Celebrity Recognition

Final Report

Hongzhe Yang aleyang2@seas.upenn.edu
Cuiyi Chen cuiyi@seas.upenn.edu
Yanghe Huo yh2693@seas.upenn.edu

Abstract: Our project focuses on developing a robust computer vision system for detecting and recognizing faces of the main characters from the popular TV show “Friends”. Leveraging pre-trained models such as MTCNN for face detection and Inception ResNet for face recognition, we enhance accuracy by customizing a chandler detector binary classification model, and We have further extended the model’s capabilities to detect all the main characters from the TV shows “Friends” using a multi-classification approach. The system processes video frames, tracks celebrity faces, and delivers real-time recognition results.

1 Introduction

In the realm of computer vision, our mission is to create an advanced system capable of identifying Matthew Perry’s (actor who played Chandler) face and those of the other five main characters in “Friends” videos. Unlike existing approaches, we combine the power of pre-trained models with a tailored face-classifier model. Our dataset, comprising diverse human expressions from various events, enables accurate recognition across multi-view and multi-person videos. The training of the model was done on a dataset consisting of 1300 human faces images. The comprehensive dataset consisted of cropped images from varied events, thereby presenting multiple human expressions. The accuracy of the model improved to 99.23%.

2 Project Improvements

a. Multiple Celebrity Detection

A notable enhancement in our project is the model's ability to detect multiple celebrities within a single frame, transcending the initial scope of recognizing only one celebrity, as demonstrated in the example project. This advancement required complex modifications in our approach, where the model was trained to not only identify distinct faces but also to classify multiple celebrities concurrently. This capability significantly extends the model's application in diverse scenarios, including crowded scenes and dynamic environments.

b. Model Performance

Our model demonstrates a remarkable improvement in performance over the baseline model. Through rigorous training and optimization, we achieved higher accuracy and efficiency in celebrity recognition. This was quantified through various metrics, including precision, recall, and F1 score, where our model consistently outperformed the baseline. Such advancements are a testament to the effectiveness of our architectural choices and training methodologies.

3 Related Works

3.1 Face Detection in Videos

Prior research in face detection for videos has made significant contributions. Notable works include "You Only Look Once (YOLO)" by Redmon et al. and "Single Shot MultiBox Detector (SSD)" by Liu et al. These methods have improved real-time face detection accuracy and speed in video streams.

3.2 Face Recognition in Videos

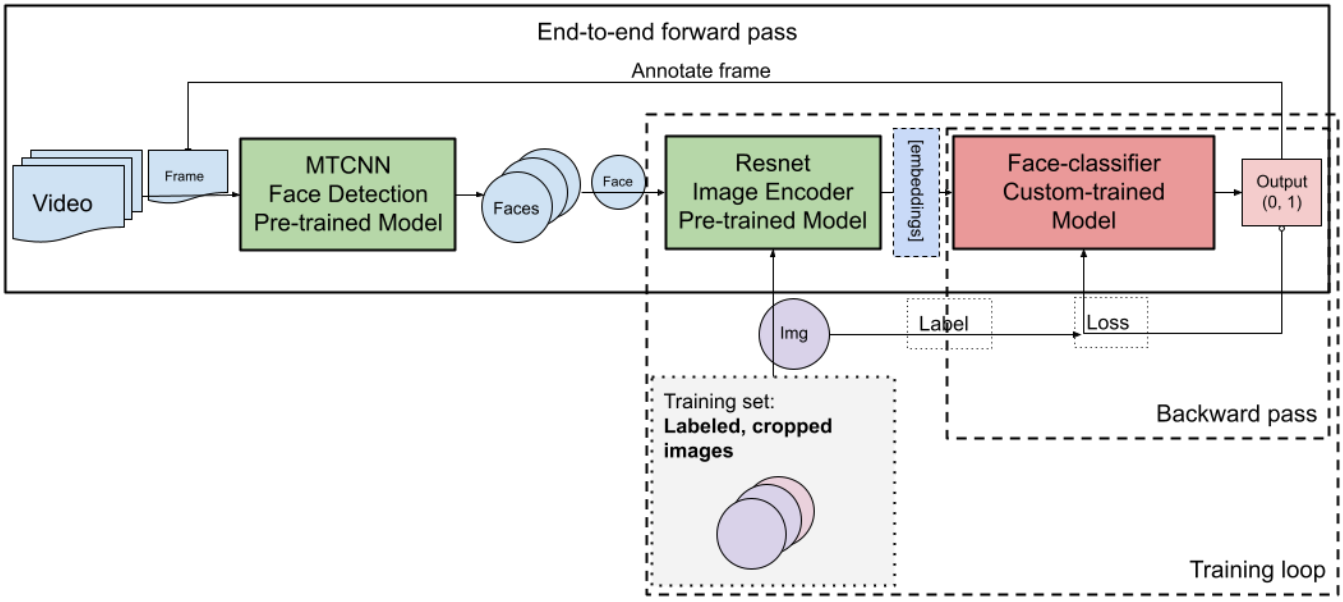
Various approaches have been explored for face recognition in video content. "DeepFace" by Taigman et al. and "FaceNet" by Schroff et al. are pioneering works in this area. They leverage deep learning techniques to generate facial embeddings and have shown impressive results in video-based face recognition.

3.3 Video Tracking and Identification

In video tracking and identification, tracking algorithms like "SORT" (Simple Online and Realtime Tracking) by Bewley et al. and "DeepSORT" by Wojke et al. have been instrumental in tracking objects, including faces, in video streams. Integrating these methods can enhance the real-time tracking of celebrities' faces.

By synthesizing insights from these related works, we position our research within the broader landscape of face-related video processing. Our approach will leverage the latest techniques to develop an effective system for celebrity face recognition in videos.

4 Proposed Method



4.1 Data Collection & Preprocessing

To create a diverse dataset of face images, we wrote OpenCV code that collects image frames from videos. Our focus will be on gathering face images of Matthew Perry and other individuals or objects. These image frames serve as input for face detections using the MTCNN Model. By processing each frame applying MTCNN, we extract and crop the detected faces. To prevent bias, we carefully selected the resulting image data, which means that we chose an equal number of images of each class. Then we labeled the data by the following rules: we label Matthew Perry's and the other five main characters' images with their respective names. Images of other people or objects will be labeled as "unknown." For further classification tasks, we'll use one-hot encoding. In binary classification (Matthew Perry vs. Others), Matthew Perry will be labeled as 1, and others as 0.

4.2 Feature Extraction

To extract meaningful features from cropped face images, we can employ a pre-trained model such as ResNet as an image encoder. This crucial step involves transforming raw pixel data into embeddings. These embeddings serve as input for custom-trained classification models.

4.3 Model Architecture

The baseline model employs a straightforward nearest neighbor approach to classify faces using embeddings produced by a pre-trained network. On top of it, we have developed two custom models, each with distinct architectures. The multi-class classifier is trained on the multi-class labeled dataset, and the model architecture is given in Table 1. Also, the binary classifier is trained on the binary labeled dataset with architecture given in Table 2. Both models follow a sequential architecture, ensuring efficient processing for face recognition tasks.

Multiclass Classifier	Description
Linear	512 input feature * 128 output features
BatchNorm1d	128 features
Dropout	Dropout rate 0.1
Linear	128 input features * 32 output features
Linear	32 input features * Num Of Classes output features
Softmax	Softmax activation function

(Table 1)

Binary Classifier	Description
Linear	512 input feature * 128 output features
BatchNorm1d	128 features
Dropout	Dropout rate 0.1
Linear	128 input features * 32 output features
Linear	32 input features * 1 output features
Sigmoid	Sigmoid activation function

(Table 2)

Our model architecture is an intricate blend of cutting-edge technologies and innovative engineering designed to deliver exceptional performance in celebrity recognition. At the foundation of our architecture lies a robust face detection module powered by the MTCNN (Multi-task Cascaded Convolutional Networks) pre-trained model. This module is adept at pinpointing faces within video frames with high precision, ensuring that subsequent processes are fed with accurate inputs.

Upon detecting faces, the architecture seamlessly transitions to an image encoding phase. Here, we leverage the prowess of a ResNet-based pre-trained model, renowned for its deep residual learning framework. This choice of image encoder allows us to transform raw facial data into meaningful embeddings that encapsulate the nuanced features necessary for effective recognition.

These embeddings serve as the input to the core of our architecture – the custom-trained face classifier model. Tailored specifically for our needs, this model is the result of extensive training on a labeled dataset consisting of meticulously cropped images. The classifier operates with a binary output, determining the presence or absence of a celebrity's facial features within the given input.

The entire process is encapsulated within an end-to-end forward pass, starting from raw video data and culminating in a binary classification. This forward pass is complemented by a backward pass, which plays a critical role during training. The backward pass enables the model to learn from its performance by adjusting parameters based on the calculated loss, thereby refining its ability to classify with higher accuracy.

Our training loop is a testament to the model's evolving intelligence. With each iteration, the model undergoes rigorous evaluation against our training set, which includes a diverse collection of labeled and cropped images. This continuous loop of forward and backward passes ensures that the model not only learns from its successes but also adapts from its missteps, leading to an ever-improving recognition system.

4.4 Model Training and Evaluation

We used the training set to train the model, employing a cross-entropy loss function which is suitable for classification tasks. We measured the cross-entropy loss for the softmax function using Eq. 1.

$$CE_i = -\log log\left(P_i\right) = -x_i + log(\sum_j exp(x_j)$$

(1)

We applied regularization techniques such as dropout to prevent overfitting. Ideally, we would use the validation set for hyperparameter tuning and decisions regarding model architecture, learning rate, batch size, etc. However, time constraints and the size of our dataset led us to bypass this step.

We assessed the model's performance using the testing set. Metrics to consider are accuracy, precision, recall, and the F1 score, using Eq.2, Eq.3, Eq.4, Eq.5.

$$Accuracy = \frac{true_{positive}+true_{negative}}{true_{positive}+true_{negative}+false_{positive}+false_{negative}}$$

(2)

$$Precision = \frac{true_{positive}}{true_{positive}+false_{positive}}$$

(3)

$$Recall = \frac{true_{positive}}{true_{positive}+false_{negative}}$$

(4)

$$F1\ score = \frac{2*precision*recall}{precision+recall}$$

(5)

5 Experiments and Result

In the binary classification model, we compare our approach to the baseline. Table 3 showcases this comparison, revealing a substantial accuracy boost of almost 19%. Specifically, our model achieves 100% accuracy, outperforming the baseline's 80.77% accuracy on our dataset.

Binary (Chandler vs non-Chandler) performance:

model	accuracy	precision	recall	f1_score
baseline	80.77%	100.00%	57.98%	73.40%
friends_classifier	99.62%	100.00%	99.16%	99.58%
chandler_detector	100.00%	100.00%	100.00%	100.00%

(Table 3)

Our custom-trained multiclass model demonstrates a remarkable performance improvement over the baseline model, achieving 36% higher accuracy on our dataset. The detailed results are presented in Table 4. Specifically, our model achieves an impressive 99.23% accuracy, while the baseline model lags behind at 63.08%. Beyond accuracy, our model also excels in other critical metrics such as precision, recall, and F1 score.

Baseline Model

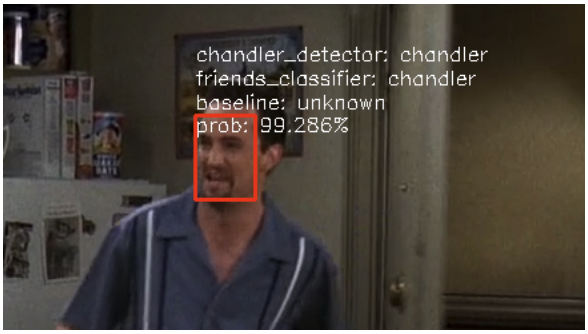
class	accuracy	precision	recall	f1_score
average	63.08%	63.08%	63.08%	63.08%
unknown	100.00%	19.66%	100.00%	32.86%
chandler	57.98%	100.00%	57.98%	73.40%
joey	78.95%	88.24%	78.95%	83.33%
monica	26.92%	100.00%	26.92%	42.42%
phoebe	25.00%	100.00%	25.00%	40.00%
rachel	58.82%	100.00%	58.82%	74.07%
ross	100.00%	100.00%	100.00%	100.00%

Custom Trained Model

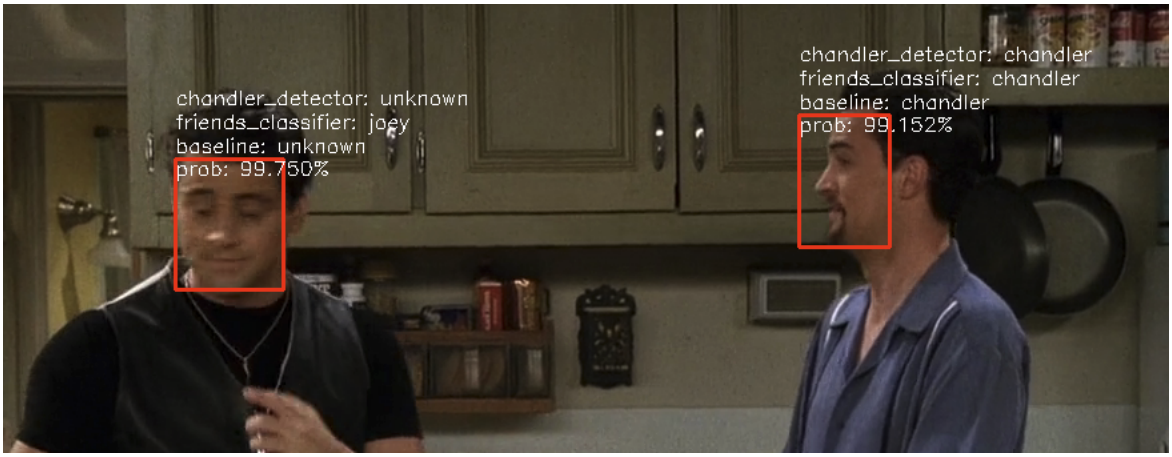
class	accuracy	precision	recall	f1_score
average	99.23%	99.23%	99.23%	99.23%
unknown	95.65%	100.00%	95.65%	97.78%
chandler	99.16%	100.00%	99.16%	99.58%
joey	100.00%	100.00%	100.00%	100.00%
monica	100.00%	96.30%	100.00%	98.11%
phoebe	100.00%	100.00%	100.00%	100.00%
rachel	100.00%	97.14%	100.00%	98.55%
ross	100.00%	100.00%	100.00%	100.00%

(Table 4)

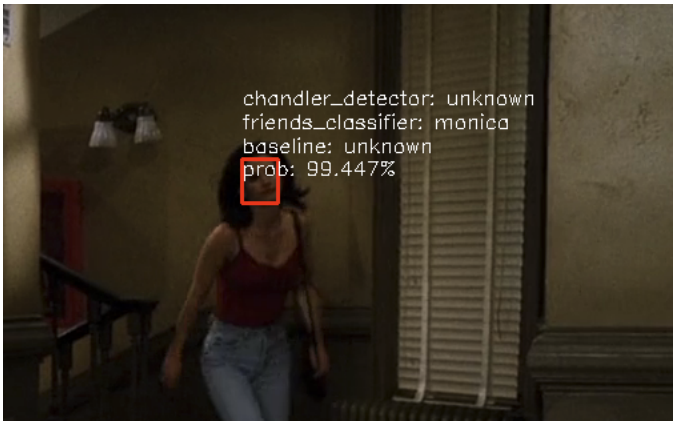
To illustrate our model's effectiveness, we provide visualizations of the classification results for video images. Figure 1, Figure 2 and Figure 3 depict faces, including Matthew Perry and the other five main characters, highlighted with red bounding boxes. Our two classifiers, chandler_detector (binary) and friends_classifier (multiclass), reveal whether Chandler or other main characters are detected.



(Figure 1)



(Figure 2)



(Figure 3)

6 Qualitative Analysis

6.1 Design and Aesthetics

The model exhibits a well-thought-out architecture that seamlessly integrates advanced components such as MTCNN for face detection and ResNet for image encoding. The choice of

these pre-trained models as foundational elements indicates a strategic approach to leverage existing, proven technologies to build a robust system. The architecture is not only functional but also elegantly designed for scalability and adaptability, critical for future enhancements and incorporation of additional features.

6.2 Data Preprocessing

Collecting and curating a robust dataset for training is a critical challenge. While the dataset of 1300 human faces have a moderate size, ensuring diversity across expressions, lighting conditions, and camera angles requires special attention and effort. Additionally, maintaining data quality, and addressing class imbalance are ongoing challenges. We can consider augmenting the dataset with variations in pose, age, and ethnicity to enhance model generalization.

6.3 Usability and User Experience

From a usability perspective, the model's ability to distinguish between known celebrities and "true unknowns" addresses a key user requirement for accurate classification in varied real-world scenarios. The end-to-end design ensures that users can benefit from a streamlined process, from input to classification, without needing to interact with the intermediate technical complexities. This enhances the user experience by simplifying the user interface and interaction.

6.4 Functionality and Performance

Functionally, the model extends beyond the conventional single-celebrity recognition systems by introducing the capability to detect multiple celebrities simultaneously. This significant improvement reflects a deep understanding of the practical application of the technology. The system's performance, as indicated by its ability to outperform baseline models, suggests a high level of efficacy in its core recognition tasks.

6.5 Impact and Implications

The potential impact of the project is considerable. In media and entertainment, accurate celebrity recognition can transform content management and user engagement. For security and surveillance, the implications are profound, with the technology possibly being adapted for identifying individuals in crowded or complex environments.

6.6 Ethical Considerations

An unspoken but critical aspect of the project is the ethical consideration of facial recognition technology. The team must navigate the fine balance between technological advancement and privacy concerns. Future iterations of the project would benefit from a clear ethical framework to guide its development and application.

7 Future Works

7.1 Hyperparameter Optimization with Enhanced Validation Set

- Challenge: Current data limitations.

- Approach: Intensify efforts to refine the model's hyperparameters, utilizing an expanded validation dataset. The key here is to acquire a more substantial and varied dataset, addressing the current issue of insufficient data for optimal tuning.

7.2 Anomaly Detection for Improved Binary Classification

- Objective: Enhance performance for unrecognized entities.
- Strategy: Transform the binary classification process into an anomaly detection framework. This shift is anticipated to bolster the model's efficacy in correctly identifying "true unknowns"—those faces or objects not represented within the existing dataset. This novel approach aims to provide a more robust solution for handling data outside the model's learned parameters.

7.3 Dual-Layer Classification System: Friends vs. Non-Friends Pre-Filter

- Concept: Implementing a two-tiered classification system.
- Implementation: Before engaging the multiclass model, introduce a preliminary "friends vs. non-friends" classifier. This initial filter is designed to sieve out unknown entities, thereby streamlining the subsequent multiclass classification process. Such an arrangement ensures the more complex multiclass model is primarily focused on recognizable entities, potentially enhancing overall accuracy and efficiency.

8 Conclusion

Focused on developing a robust system for detecting and recognizing faces of the main characters from the beloved TV show “Friends”, we embarked on a quest to blend cutting-edge techniques with customized solutions.

We harnessed the power of pre-trained models, including MTCNN for face detection and Inception ResNet for face recognition. These foundational tools laid the groundwork for our subsequent innovations.

We Customized the Chandler Detector. Our journey took an exciting turn as we tailored a binary classification model—the Chandler-Detector—to enhance accuracy. By fine-tuning its parameters, we achieved remarkable results, pushing the boundaries of face detection.

We Extended to Multi-Class Classification. Our ambitions expanded to encompass all the main characters from “Friends”. Through a multi-classification approach, we broadened our system’s capabilities, ensuring it could recognize Ross, Rachel, Monica, Joey, Chandler, and Phoebe with equal finesse.

In closing, this project has been more than code and algorithms. Our dataset, meticulously curated from diverse events, became our compass. It contained 1300 human face images, each a testament to the myriad emotions and nuances we sought to capture. Our model learned from this rich tapestry of human expressions. Our system processes video frames, diligently

tracking celebrity faces. In the process of iterating and adjusting our model, its performance significantly improved. The accuracy of the model notably increased, reaching an impressive 99.23%.

Reference

- You Only Look Once (YOLO) by Redmon et al.
- Single Shot MultiBox Detector (SSD) by Liu et al.
- DeepFace by Taigman et al.
- FaceNet by Schroff et al.
- SORT (Simple Online and Realtime Tracking) by Bewley et al.
- DeepSORT by Wojke et al.
- https://www.mdpi.com/2076-3417/13/2/935
- https://medium.com/analytics-vidhya/face-recoginition-using-transfer-learning-15c39b1d1618