Intro to Deep Learning *CSE591, Spring 2017*

Unfinalized: ASL Command Recognition using Deep Neural Networks

1 List of Group Members

1. Prajwal Paudyal

2 List of Abbreviations

1. ROS: Robot Operating System
2. RGB-D - Red Green Blue Depth Sensor
3. HCI - Human Computer Interaction

3 Background

I am working with two other people in this project who are not in the class. Their focus will be in the getting data, setting up, HCI aspects and the robotics aspect. I will focus on video processing and perhaps using CNN or RNN to detect simple visual commands issued using a Natural Language. I will finalize this soon. This document is being used as a project proposal for a robotics class as well which I am not a part of.

Visual story understanding is a field of computer vision that is gaining a lot of research. Understanding of visual gestures can be understood as a story understanding problem where the recognized gestures in the dialogue are directly used for tagging a certain video sequence. These tags can be then used for various purposes including human computer interaction. In this project we will focus on extracting commands from a video sequence consisting of a person issuing a command. [4]

3.1 State-of-the art

1. Speech-based Interfaces
   To combat social healthcare challenges due to ageing population, the proposed system [7] can detect a situation of distress when the user asks for help vocally. [5] use voice commands based on military jargon to position objects on digitized maps by creating labels and drawing symbols. Android Based Smart Home System with Control via Bluetooth and Internet Connectivity to control home appliances via voice commands [6].
2. Gesture based HCI interfaces
   GestureTek [1] product is used for applications like controlling pc, mobile or console applications using camera or phone. It is based on the technology of 3D camera for computer vision, camera in mobile device and pointing frame. HandGET [2] has the toolkit that facilities integration of hand gesture control with games and VE applications. EyeSights [3] has a product that uses the power of intuitive hand gesture to control the wide array of consumer electronics and digital devices.

3.2 Novelty

Use of Natural Language for Gesture Processing: The current systems are either based on speech as discussed in subsection Speech based interfaces or use a generic set of gestures for HCI as discussed in subsection

Gesture based HCI interfaces. However, there isnt much work towards an interface that allows the use of natural language gestures for HCI. For instance a deaf or hard of hearing person may wish to use her native language like ASL for interacting with computers. Inclusion of Natural Language in control interfaces allows a greater level of fluidity and higher expressibility. This comes at the price of newer challenges associated with Natural Language Processing and Sign Language Recognition.

4 Scope

4.1 Basic assumptions and constraints:

1. This project will utilize concepts from natural language understanding, dialogue systems, computer vision and robotics to facilitate a project where a mounted camera system is able to process simple commands issued using a sign language.
2. To keep the project within the scope of this class, we will utilize a subset of commands/ dialogues conceivable for a computer interface control. For example we can have gestures like open, close, pause, stop, music, file etc which can be combine to create commands to operate applications like music, text editor. There can be some commands which cant work on a certain operating systems.
3. Due to possible lack of availability of robots, we will focus on the computer vision and language understanding part as core of the project, and will bring in the robot as a tool for proper alignment of the mounted camera later.
4. The system will use a unix computer to facilitate usage of ROS, a kinect or a RGB-D camera for vision and the same computer or a server for processing (if need be)
5. There are no assumptions for the system to work on any gestures not trained on, or to be portable to other systems. Timing constraints, real time or not. There will be only one user at a time in the camera frame and there is no moving object in the background.

4.2 Functionalities:

4.2.1 Command System Interface (Scrum Lead - Shweta)

1. The system will have a list of ASL words that it understands..
2. The system will recognize any combination of these words as a command.
3. The system will determine if the identified command is valid based on the list of accepted functionalities for the various applications it supports. If the command is not valid, the system will suggest alternatives.
4. The system will then perform a command on an application if appropriate or only display the command on the screen if not. The system will also speak out the command.

4.2.2 Feature Extraction and Selection (Scrum Lead - Amsal)

1. The system will use a Kinect or RGBD camera to obtain color and depth information.
2. Features that are useful for recognition of gestures will be extracted and cleaned.
3. Pre-processing will be performed on these features to get them in the format that is required by the learning algorithms used in step 4.2.3.

4.2.3 Deep Learning for Training and Testing (Scrum Lead - Prajwal)

1. Features obtained from step 4.2.2 will be fed into a deep learning framework for training.
2. Training sessions will be performed for obtaining the data
3. Analysis will be done on gesture data available from elsewhere if they can be leveraged for training this system.
4. Matched output will be returned to the Command System Interface part for output to the user.

5 Other technical requirements

5.1 Usability

1. The user need not speak or type to the system as all commands will be taken by vision.
2. The user need not worry about the background color or lighting differences
3. The user can sit / stand in their regular positions while using a computer as long as they are completely in the frame of RGBD / Kinect Camera

5.2 Requirement on ease of use

5.2.1 Effective

1. The system will be effective and successful in recognizing and performing the commands that it recognizes.
2. The effectiveness of the system will be evaluated in terms of accuracy .

5.2.2 Efficient

1. Users have to be able to use the system in a fast and efficient manner.
2. This will be evaluated in terms of precision and recalll.
3. This will be evaluated also in terms of recognition speed.

5.2.3 Engaging

1. The system will be satisfying to use and all confusion should be minimized.
2. This will be evaluated by user feedback

5.2.4 Error Tolerant

1. There will be a mechanism for the system to recover from all foreseeable errors that can be identified.
2. This will be evaluated by testing the system by creating exception scenarios.

5.2.5 Easy to learn

1. The system should be intuitive and not overly complex.
2. This will be evaluated by user surveys.

5.3 Safety requirement

1. Does not start malicious scripts/programs.
2. The system will keep an ongoing list of acceptable programs
3. The system will not have sudo capabilities.
4. The system can be extended to have authentication (future work).

5.4 Scalability

1. The system will be trained incrementally thus it should be feasible to add more gestures
2. The machine learning algorithms will be trained using all the data available so the system should be able to accept input from multiple cameras.
3. The system should have a reasonable accuracy for inter-user recognition.

6 Project Plan

Table 1: Proposed Work Plan

Weeks	Milestones
0-4	We will finish Project Proposal and Scopes of Work
4-6	Camera calibrations, preprocessing of recorded video, gather data for training.
6-8	Extract features from the recorded data from camera.
8-10	Mid term presentation on March 14
10-12	Recognize the gestures and commands
12-14	Respond to the commands if valid, iteratively update the features and commands to improve performance.

References

[1] GestureTek. `http://www.gesturetek.com/`, 2008. [Online; accessed 8-February-2017].

[2] HandGKET. `https://sites.google.com/site/kinectapps/handgket/`, 2011. [Online; accessed 8-February-2017].

[3] Eyesight. `http://www.eyesight-tech.com/`, 2012. [Online; accessed 8-February-2017].

[4] Grazia Cicirelli, Carmela Attolico, Cataldo Guaragnella, and Tiziana D'Orazio. A kinect-based gesture recognition approach for a natural human robot interface. *International Journal of Advanced Robotic Systems*, 12(3):22, 2015.

[5] Philip R Cohen, Edward C Kaiser, M Cecelia Buchanan, Scott Lind, Michael J Corrigan, and R Matthews Wesson. Sketch-thru-plan: a multimodal interface for command and control. *Communications of the ACM*, 58(4):56–65, 2015.

[6] Shiu Kumar and Seong Ro Lee. Android based smart home system with control via bluetooth and internet connectivity. In *Consumer Electronics (ISCE 2014), The 18th IEEE International Symposium on*, pages 1–2. IEEE, 2014.

[7] Emanuele Principi, Stefano Squartini, Roberto Bonfigli, Giacomo Ferroni, and Francesco Piazza. An integrated system for voice command recognition and emergency detection based on audio signals. *Expert Systems with Applications*, 42(13):5668–5683, 2015.