Image Clustering Over Time

Dylan Lee, Zach McMichael, Viktor Zenkov, Adam Tutko

Abstract— **Image analysis has become a large part of deep learning and machine learning research. In this work we will apply some of these deep learning methods to generate image representations which will then be clustered by some generic clustering algorithms, such as K-Means. The goal of this will be creating clear and concise clusters that track the same feature in an image, even when that feature shows change over time. With successful results this will further the work currently being done on human decomposition images and show that these algorithms can be generically applied to any similar data set or domain.**

I. Objective

The goal of this project is to apply deep learning and clustering algorithms to create groupings of images that track the same feature over time. One of the main objectives related to creating these clusters and analysis will be cultivating or discovering an appropriate image data set that tracks a feature through time, and preferably is labeled and contains some type of metadata. Once a data set is found, then the goal will be applying the already developed deep learning model and clustering methods to determine how successful and accurate these methods can cluster other image sets from different domains.

This project will further develop the deep learning model used to generate image representations and apply similar methods as used in the ICPUTRD research to take advantage of metadata to improve cluster accuracy. This includes using the dates of when the images were taken with the goal of merging clusters between days to improve on feature tracking over time. Overall, this project will dip into all the focus areas of the class with a direct focus on data discovery and analysis.

II. Motivation

The motivation behind this project is related to work currently being done on an image data set from the body farm that tracks human decomposition (ICPUTRD). While the results from this work have shown promise, this project will test the algorithms and methods to see how well they work in other domains on different data sets. This project was also chosen due to the interest in deep learning and CNN's currently being used to create the image representation. Another motivating factor is the challenge of finding an appropriate data set for the project. While there are many labeled data sets available, finding one that has images of the same feature, where that feature shows changes over time, is a much more difficult task. The final motivation of the team was to choose a project in which our work will benefit and enhance other research being done in this field. Since a team member is already involved in research on this topic, there is a natural easy overlap where this project will hopefully benefit the research and vice-versa.

III. Data

A very important part of this project is data collection. Since the algorithms that have already been developed and tested are centered around a specific data set, it is a difficult challenge to find a large collection of images that also track a feature that changes over time. Some early thoughts of the team were finding image sets that track different types of mushrooms or plants growing, or even using a data set from NASA tracking constellation and planet movement. However, the most promising data set that has been found so far is an image data set named MORPH that is a collection of "mug-shot" photos that track different subjects over time. This image data set contains thousands of different individuals with multiple photos ranging from a few days to a few years apart. This provides a similar challenge to the body farm images, where the goal is to track features that changes over time; the face is the feature and the effects of aging over time take the place of decomposition. This data set also provides a good amount of metadata such as age, race, and date that can be used to improve the clustering.

The downside to this data set is it is not open source, and is only available to license for commercial use or purchase for academic and research use. Another issue is that this image set only shows one feature, in this case, a person's face. While this is not as unique and vast a data set as the ICPUTRD images, it still provides the best progression of a feature over time that the team has been able to find. To obtain access to this data set it must be purchased for $99.99 from The University of North Carolina Wilmington. Once obtained the team will need to extract the image metadata and create a database of the image paths that can be fed into the deep learning models to generate the representations.

IV. Team

The team will be made up of four senior Computer Science majors that all have experience in machine learning but little experience in image analysis and CNN's. Zach McMichael has worked with many machine learning libraries while doing text analysis on tweets and is experienced in C++, Python, and Java. Dylan Lee has worked on the ICPUTRD research team and gotten experience with using the deep learning models and modifying clustering algorithms to take advantage of metadata, he is experienced with C++, Python, and Java. Adam Tutko has experience with text analysis on open source git repositories, and has worked in C++, Python, and Java. Viktor Zenkov has experience with machine learning for text analysis and regression and has experience with C++ and Python.

Name	Role
Dylan Lee	Project Lead
Zach McMichael	Machine Learning Expert
Viktor Zenkov	Model Validation / Documentation
Adam Tutko	Data Set Augmentation / Documentation

V. Timeline

The proposed timeline for the project:

Week of	Milestone
9/30	Find or cultivate a data set
10/07	Clean and prepare image metadata for use in clustering
10/14	Generate image representations using deep learning
10/21	Using image metadata modify the clustering methods to improve cluster accuracy
10/28	Evaluate the clusters and continue to modify the algorithms
11/4	Based on results, either continue modification on algorithm or begin results analysis
11/11	Generate results analysis graph and prepare presentation of results

VI. Expected Results

The expected results for this project should match the results that have been seen from work on the ICPUTRD image set. This means creating clusters that track a feature that changes over time. The feature to track in this project would be the faces in the mug-shots and the changes would be the results of aging. This is similar to the ICPUTRD data set, with the difference being that the MORPH data set has less extreme time distortion. Since the MORPH data set that will be used for this project is more controlled and has less variables than the ICPUTRD dataset we would expect better cluster accuracy. These results would provide validity to the ICPUTRD research and show that the clustering algorithms used are applicable to other domains. Other expected results would include improving the existing algorithms ability to recognize and cluster faces of the same individual over time.