Applying Computer Vision and Deep Learning to the Art of Boxing

Evan Huang
Cedric Fraces Gasmi
Stanford University
evhuang@stanford.edu
cfraces@stanford.edu

CS 230: Deep Learning, Winter 2018

Abstract - **This project focuses on the identification of boxing poses and the recognition of styles of famous boxers. We use a series of deep learning methods to identify, classify and recognize different boxing movements. We use transfer learning at various levels to encode human body postures using a pose extraction approach. We then build a labeled dataset by gradually applying classifying neural networks to series of image frames and videos borrowed from boxing and non-boxing footages. We apply unsupervised learning methods to cluster boxers based on their styles. We used the trained algorithms to refine our classification algorithm to be able to utilize larger and more generic datasets. The classification was tested on two of the famous boxers of the 20th century. This tool can be used to help boxers around the world assess their performances and track their progress.**

Index Terms – Sports, Boxing, Computer Vision, Style, Open Pose

INTRODUCTION

AI is radically changing the world of sports. From analyzing advanced statistics on team sports to devising better tactics on teams' positions through deep imitation learning. At the individual level, athletes can benefit from the progresses of AI too. Personal assistants already exist for sports such as Golf, Tennis and Baseball. They help correct and calibrate movements to improve technique. We propose the use of computer vision in boxing to identify styles, areas of weakness, and useful metrics for boxers of interest. With successful identification of these factors, we can eventually create a tool that would act as a personal assistant to help improve people's technique. This tool would furthermore create a competitive advantage, and would be incredibly useful to keep or gain standing in the boxing world.

RELATED WORK

The sports industry has previously taken advantage of deep learning for additional analysis and statistics. Analysts working in baseball have already started implemented deep learning models on the plethora of advanced statistics to improve team compositions and strategies. In soccer, analysts have found some success using LSTM networks for ghosting analysis. And various applications and startups have looked at the use of computer vision for fixing stance, posture, and technique. This includes companies like Swingbyte, a golf assistant that tracks the path of a golf swing to provide advanced analytics for the user.

The related work is therefore vast, but a lack of accessible data has hindered the development of complex deep learning in sports. Boxing in particular has yet to indulge in the deep learning craze, creating an opening for novel and impactful analysis.

DATASET AND FEATURES

In this problem, we are interested in capturing the styles of boxers. An important first step in the accomplishment of this task is to recognize boxers from non-boxers based on the way they move and position themselves. This method of recognizing boxers is nontrivial, as correct identification could lead to automated scorekeeping, automated penalties, video tracking of boxers, and the generation of statistics. In order to better capture the information related to the pose of a person, we chose a form of encoding called Human Motion Capture (HMC) and more specifically Human Body Key Points Detection (HBKPD). Two main research groups have published extended work and released libraries that allow to capture body parts from a picture or a video. The Open Pose project at Carnegie Mellon University (CMU) Perceptual Computing Lab and Detectron at Facebook AI. The later seems more geared towards whole body segmentation. We benchmarked the two libraries and finally opted for OpenPose. Detectron would provide good information on the distance between boxers and general attack/defense sequences. Since we are more interested in style than strategy, OpenPose offers more granularity on the movements of the individuals (body parts).

The work started with a fairly small dataset. 26,000 pictures sampled from boxing (6,000) and non-boxing (20,000) datasets were collected from the ImageNet database. Non-boxing pictures were taken from other sports such as baseball, basketball, cross country training, gymnastics, soccer, tennis, wrestling, and regular activities (walking, sitting, standing …). The goal was to rapidly evaluate whether it was feasible to recognize boxing from non-boxing poses and set a benchmark for the rest of the study. Data seems available in abundance (Imagenet, CMU database, Youtube) but the amount of labelled data we could get within the timeframe of this project was quite limited. We also reached out to various parties susceptible to provide more data. Local and foreign boxing gyms, CMU for their motion capture database, professor Jeff Marsh at University of Colorado Colorado Springs (UCCS) shared some of his insights on HMC. The boxing sample were taken from fighting and training, amateur and professional photos. Each picture was analyzed using the OpenPose open source library. The code was modified to run on batch of images and output 3D pose representations of humans in a csv format. Figure 1 shows such a capture for two of the examples cited above.

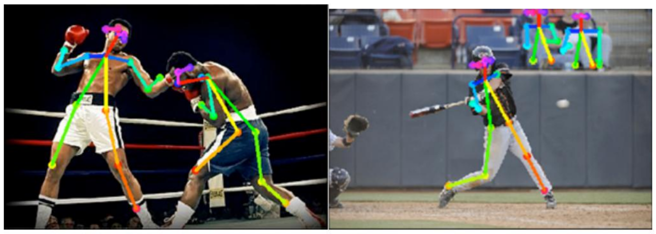

Figure 1: Two captures of human bodies using Open Pose. On the left a capture of a famous professional boxing match (Ali vs Frazier), on the right a professional baseball player

On these frames, we can see that the captures present some challenges. The right arm and left leg of Frazier are not captured. Neither is the bottom right leg of Ali. The baseball shot shows the capture of two spectators sitting in the tribune. These issues can contribute to make the data noisier and mislead the classifier. These challenges in collecting clean data lead us to consider a "simpler" problem at first. The ability to recognize a boxer vs a non-boxer on static frames.

Each skeleton is composed of 17 key points placed at the main joints of the body: (Pelvis, Right Hip, Right Leg, Right Foot, Left Hip, Left Leg, Left Foot, Belly, Thorax, Neck, Head, Left Shoulder, Left Arm, Left Hand, Right Shoulder, Right Arm, Right Hand). We decided to use 3D human body captures so each point has three coordinates (x,y,z). Each body is now encoded into a 51 features vector that we can use for our classification problem. Figure 2 shows the poses captured for a typical boxer and non-boxer pose.

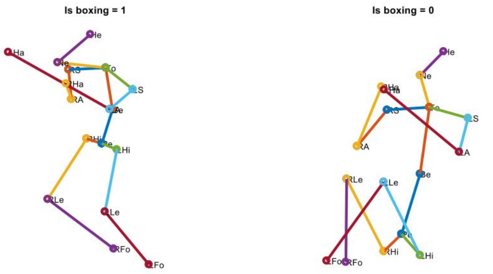

Figure 2: 3D capture of key body points for a typical boxing pose (left) and non-boxing pose (right)

The two main challenges to address in getting a clean labeled dataset were the following:

- Eliminate poses that did not correspond to boxers
- Recognize the right boxer from a frame generally showing two of them fighting

We addressed the first challenge by using a combination of heuristic rules and the classification model we had trained with a clean set of photos and videos. The heuristic rules are a set of hard coded conditions that allowed us to discard skeletons that were obviously doing something else (seating in a tribune) or the 'grotesque poses' meaning the poses that could not have been human and were often the effects of perspective rendering from the Openpose library (abnormally large limbs or necks, dislocated spines,…). For the later, we used the rules of divine proportions established by Da Vinci in its Vitruvian Man[*]

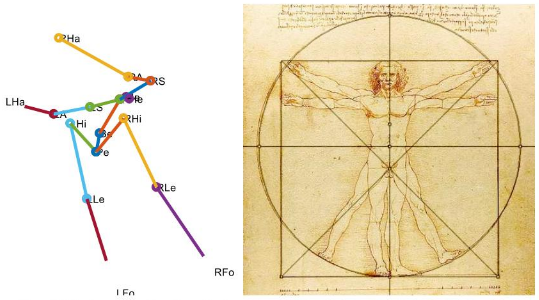

Figure 3: Distorted skeleton with head rolled under arm pit on the left. Vitruvian man embodying divine proportions on the right

The second challenge was less obvious given how boxers change positions and angles within video. This with the large number of failed extractions make the designation and tracking of a skeleton impractical. We decided to stack various models in order to accurately label boxers before training on style.

METHODS

Given our choice to operate from key points data, we would need to label different boxers for style classification. Additionally, we need a model which identifies boxer poses. That is, a classifier trained on clean boxing data to label the boxing vs non-boxing sequence in a video would help separate external poses in boxing pictures, as shown in Figure 1. The justification for a two-step approach is that the majority of the data is "polluted" by these artefacts and we need to be able to confidently filter out the noise before we can train a robust algorithm for classifying boxing styles. We will look at different methods for each.

Labeling Boxer Identities

I. Face Recognition

Our initial idea to identify specific boxers included the use of a face recognition package. We wanted to analyze face recognition results to form a baseline for our style identifier. Face recognition was to be used to anchor the skeletons to a given boxer hence giving us a labeled dataset for training a style classifier.

We used a face recognition API by Adam Geitgey which included functions for face recognition, locating faces, and drawing bounding boxes for identified faces. The API was built using dlib's face recognition tool. Usage of the face recognition gave us some correct labels on videos and images, but problems occurred. Quality on video was low in many frames, and face locating failed in cases where a boxers head was turned. Another issue was the identification of referee, and audience members, of which we are not interested in. Though the recognition algorithm would not consider these individuals, the additional computation time with these faces got lengthy and made the process impractical. Nonetheless, the facial recognition package worked wonders when boxers were directly facing the camera, and had success on boxers such as Mayweather, Ali, Tyson, and Mcgregor. However, the facial recognition package proved to not be accurate enough for consistent labeling, even when anchoring poses to location of the last recognized face. There were simply not enough accurate recognitions at a fast-enough rate.

The face recognition issue brings to light the difficulty in identifying individuals. When the face is not clearly visible, the body can additionally be taken into account. However, low accuracy when looking at just body by humans makes this process difficult. This process would also require a large amount of data to train, of which we did not have. Thus, a different approach was likely needed and chosen.

Transfer Learning & Model Design

One concept taken advantage of multiple times in this project was transfer learning. In order to get pose data, we used the OpenPose package. The model itself is a deep learning algorithm that converts the initial data type of images into poses. Thus our classifier is effectively adding on extra layers to output boxer or non-boxer.

Given that we are classifying pose data, we looked at the architecture of other OpenPose classifiers. These gave us a baseline for some parameters of the model. One such model is pose detection by Srini Aananthakrishnan which used an Adam optimizer, dropout = 0.1, and a relu activation function. From there on we adjusted the model to fit our specific application.

The ultimate intended model design is a combination of separate models. Such a design consists of linking the OpenPose results to a classifier model (boxing vs. nonboxing), then taking those results with the boxer identifier to train a style predictor. That is, when we create the style model, we will need to feed in both the boxer identifier and boxing classifier to generate usable data for training.

Distinguishing Boxers from Non-Boxers

I. Neural Network Architecture

We train a neural network on boxing and non-boxing poses in pictures from Imagenet and videos from boxers training by themselves. We used a 6 layers dense network with decreasing number of hidden units. We used ReLu activation functions on all layers but the last (where a sigmoid logistic regression was used). This method worked and we reached 89% accuracy in classification with a raising trend as shown in Figure 4.

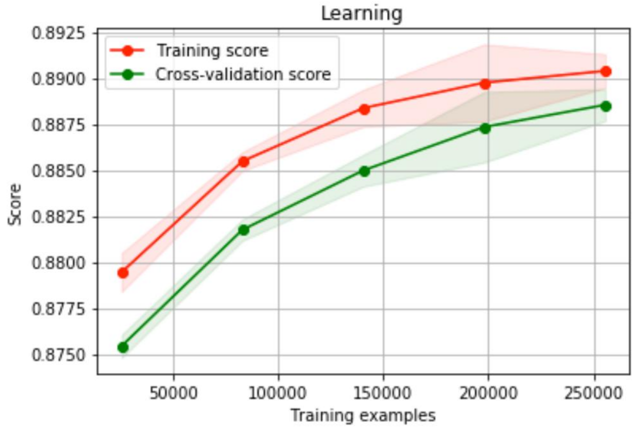

Figure 4: Training and testing accuracy vs number of training examples for boxing vs non-boxing classification problem

Considering the limited access to clean data, we employed data augmentation techniques to improve our model's performance. We added random noise to the existing data and ran the same training. The results showed a substantial improvement. By tripling the size of the training dataset, we were able to reach a 96% testing accuracy as seen in Figure 5

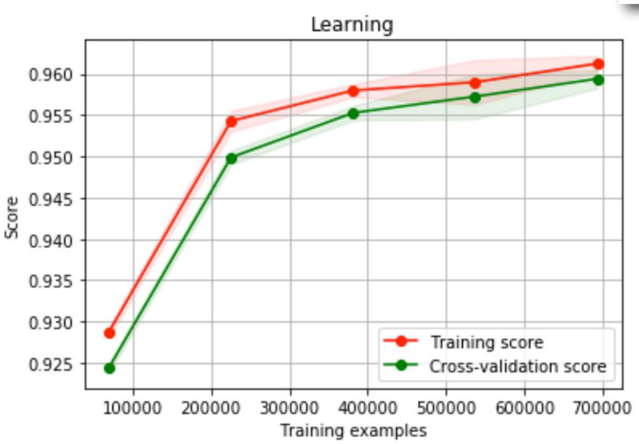

Figure 5: Training and testing accuracy vs number of training examples for boxing vs non-boxing classification problem with data augmentation

This result shows that getting clean labelled data is an essential step of the project. Because the actual number of valid boxing poses runs scarce after we have processed the few clean videos we were able to collect through Youtube, it becomes important to have a good classifier and utilize the large amount of videos (of not so good quality) available. We then employ this model to filter the boxing poses in the larger boxing movies dataset we processed using OpenPose. We used an AWS instance with GPU's to accelerate the capture process. The speed-up we could observe was a 5-10 folds depending on the quality of videos. We let the pose capture run for a full week. This allowed us to process a few hours of videos featuring Muhammad Ali and Mike Tyson, two of the most emblematic boxers of the century with very distinct styles.

We implemented our models using both sklearn's neural network tool and tensorflow in python. Though tensorflow offered more flexibility, we found that results were similar. We initially started at a large number of 57 layers with 50 neurons each to match other openpose classifiers detailed in online blog posts. However, the results were overfit on noise. Models done by others used much cleaner datasets (such as sitting vs. standing), and consisted of many additional nodes in the openpose model such as facial features and fingers. Adding regularization helped a bit, but ultimately using a less complex model for this low dimensional data proved most effective.

II. Unsupervised Learning/Clustering

One of the main challenges in training an algorithm to recognize boxers is that the majority of boxing footages feature two boxers. If we want to be able to use this vast source of information, we need to develop a method to recognize the boxer from its poses. This task is not quite as complex as the original one as we know that one of the boxers is the boxer of interest. We "just" need to find out which one it is. One approach is to use unsupervised learning to cluster poses on a large dataset (a series of fights between Ali and other boxers) where we know that most of the boxing poses captured will be ones of Ali. We use the model trained at the previous step to filter out all poses that do not represent a clean boxer stance. We use a combination of K-means and Mixture of Gaussians algorithms to cluster the various poses and a PCA transform to represent them. We separate the dataset in three clusters. We note that the k-means algorithm could segregate clusters that are spatially distinct but not latent distributions. In order to do so, we combined the k-means clustering with a mixture of Gaussian (EM algorithm). Results are shown in Figure 6.

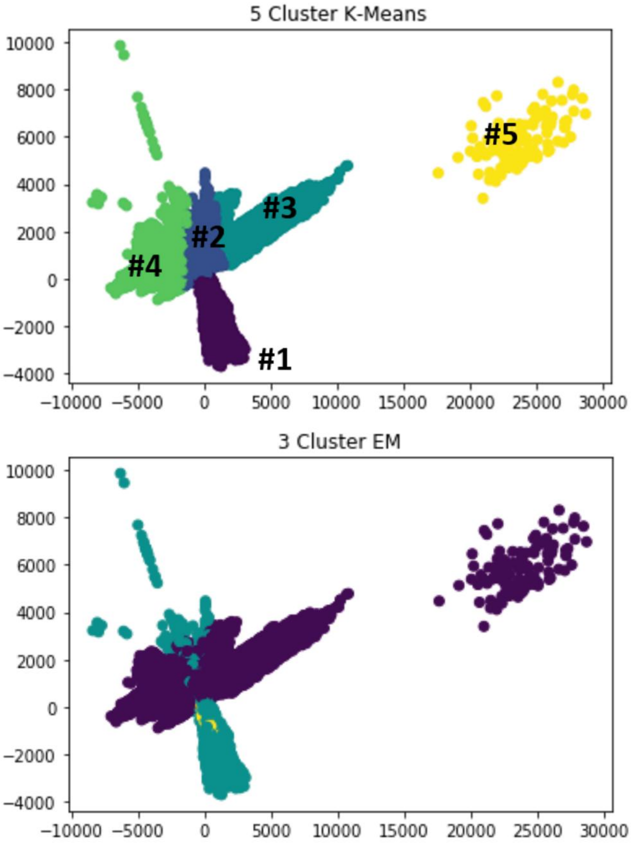

Figure 6: Clustering of Ali's poses using k-means (top) and Mixture of Gaussians (bottom) algorithms

This step allows to increase our predictions of a pose belonging to a certain boxer. In Figure 6 for example. The dominant trend could be representing the poses of Ali. We do not have a systematic way of measuring the performance of this but checked some of them manually. We present some of them in Figure 7

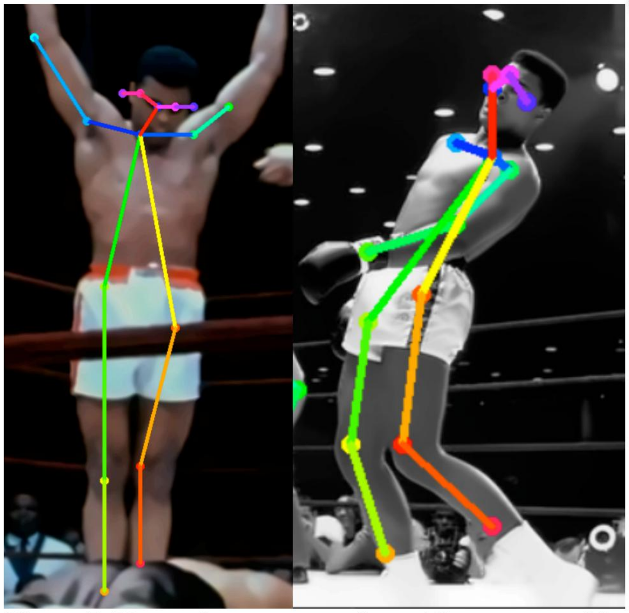

Figure 7: Two reference poses from Ali represented in Cluster#5 (left) and Cluster#3 (right)

As an example, cluster #5 in Figure 6 is mainly composed of poses that have their hands in the air. This move is typical of Ali when he wins or teases his opponent. On another hand, cluster #3 mostly shows poses which are leaning back. This too is a signature move of Ali and something prescribed in all boxing schools as it opens the boxer's guard to body shots…

RESULTS & DISCUSSION

Model Results

In order to build a robust framework around the problem of style recognition, we adopted a combination of approaches. The first challenge was to distinguish human vs non-human poses. For this, a heuristic approach gave the best results. The baseline against which we compare is a random guess. The second problem was to compare poses of boxers vs non boxers. For this task, we trained a 6 layers dense network on videos of boxers training by themselves in a gym. We then tested the algorithm on a set of poses including general footage with boxing and non-boxing (spectators, referee) poses along with other random poses (other sports, original image set). The accuracy of this model exceeds 96%. The last step was to identify the boxers based on their styles. For this task, we picked 2 boxers and our baseline was the ability to recognize who the pose belongs to based on its movements. Our accuracy was of 75%. We estimate that an average human does fairly well based on full footage but remarkably poorly (slightly above average) based on poses only.

Application

Given our resulting model, we could experiment and create a live webcam boxer classifier. This determines if a person in the frame is currently boxing or not. We can imagine several applications of this. One of them would be to recognize if a person in front of you has a hostile pose and warn you… Other uses are more related to the recording of statistics or positions (how many punches were thrown, what type…). By generating quantitative data and metrics on the moves, we can start measuring and tracking progress. This would be very useful for personal assistants and coaches who work with boxers. Finally, there is also potential use in professional boxing to assist referees and judges and help reduce the error rate and misjudgment that are inherent to human assessment.

Next Steps

Our next steps are to improve the style classifier using deep networks. Though we had some success, the deep learning models we tried to apply did not work out for styles. We simply did not have accurately labeled data that could be trusted for the implementation. However, we plan to look for solutions to this problem to eventually create style classifiers using RNN on series of frames. We eventually want to partner with boxers and coaches to actively create a useful tool through continued testing.

CONCLUSIONS

This work focused on the classification of boxers based on patterns of movements or poses. We experimented various methods and addressed the problem step by step. We first attempted to recognize human then boxers vs non boxers poses. We then classified various poses based on the classes of activities they represented (offense, defense,…). We finally were able to identify clusters of moves that were representative of two emblematic boxers (Muhammad Ali and Mike Tyson). The main goal of this work was to define an encoding of style base on visual attributes (no sensors other than a camera). We built a low dimension parametric representation of a human body based on the positions of its key connecting points. We were able to extract the characteristic stances of a boxer and even some signature moves that can be associated with famous professional boxers. This approach provides a consistent way of comparing boxers. It sets the path for a quantitative assessment of technical skills. We started with a simple problem and progressively expanded. We are confident that this work can be pursued and expanded to further improve technique evaluation and correction.