Age Estimation and Gender Classification using various Image Processing APIs

David Tang*, Waley Chen*, Karmveer Sidhu*
*Department of Electrical and Computer Engineering, University of Calgary, Canada
{david.tang, wachen, ksidhu} @ucalgary.ca

***Abstract*—Advancements in machine learning, particularly deep learning has paved the way for widespread adoption of facial recognition technology. However, it is still difficult to estimate age and gender from images due to high variance in facial features and image quality. The increased performance in predicting age and gender in images is due to leaps in deep learning methods using Convolutional Neural Networks (CNNs). Large technology companies such as Google and Amazon have released their own image processing tools that employ pre-trained deep learning models. This paper will utilize a large labeled data set of facial images scraped from IMDb to determine the effectiveness of Google's Cloud Vision (GCV) and Amazon Rekognition (AR) APIs. Specifically, it aims to determine the accuracy of predicting age and gender while using these APIs. Furthermore, this paper will assess how GCV analyzes a different data set of images and the effect of image quality. A model will be created using Tensorflow and the Keras API in Python to compare with the GCV results. We expect GCV and AR to perform well at age and gender prediction as they are leaders in image processing.**

***Index Terms*—Gender classification, age estimation, Google's Cloud Vision, machine learning, Amazon Rekognition**

I. Introduction

Facial recognition and analysis are some of the most important topics in computer vision. Automatic analysis of facial images have become increasingly important and are being used in areas like social media, video surveillance, business intelligence, and security. However, age estimation and gender recognition are challenging problems within the field of computer vision as they rely on interpretation of facial features [1]. Many facial images exhibit high variance even within the same age range. These variances include personal traits of each individual, nature of aging process, different race and genders, different pose variations, bad illuminations, effect of cosmetics, haircuts, and image quality [2]. Even humans can only give a rough estimation of someone's age by looking at their face. Due to these image variations, is it hard to gather complete and sufficient data to train accurate models.

Within the last few years, Convolutional Neural Networks (CNNs) within the field of deep learning have proven to be effective for computer vision. CNNs are widely used for both gender recognition, age estimation and have significantly improved object detection and recognition, and image captioning [3]. These CNNs can be trained through the use of Tensorflow and the Keras API in Python.

Google and Amazon also offer image analysis through their own proprietary APIs: Google's Cloud Vision and Amazon's Rekognition. Both of these APIs are pre-trained and offer label detection and image attributes [4], [5]. In this paper, we will conduct an analysis on these APIs to determine the accuracy of these models for gender recognition and age estimation. The main contributions of this work will be as follows:

- Use GCV and AR APIs for age and gender prediction and determine the effect of image quality on these models.
- Compare the performance of these APIs to determine which model identified features more accurately.
- Determine how well Google's AutoML Vision model classifies data from another data set.
- Analyze the effect of human bias and whether it results in a poorly trained GCV model.

II. Related Work

Classic age estimation methods involve a two-stage pipeline: local binary patterns are determined for feature extraction and then machine learning algorithms like Support Vector Machine are used to predict the age [6]. Early studies found that age estimation can be substantially improved by incorporating other facial traits like gender and race information; furthermore, the age estimation error can be reduced by 20% if trained separately on male and female [2].

Currently, most image recognition models rely on deep learning technology through the use of CNNs. Neural networks use algorithms that are layered next to each other and makes each algorithm contingent on the outcome of other surrounding algorithms [7]. A neural network uses convolution by merging multiple sets of information, pooling them together to create an accurate representation of an image. After pooling, the neural network uses the data to make predicts about the image [7]. CNNs incorporate the aforementioned two-stage pipeline into one step: the network extracts the best features and classifies these features into age categories or performs age regression.

The ChaLearn Looking at People Workshop in 2015 challenged teams to perform age estimation on a large data set [1]. The organizers of ChaLearn provided one of the largest data sets known to date of images with age and gender annotations. Of the top ten teams, nine used CNN models for age estimation. Rothe et al. employed a CNN network that uses VGG-16 architecture (16 layered model created by Visual Geometry Group) and pre-trained on ImageNet for image classification [8].

VGG models have been further improved since 2015. Xing et al. incorporated a newly proposed deep multi-task learning architecture which resulted in a high-accuracy race and gender classification. Rodriguez et al. proposed a novel feed-forward attention mechanism that is able to discover the most information and reliable parts of a given face for improving age and gender classification [6].

GCV and AR also employ CNNs using their own deep learning algorithms. Disney uses the GCV AutoML technology to build vision models to annotate products with Disney characters, product categories, and colors [4]. Motorola Solutions uses the video analytic features of AR to help with search of historical and real time video for persons-of-interest [5].

III. Methodology

A. Research Questions

This research aims at addressing the following Research Questions (RQs).

RQ1– *Can image quality affect the accuracy of GCV when predicting age and gender?*
This question aims to explore how much image quality can be lost before GCV API fails to predict the correct age and gender from the celebrity's face. Celebrity faces were chosen due to the abundance of publicly available information on them.

RQ2– *Are there differences in predictions when using Amazon vs Google's image recognition API?*
Image recognition models are challenging to optimize and will most likely have their own individual strengths and weaknesses. With this question, AR and GCV are compared to see which features are better identified in each model.

RQ3– *How well does a trained AutoML Vision model classify data from another data set?*
GCV relies on their repository of images to aid in classifying the images. Celebrity faces are over saturated and it is highly likely that their model is well tuned to detecting celebrity faces. By feeding a data set containing faces scarcely present on their database, it may lead to a loss of accuracy for certain features.

RQ4– *Can human bias result in a poorly trained model vs GCV?*
When labeling these images with emotional expressions, bias can be easily introduced because labelling the data is highly dependent on the on the label writer. The resulting mislabeled images can introduce bias in the machine learning model

B. Data Collection and Preparation

For this analysis, a large data set of facial images are used that have been scraped from IMDb for another study and have been since made available [8]. The data set contains 461,871 images with the following provided information:

- **dob**: Date of Birth.
- **photo_taken**: Date of Photo Taken. Used to calculate the subjects age when pictured.
- **full_path**: Path to File.
- **gender**: Gender of Subject. 0 for female, 1 for male, *NaN* for unknown.
- **name**: Name of Subject.
- **face_location**: Location of Face within Image.
- **face_score**: Score of Detector from DEX project [8]. *Inf* for no found face
- **second_face_score**: Detector Score for Secondary Face. *NaN* if only single face present.
- **celeb_names**: Subject Name.
- **celeb_id**: Subject ID.

To prepare the data for testing, approximately 600-900 images will be manually labelled with additional information such as eye color, skin tone, emotion (happy, content, sad). Since the data is comprised of celebrity images, it is expected that majority will be happy. To create an even distribution, the data will be manually crawled and labelled. The external data-set will be taken from personal libraries and from ones where permission has been received. For scaling down the images for the first research question, Apache Spark will be utilized, being run on the University of Calgary ARC cluster. Additionally, due to the larger size of image files, they will be modified and only the faces will be extracted. This will reduce needed disk space.

C. Data Analysis

The analysis of the collected IMDb data will be done using GCV API and AR. AutoML Vision from Google will be used for supervised machine learning where it will be trained to predict other labels not provided by Cloud Vision. A portion of the collected IMDb images will be manually labelled with some of the following labels: happy, sad, or angry. Our supervised model will be trained using this data before being applied to new unlabelled images of both celebrities and normal people. GCV and Amazon's Rekognition will be used to obtain labels which can then be compared to see which one is more accurate. Finally images will be down-scaled to 50% and 25% size and then reanalyzed with GCV. The results will then be compared to the results of the image at their initial resolutions.

IV. Expected Results

RQ1– Even though GCV API is a powerful tool for extracting data from images, it is still dependent on the image for what information can be extracted. If the images have very high resolution, a 75% reduction in resolution may not affect the output. However, if the image resolution was originally low to begin with, any reduction will result in loss of data. Depending on how the image was cropped, the celebrities' face's original resolution would determine if any reduction would result in different predictions. If the resolution becomes too low, details such as wrinkles and hair can blur leading to inaccuracies when predicting age and gender.

RQ2– Amazon & Google are premier machine learning companies so it is expected that they both excel at classifying majority of the features. Both models are expected to score highly but there might be some features that one model will perform better at. Overall, GCV might be the more advanced model because they have had a significantly longer time to grow their training database.

RQ3– It is expected that the drop in accuracy will be minimal as the model is likely to be highly tuned to many other features. Areas where the model might be significantly weaker are cases where the subject is not focused properly. IMDb pictures are generally well focused, with proper lighting, and by a professional photographer.

RQ4– Human bias can be introduced when manually labeling the images since studying human behaviour and emotions is an imperfect science. The professionally trained model by GCV could have involved specialists that were able to determine all present emotions on a human face. Majority of the images seems to be easy to classify since the celebrities are mostly smiling therefore they are happy. Our model results should not differ too much from Google results since we should be able to identify most emotions in the picture.

V. Discussion

A. Key Findings

RQ1– The IMDb database contains a wide range of celebrities which include the young and old. The labels are already provided so the model only needs to be trained and then tested for each image at reduced resolutions. The % reduction in resolution that leads to a significant amount of inaccurate predictions will be the desired results.

RQ2– The information trying to be extracted are the strengths & weaknesses of AR and GCV as well as the difference between the accuracy of the two. A large difference will allow a claim of which features each model specializes at. Features with an insignificant difference will be neutral between the two models.

RQ3– The difference in prediction accuracy with the celebrity only trained model for predicting non-celebrity images to celebrity images. Images will a significant loss of accuracy will be analyzed to determine what might have caused such as result. This could include being out of focus, covered face, and poor lighting or image quality.

RQ4– The model trained using manually generated labels will be tested on new data. The same data will also be used to test GCV then the results from both tests will be compared for differences. Any statistically significant differences in the results would mean that the manually trained model could contain bias.

B. Limitations

There are multiple limitations present for this study. The amount of images that will be processed are limited due to the financial restrictions of using both GCV and AR. Both services only provide 1,000 images to be processed before charging for each subsequent 1,000. As such, only enough testing will be performed to reach conclusive result. Secondly, adding labels manually requires significant time investment so the amount of labeled data can vary drastically in this study. When doing this manual labelling, personal bias is added. Different people might classify certain traits of a picture differently such as the emotion shown and ethnicity. File storage also poses a problem due to images requiring significantly more memory allocation than text.

VI. Conclusion

Image processing is a powerful tool that is being utilized in various fields such as security and surveillance. Deep learning advancements has made it possible to perform age estimation and gender classification on sets of images. However, this can be difficult due to the variance in an image data set. A large labeled data set similar to the one analyzed in this paper will improve the predictions of any CNN model.

This paper is focused on analyzing the output of three separate image processing APIs: Google's Cloud Vision, Amazon Rekognition, and a custom model created using Tensorflow and Keras. The conclusions drawn from this study can determine the effectiveness of each API to provide recommendations for future usage. This will allow others without significant deep learning experience to utilize readily-available APIs to perform image processing. Lastly, future research can be conducted on other user-friendly image processing APIs to determine their ability to predict age and gender, or any custom label.

References

[1] S. Escalera, J. Fabian, P. Pardo, X. Baró, J. Gonzalez, H. J. Escalante, D. Misevic, U. Steiner, and I. Guyon, "Chalearn looking at people 2015: Apparent age and cultural event recognition datasets and results," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2015, pp. 1–9.

[2] J. Xing, K. Li, W. Hu, C. Yuan, and H. Ling, "Diagnosing deep learning models for high accuracy age estimation from a single image," *Pattern Recognition*, vol. 66, pp. 106–116, 2017.

[3] G. Antipov, M. Baccouche, S.-A. Berrani, and J.-L. Dugelay, "Effective training of convolutional neural networks for face-based gender and age prediction," *Pattern Recognition*, vol. 72, pp. 15–26, 2017.

[4] Google Cloud, "Cloud Vision." [Online]. Available: https://cloud.google.com/vision/

[5] Amazon Web Services, "Amazon Rekognition." [Online]. Available: https://aws.amazon.com/rekognition/

[6] P. Rodríguez, G. Cucurull, J. M. Gonfaus, F. X. Roca, and J. Gonzalez, "Age and gender recognition in the wild with deep attention," *Pattern Recognition*, vol. 72, pp. 563–571, 2017.

[7] Exxact, "Beginner's guide: Image recognition and deep learning." [Online]. Available: hhttps://blog.exxactcorp.com/how-does-image-recognition-work-deep-learning-basics/

[8] R. Rothe, R. Timofte, and L. Van Gool, "Dex: Deep expectation of apparent age from a single image," in *The IEEE International Conference on Computer Vision (ICCV) Workshops*, December 2015.