environment, allowing us to store hand gesture data from numerous angles while only recording from one stationary camera location in real life.
- Utilizes the bone/joint location detection within the Leap Motion Hand Tracking software, allowing us to map accurate gesture animations within Unity.
- Applies gesture animations to hundreds of randomly generated human models allowing for the neural network to train using comprehensive data of varying body structures, skin tones, etc.
- Human data recording is not always legal. At a minimum, human data collection requires some amount of paperwork for each participant. Using our research, we reduce legal risk.

5.2. Disadvantages and Limitations

Within our project, we can think of a few potential limitations that include:

- There is one person performing a gesture, then the recorded animation is applied to different human models. As a result, the recognition accuracy for a certain hand gesture might change depending on how a person may perform the gesture differently from others.
    - This can be mitigated by recording multiple versions of the same gesture to account for the randomness in the way a gesture can be performed. Another method is to use one or two more participants to perform each gesture. In both methods, a random version of the gesture recording would be applied to the virtual human model at run time.
- The Leap Motion Controller might not accurately depict the hand gesture being performed by a user into Unity. As a result, a user might have to record his/her gestures multiple times or from different angles to provide the best representation of the real gesture before proceeding to take images and videos for the train set.

5.3. Alternatives

Alternatives to our virtual hand gesture train set generation platform include:

- Traditional process as it stands now: creating a training set using “real” data. Different people perform a gesture while being recorded from either a mounted camera or a variety of cameras. Images and videos of the physical actions are then used to train a neural network. Time consuming and resource intensive.
- Different methods to create virtual datasets. We could use a different game engine software, human model generating software, and motion sensor/tracking software for recording gestures.

5.4. Impact

Our project has a significant influence on society in technological advancement and availability.

By validating the process of creating virtual training sets for neural networks in hand gesture recognition, the possibilities for future virtual training sets expands greatly. Developers could use our research to create diverse gesture recognition training sets much more easily than before. They could also use our research to create new virtual environments for different applications beyond hand gesture recognition.

Through our research, the availability of training set data would drastically increase. As more people begin to use our virtual environment to create training sets, more diverse training sets of certain gesture datasets will be readily available online without needing to find the money and participants to gather data of real individuals. This will also boost the technological advancement of machine learning and AI as more research with training neural networks can be done on much lower budgets.

In addition, our research has certain legal impacts as mentioned previously. Without the need for large amounts of human participants, the creation of a virtual human training set requires far less paperwork and legal permissions to use private data of the participants since only one human is required to perform each gesture that will eventually be applied to a large number of virtual humans.