**Neural Nets and 3D Mapping of Human Motion**

Team Members: Shahbaz Chaudhary, Emily Coppess, Jay Ong

**Executive Summary:**

For our capstone project, we will be building the basis of an eventual yoga training tool, which will provide a cheap and efficient way for users to have a personalized remote learning experience. This particular project will be focusing on building a system capable of identifying yoga poses from static images - with a focus on eventually connecting this identification system into a real-time feedback system.

**Background:**

Yoga is widely recognized as a convenient form of exercise that is effective in improving and maintaining our physical and mental well-being. According to a 2016 industry-wide study, there are an estimated 40 million people in the U.S. who are open to doing yoga for the first time in the next 12 months. So why hasn’t yoga gone mainstream resulting in all of us living healthier lives? This is due to the major barriers to entry that every aspiring yoga practitioner faces:

1. The lack of financial capacity to attend expensive yoga classes or to hire a personal yoga trainer.
2. The fear of being judged by others while performing the seemingly ‘awkward’ yoga positions for the first time.
3. The lack of a community of peers who share the same lifestyle and therefore challenges in practicing regularly.

While there have been products developed to try to tackle the issue of remote fitness learning (e.g. exercise video and wii fitness games), these systems either do not provide personalized feedback to assist the user or require the purchase of specific equipment. In our capstone project, we wish to apply recent advancements in the field of computer vision in recognizing human poses to provide a feedback system that is technologically agnostic. Current developments in the field of human-computer interaction and computer vision provide a basic framework for us to build such a tool to increase access to yoga (and related fitness programs) to a wider consumer base.

***Natural User Interfaces:*** The goal of a lot of developments in human-computer interaction is to make these interactions more intuitive and require less technical training for human users. In other words, developers want to build technology that adapts/learns human behavior rather than the other way around. Due to this goal, there have been several technological developments focused on creating “natural user interfaces” (NUI) – especially in the video/computer game industry. Some of these initial developments focused on a more tactile approach to building NUIs (e.g. Perceptive Pixel, Microsoft PixelSense, and 3D Immersive Touch). These interfaces rely on physical contact between the user and the computer interface. On the other hand, there have been several developments of NUIs which use motion-sensing instead of the tactile approach (e.g., Nintendo’s Wii Remote, Xbox Kinect). Recently, there has been a focus on improving the video/computer gaming experience by using a 3D mapping based on video input to incorporate the user’s motions into the action of the game (similar to the concept of virtual reality).

***Skeletal Mapping***: In recent decades, computer vision researchers have been working on models which can infer internal skeletal structure (i.e., discerning joint placement and angles to classify poses) from static and active sources (i.e., from pictures and video). Much of this research has been focused on classification tasks. The researchers are either trying to identify the pose or gesture of the subject (i.e., trying to identify what the subject is doing) or identifying a specific subject (e.g., combining facial recognition and gait analysis). Additionally, there has been a focus on the application of neural networks in creating maps of human poses (e.g., DeepPose project at Google).

These technological developments have a huge potential impact on increasing access to fitness and health resources - providing fitness access to people who do not have the financial capacity or the geographic proximity which would enable them to utilize resources such as physical trainers or physical therapists. 3D skeletal mapping has been utilized by some NUIs in the context of fitness programs enabling people exercising at home to determine if they are correctly following the exercise instructions at home in the absence of an in-person instructor. However, these applications will only function with brand-specific technology. The requirement of specific equipment drives up the cost of these video fitness programs. Additionally, these programs also only inform the user of how accurately they are matching the instructor. They do not provide a way for users to measure their progress towards a goal or when they are in danger of injury.

**Problem Statement:**

We would like to contribute to introducing yoga to millions of individuals and households. To solve the issue of lack of access to traditional training facilities, we would like to create a cheap and efficient way for anybody to have access to yoga training. Our solution will convert images to a data set of poses and those poses will be classified into various yoga poses. While our main focus is on assisting beginning yoga users, we believe that this tool could be eventually adapted to serve all levels of yoga users - not only making sure that they are performing poses correctly, but also help them set and realize personal fitness goals.

**Research Purpose:**

There are several key tasks that will precede a full working prototype of our vision:

1. Recognize 13[1] major joints on the human body from images: ankle, knee, hip, wrist, elbow, shoulder, neck
2. Use joint position and angle data to classify 9 poses from a common[2] yoga flow: down dog, three legged down dog, knee to nose, high lunge, warrior II, double blocks A, double blocks B, spearing warrior and reverse warrior

We ultimately aim to use the classification models built for (2) and the joint data to be able to build a model using a distance metric to determine where a person is deviating from the ideal pose (i.e. making a mistake). Our current models will be focused on static images, but we will eventually use our work on the static images to create a real time system.

**Research Questions:**

1. Are we able to extract pose information using existing methodologies?

1. Are we able to extract pose information in real-time, using existing methodologies?

1. How do we manage the element of time while aligning poses and while combining information sources?

1. Is the third dimension (depth) necessary for the classification task or the distance metric?

1. Is a second camera necessary or can we build an accurate and precise model with only one?

**Research Hypotheses:**

1. We predict that we will be able to extract pose information from an image without the addition of a second camera and that the depth dimension will not be critical for correct classification.

1. If successful with building our classification model, we expect to be able to adapt our model to detect these poses in real-time.

**Preliminary Data requirements:**

For the initial phase of our project, we will utilize the “MPII Human Pose” dataset from the Max Planck Institut für Informatik which contains twenty five thousand images containing over forty thousand people. These pictures capture over 410 human activities. This dataset is already being used to drive the latest research in the field. We will use this dataset to refine the pre-built model we pick from the latest research papers. Our initial attempt will be to use this model and test it against yoga specific videos from the internet.

Once we are satisfied with the quality of detection, we will record our selves performing yoga poses and evaluating the quality of pose detection. After we are satisfied with the quality of the model, we will introduce the set-up to a commercial yoga studio to get professional feedback.

**Preliminary Data collection techniques:**

While the previous section lays out the ideal case scenario, if our models are not accurate enough, we will have to collect some data ourselves. One method we propose is to generate human poses, in the form of virtual avatars, rendering them and using our models to detect poses. Since these poses will be computer generated in the first place, we will have absolutely perfect training set, against which we can train our models. We will be able to generate essentially unlimited number of poses and feed deep learning models (which are notorious for need large training observations)

**Definition of Variables:**

In order to build our representation of the human body, we will need information about the following variables:

1. Joint angles, for all major joints in the body (excluding small joints such as fingers)
2. Locations of all joints (from which lengths of limbs can be extracted)

These variables will serve as our initial variable for our pose classification model. Potentially, we might need to include variables related to the pose (such as pose type, standing versus sitting, prop involvement, etc). However, for the initial models, we will be focusing on joint information.

**Quantitative Methodology:**

Our problem can be divided into three main sections: skeletal mapping, joint feature engineering, and pose classification.

***Skeletal mapping:*** Almost all successful results in modern computer vision are being obtained using Deep Learning frameworks. Image data is best handled by a subcategory of deep learning, called Convolutional Neural Networks (CNNs). We expect that CNNs will suffice for converting images to skeletal pose information. By pose information, we specifically mean the coordinates of individual body parts in the image. Our deep learning model will be trained and tested (holdout validated) on data consisting of image-coordinate pairings. The datasets will contain a single person approximately 150 pixels in length annotated with coordinates of the following parts:

Ankles, Knees, Hips, Wrists, Elbows, Shoulders, Neck and Head.

Ultimately the main purpose of our CNN is to consume images and produce a file containing all the necessary features for further classification models to work.

***Joint feature engineering:*** We will convert the output of skeletal mappings into a vector containing joint angles by using the location of key body parts. For example, the locations of the right wrist, the right elbow, and the right shoulder would be used to calculate the angle of the right elbow (by creating a triangle). We decided to use the joint angle for pose detection because we feel it will deal with the issue of individual differences in things like arm length and height. Also, for yoga specifically, proper joint alignment is key, so angles would be the main indicator of problems in this area.

***Pose classification:*** Once a pose is extracted, we will need to run more established algorithms to detect and classify yoga poses. Since we wish to ultimately build a system capable of giving feedback, we want to focus on classification algorithms that are highly interpretable. We will begin by looking at decision trees, random forest, and logistic regression. Model selection will be based on primarily on accuracy, but we will also take into account other measures of fit and runtime (which will be critical to a real-time system). We suspect the best method to validate the performances of our classification models (and inform our final model) would be k-folds cross-validation as it reduces selection bias as well as the variance in prediction power of our final model due to overfitting. K-folds cross-validation also provides the full benefit of using all of our data.

Once classification is done, and if time and resources are available, we will use a Markov model to detect patterns in video sections, rather than just in static images.

**Expected Results:**

Based on our preliminary efforts to train neural nets using the DeepPose framework, we predict that the most efficient method will involve adding a yoga specific “calibration” layer to an existing pre-trained neural net. Our skeleton model will provide a 2D map of joint positions and joint angles. The joint angles calculated will not be the real angle of the joint because we will be missing the depth dimension. However, we suspect that the depth dimension will not end up being a crucial factor in our classification efforts.

We predict high classification accuracy using our 2D angles. Given our understanding of the computation problem our model is trying to solve, it is highly likely that a classification tree model will probably make the most sense. Its ease of interpretability makes it a highly attractive prospect. It is also possible that Markov models could maybe be utilized (making the transition to real time simpler).

Once we have a high rate of accuracy in classification, we think that we will be able to build a real time system. A real time system will probably involve the creation of Markov chains. We think that the classification model will serve as a good indicator when a specific state of the Markov chain has been achieved (i.e. pinpointing completed poses versus bodies still in transition between poses).

**PROPOSED SCHEDULE:**

**Implementation:**

Summer Quarter:

Week 1-4: Data acquisition, cleaning, and labeling

Week 5-6: Neural Net Training, Building Skeleton Vectors

Week 7: Calculating angles, determining if 2D approach to angles sufficient for classification

Week 8-9: Classification Models - week 8 focusing on model selection

Week 10: Identify metric to use for level of pose correctness

September: Building program to identify pose and level of pose correctness

**Writing:**

September: Literature Review, Revising/updating Project Proposal sections to be included in final report

Fall Quarter:

Week 1-2: Quantitative Methodology, Analysis of Results

Week 3-4: Conclusions, Recommendations

Week 5-6: Business implications, Future Directions (?)

Week 7: Revisions

Week 8: First Draft of Capstone Presentation

Week 9: Final Draft of Capstone Presentation

Week 10: Capstone Presentation

**References:**

[1] <https://annania.wordpress.com/2011/08/21/the-bodys-13-major-joints/>

[2] <http://templeturmeric.com/movement/primal-warrior-ii-yoga-flow/>

**Appendix:**

