University of East Anglia

Audio-visual speech recognition

Mohammad Alhnaity

Introduction

Exploring Speech Recognition : this work focuses on combining audio and visual data for improved name recognition, navigating the complexities of multimodal speech analysis.

- Integrating Audio-Visual Data for Speech Recognition : Utilizes audio data for feature extraction and employs a 1D Convolutional Neural Network as a classifier, blending visual information from lip movements to enhance the recognition of spoken names.

Holistic Approach

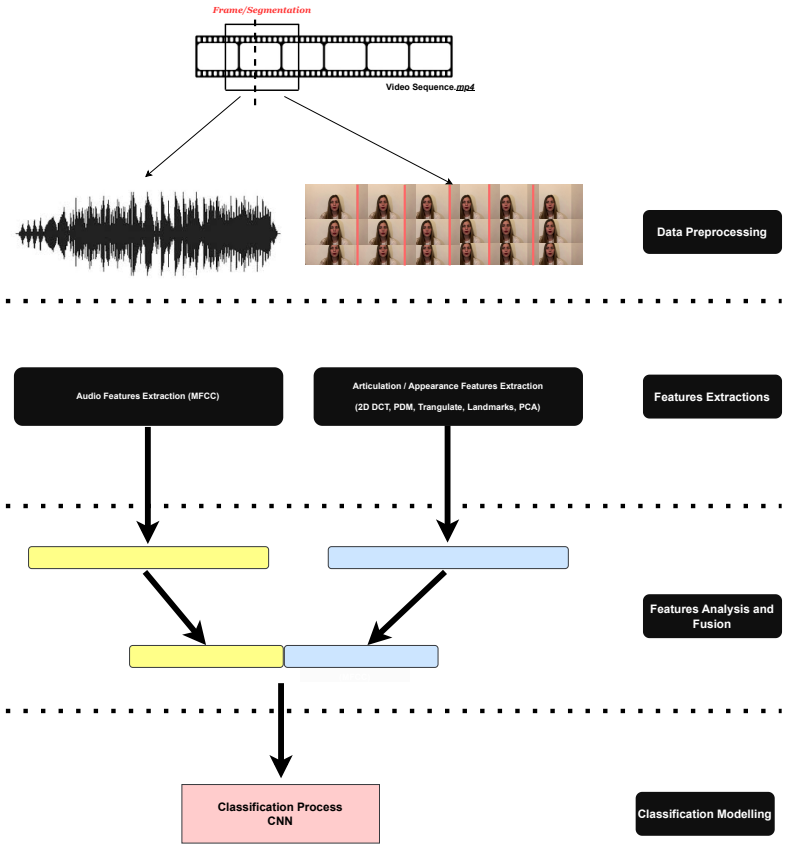

Figure – Holistic approach

Articulatory type Features

- Advanced Face and Landmark Detection : Utilizes a Deep Neural Network (DNN) model for accurate face detection in images and Dlib's landmark predictor for identifying specific facial features. This combination enables precise localization of mouth regions in diverse images, essential for detailed shape analysis.
- Mouth Articulation Feature Extraction : Focuses on extracting critical mouth shape features – width, height, area, and perimeter – from the identified mouth regions. This data is crucial for applications in speech analysis, language studies, and facial expression recognition, providing valuable insights into mouth articulation patterns.

Articulatory type Features

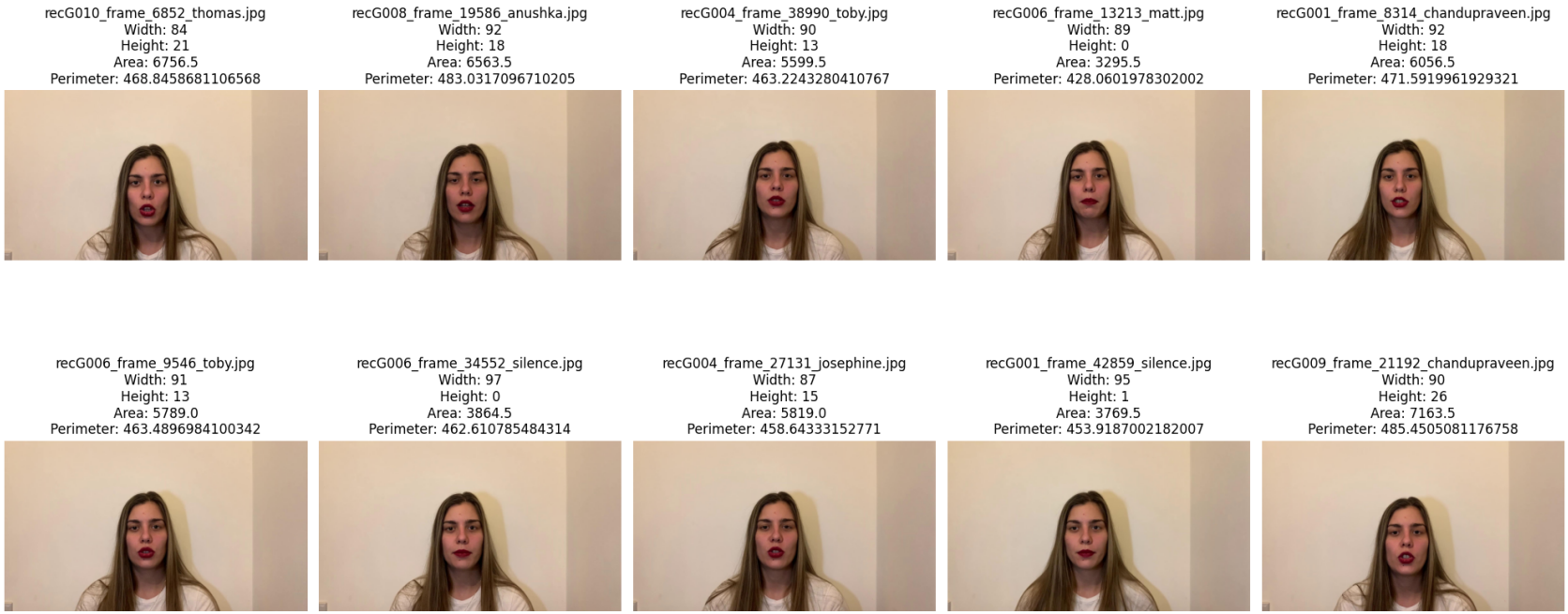

Figure – Shape features

Appearance type Features (PDM, Triangulation)

- Integrated Facial Landmark Detection and Image Warping : Using face detector and landmark predictor to identify facial features in images, and then applies an image warping technique to align these features to a mean shape. This process standardizes facial features across different images, making them suitable for consistent analysis and comparison.
- Feature Dimensionality Reduction Using Incremental PCA : Implements Incremental Principal Component Analysis (IPCA) to efficiently reduce the dimensionality of the warped facial images. This technique transforms high-dimensional pixel data into a lower-dimensional space, capturing the most significant features.

Appearance type Features (PDM, Triangulation)

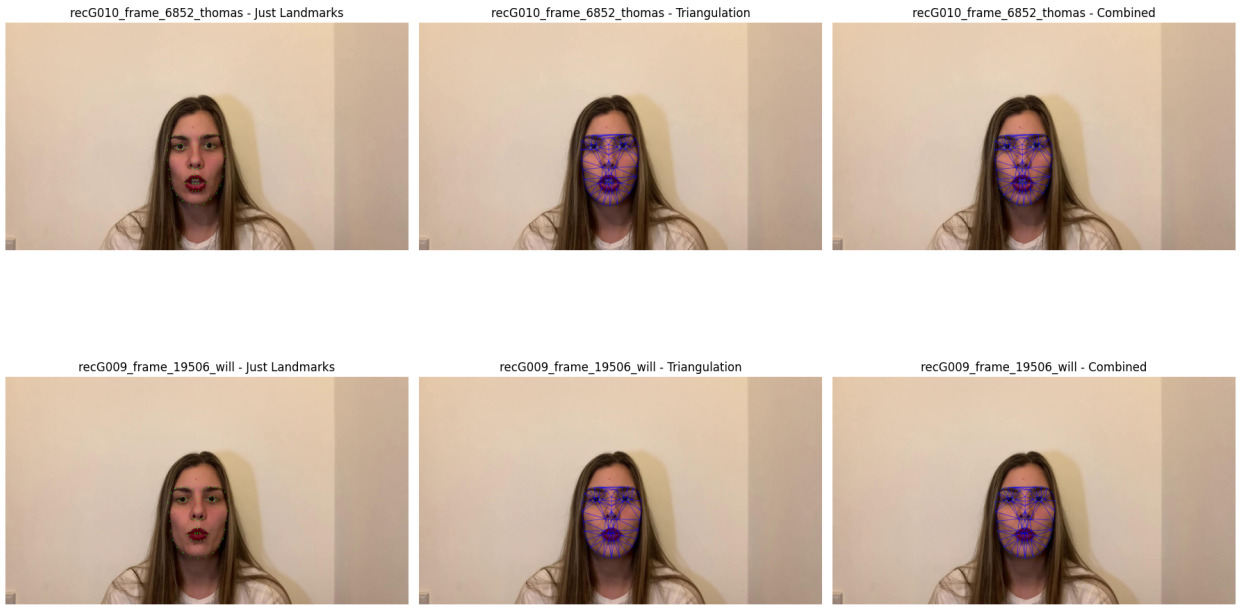

Figure – Appearance features

Appearance type Features (2D DCT)

- Efficient Image Feature Extraction Using 2D DCT : Applying a two-dimensional Discrete Cosine Transform (2D DCT) to grayscale images to extract frequency components. It then focuses on the top left 8x8 block of the DCT coefficients, which captures the most significant information of the image. This method is particularly effective for reducing the data size while retaining key features of the images.
- Dimensionality Reduction with PCA for Feature Analysis : After extracting DCT features, Principal Component Analysis (PCA) is used to further reduce the dimensionality of these features. This step transforms the high-dimensional DCT data into a lower-dimensional space, represented by the principal components. This process is crucial for efficient data analysis and is particularly useful in machine learning and pattern recognition tasks involving large image datasets.

Appearance type Features (2D DCT)

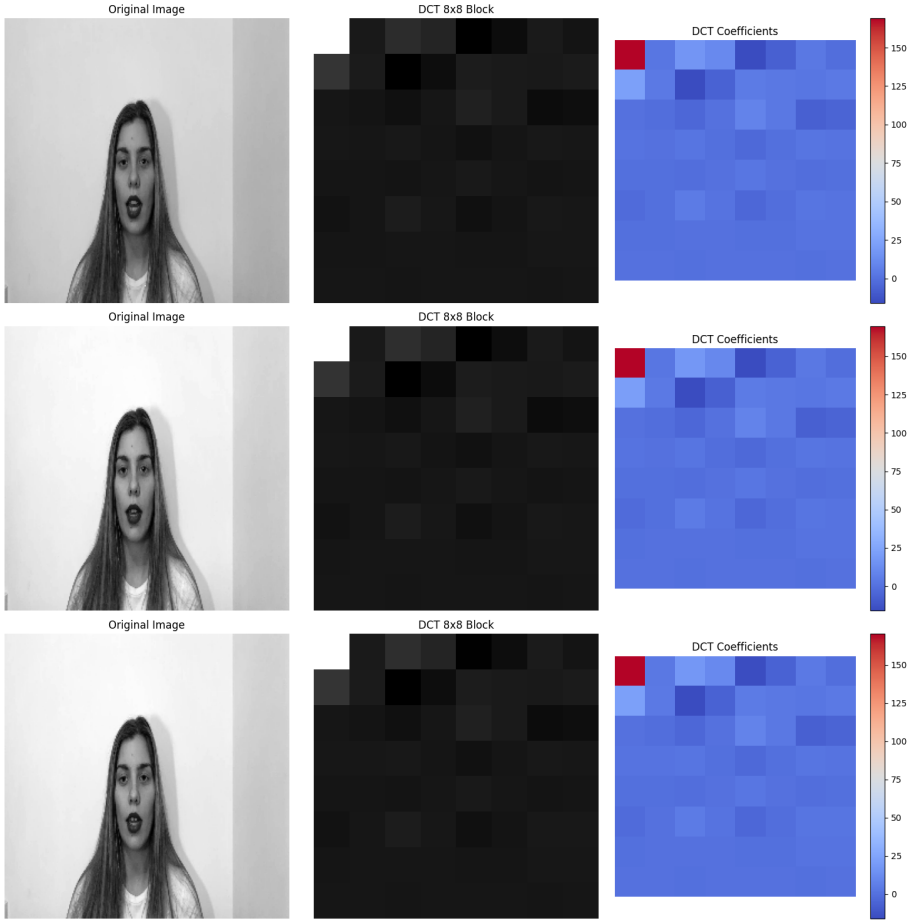

Figure – Appearance features

Assessment, Integration and Features Processing

- Dataset Balancing by Oversampling : balancing a dataset by oversampling blocks of data. For each unique label within a block, it replicates entries until a minimum count is reached, ensuring that all classes are adequately represented.
- Block Processing for Data Uniformity : It processes the dataset in defined block sizes, applying the balancing operation to each block. This method ensures that the balance is maintained throughout the dataset, enhancing the uniformity and reliability of data for analysis.

Classification Model

- 1D CNN Model Design : Utilizes a 1D Convolutional Neural Network with layers designed for detailed feature extraction from lip movement data, comprising convolutional layers, max pooling, flattening, and dense layers with dropout for optimization.
- Efficient Training and Evaluation : The model is meticulously trained and evaluated on reshaped data, using the 'adam' optimizer and 'sparse categorical crossentropy' for multi-class classification, ensuring precise learning and accurate performance assessment.

Shape Features Classification Model Performance Measurements

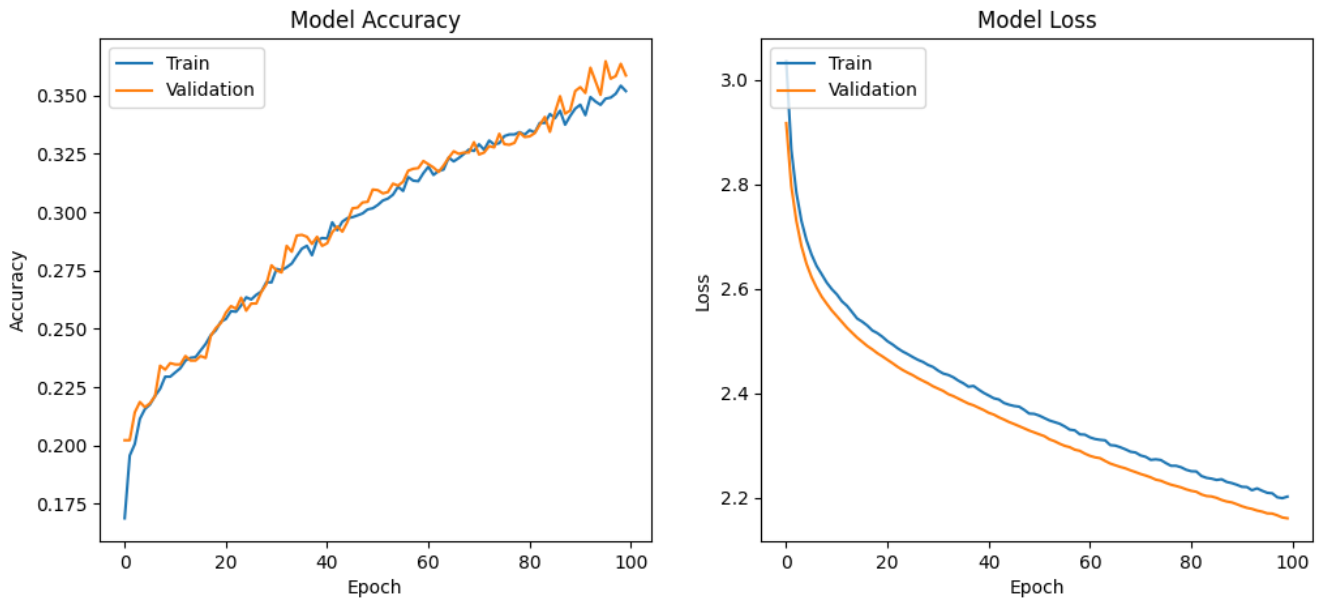

Figure – Shape features model (loss and accuracy

Shape Features Classification Model Performance Measurements

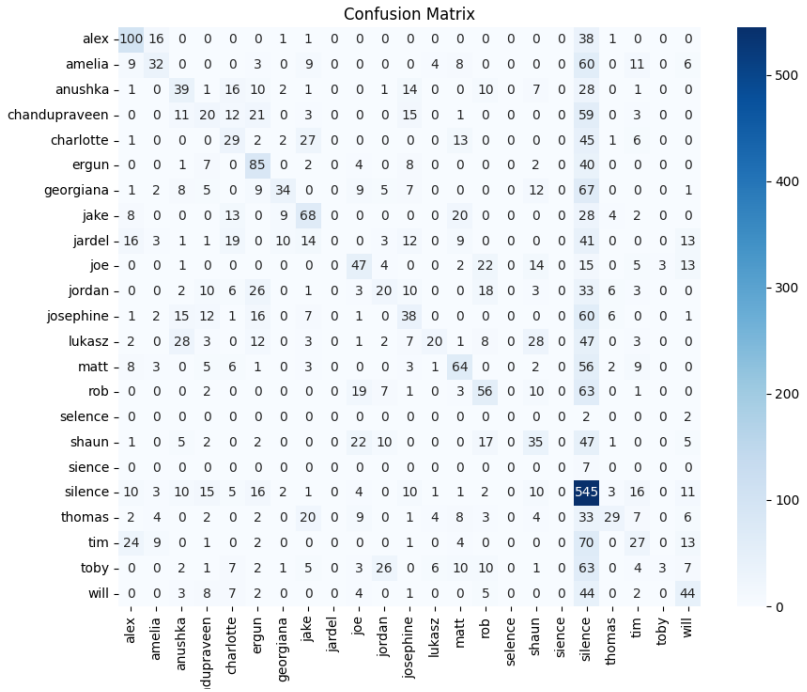

Figure – Shape features model confusion matrix

Shape Features Classification Model Performance Measurements

Classification Report

	precision	recall	f1-score	support
alex	0.5434782608695652	0.6369426751592356	0.5865102639296187	157.0
amelia	0.43243243243243246	0.22535211267605634	0.2962962962962963	142.0
anushka	0.30952380952380953	0.29770992366412213	0.30350194552529186	131.0
chandupraveen	0.21052631578947367	0.13793103448275862	0.16666666666666666	145.0
charlotte	0.2396694214876033	0.23015873015873015	0.23481781376518218	126.0
ergun	0.4028436018957346	0.5704697986577181	0.47222222222222215	149.0
georgiana	0.5573770491803278	0.2125	0.30769230769230765	160.0
jake	0.4121212121212121	0.4473684210526316	0.42902208201892744	152.0
jardel	0.0	0.0	0.0	142.0
joe	0.373015873015873	0.373015873015873	0.373015873015873	126.0
jordan	0.2564102564102564	0.14184397163120568	0.18264840182648404	141.0
josephine	0.296875	0.2375	0.2638888888888889	160.0
lukasz	0.5555555555555556	0.12121212121212122	0.1990049751243781	165.0
matt	0.4444444444444444	0.39263803680981596	0.41693811074918574	163.0
rob	0.3708609271523179	0.345679012345679	0.35782747603833864	162.0
selence	0.0	0.0	0.0	4.0
shaun	0.2734375	0.23809523809523808	0.25454545454545446	147.0
sience	0.0	0.0	0.0	7.0
silence	0.3655264922870557	0.8195488721804511	0.50556586270872	665.0
thomas	0.5471698113207547	0.21641791044776118	0.31016042780748665	134.0
tim	0.27	0.17880794701986755	0.21513944223107567	151.0
toby	0.5	0.019867549668874173	0.038216560509554146	151.0
will	0.36065573770491804	0.36666666666666664	0.36363636363636365	120.0
accuracy	0.37083333333333335	0.37083333333333335	0.37083333333333335	0.37083333333333335
macro avg	0.33573581309527545	0.26998808238890465	0.2729268450086225	3600.0
weighted avg	0.3696762565767487	0.37083333333333335	0.32869067145171216	3600.0

Figure – Classification report

Appearance Features Model Performance Measurements(PDM,T, PCA)

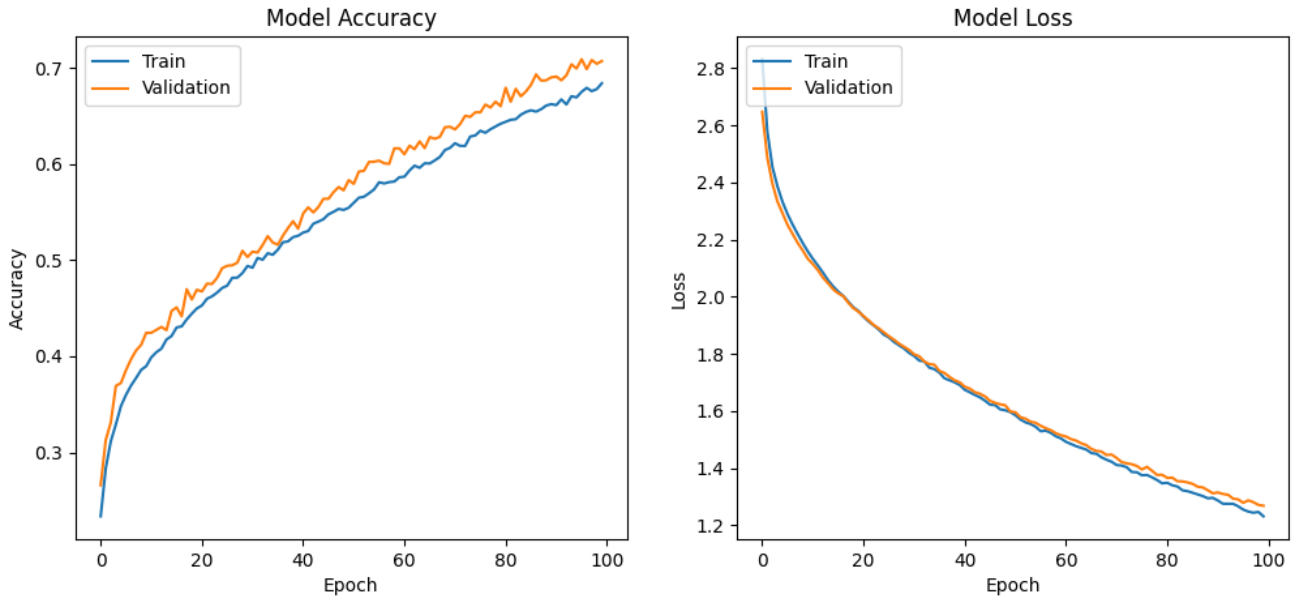

Figure – loss and accuracy

Appearance Features Model Performance Measurements(PDM,T, PCA)

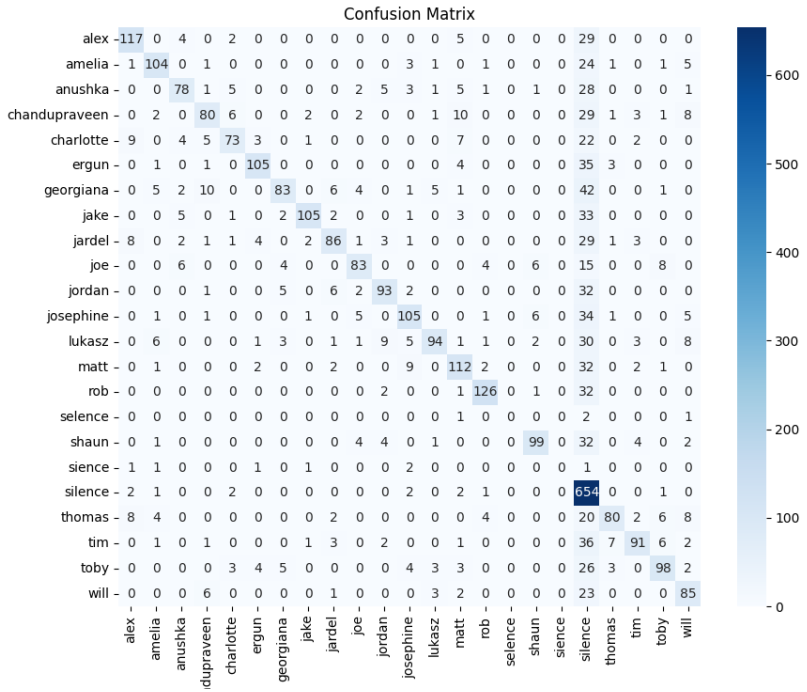

Figure – Confusion matrix

Appearance Features Model Performance Measurements(PDM,T, PCA)

Classification Report

	precision	recall	f1-score	support
alex	0.8013698630136986	0.7452229299363057	0.7722772277227723	157.0
amelia	0.8125	0.7323943661971831	0.7703703703703704	142.0
anushka	0.7722772277227723	0.5954198473282443	0.6724137931034482	131.0
chandupraveen	0.7407407407407407	0.5517241379310345	0.6324110671936759	145.0
charlotte	0.7849462365591398	0.5793650793650794	0.6666666666666667	126.0
ergun	0.875	0.7046979865771812	0.7806691449814126	149.0
georgiana	0.8137254901960784	0.51875	0.633587786259542	160.0
jake	0.9292035398230089	0.6907894736842105	0.7924528301886792	152.0
jardel	0.7889908256880734	0.6056338028169014	0.6852589641434262	142.0
joe	0.7980769230769231	0.6587301587301587	0.7217391304347825	126.0
jordan	0.788135593220339	0.6595744680851063	0.718146718146718	141.0
josephine	0.7608695652173914	0.65625	0.7046979865771812	160.0
lukasz	0.8623853211009175	0.5696969696969697	0.6861313868613139	165.0
matt	0.7088607594936709	0.6871165644171779	0.6978193146417445	163.0
rob	0.8936170212765957	0.7777777777777778	0.8316831683168316	162.0
selence	0.0	0.0	0.0	4.0
shaun	0.8608695652173913	0.673469387755102	0.7557251908396946	147.0
sience	0.0	0.0	0.0	7.0
silence	0.5274193548387097	0.9834586466165414	0.6866141732283466	665.0
thomas	0.8247422680412371	0.5970149253731343	0.6926406926406926	134.0
tim	0.8272727272727273	0.6026490066225165	0.6973180076628352	151.0
toby	0.7967479674796748	0.6490066225165563	0.7153284671532848	151.0
will	0.6692913385826772	0.7083333333333334	0.688259109311741	120.0
accuracy	0.7086111111111111	0.7086111111111111	0.7086111111111111	0.7086111111111111
macro avg	0.7233496664592074	0.6063945862939355	0.6522700520193547	3600.0
weighted avg	0.753300766498083	0.7086111111111111	0.7091450745267093	3600.0

Figure – Classification report

Appearance Features Model Performance Measurements(2D DCT, PCA)

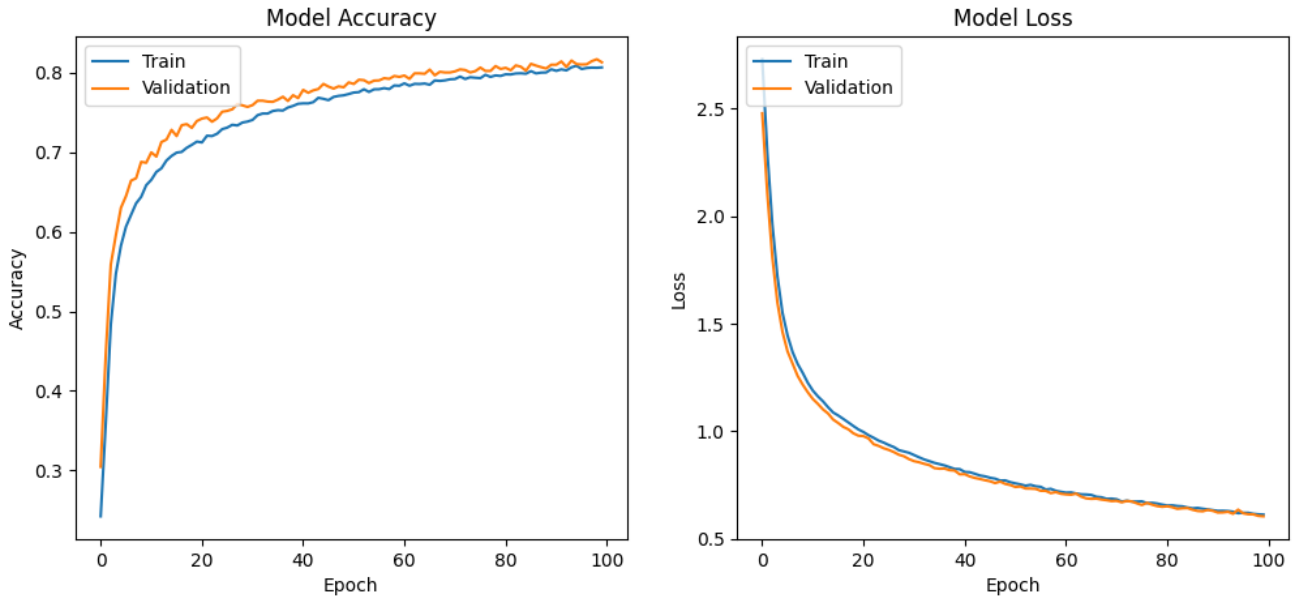

Figure – loss and accuracy

Appearance Features Model Performance Measurements(2D DCT, PCA)

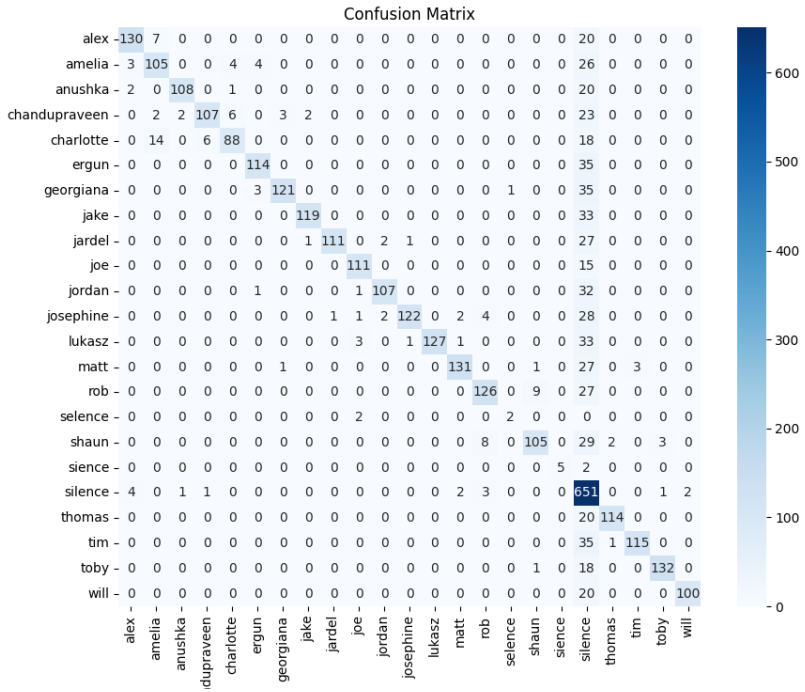

Figure – Confusion matrix

Appearance Features Model Performance Measurements(2D DCT, PCA)

Classification Report

	precision	recall	f1-score	support
alex	0.935251798561151	0.8280254777070064	0.8783783783783783	157.0
amelia	0.8203125	0.7394366197183099	0.7777777777777778	142.0
anushka	0.972972972972973	0.8244274809160306	0.8925619834710745	131.0
chandupraveen	0.9385964912280702	0.7379310344827587	0.8262548262548263	145.0
charlotte	0.8888888888888888	0.6984126984126984	0.7822222222222223	126.0
ergun	0.9344262295081968	0.7651006711409396	0.8413284132841329	149.0
georgiana	0.968	0.75625	0.8491228070175438	160.0
jake	0.9754098360655737	0.7828947368421053	0.8686131386861314	152.0
jardel	0.9910714285714286	0.7816901408450704	0.8740157480314961	142.0
joe	0.940677966101695	0.8809523809523809	0.9098360655737705	126.0
jordan	0.963963963963964	0.7588652482269503	0.8492063492063492	141.0
josephine	0.9838709677419355	0.7625	0.8591549295774648	160.0
lukasz	1.0	0.7696969696969697	0.8698630136986302	165.0
matt	0.9632352941176471	0.803680981595092	0.8762541806020068	163.0
rob	0.8936170212765957	0.7777777777777778	0.8316831683168316	162.0
selence	0.6666666666666666	0.5	0.5714285714285715	4.0
shaun	0.9051724137931034	0.7142857142857143	0.7984790874524715	147.0
sience	1.0	0.7142857142857143	0.8333333333333333	7.0
silence	0.5545144804088586	0.9789473684210527	0.7079934747145189	665.0
thomas	0.9743589743589743	0.8507462686567164	0.9083665338645418	134.0
tim	0.9745762711864406	0.7615894039735099	0.8550185873605948	151.0
toby	0.9705882352941176	0.8741721854304636	0.9198606271777003	151.0
will	0.9803921568627451	0.8333333333333334	0.9009009009009009	120.0
accuracy	0.8197222222222222	0.8197222222222222	0.8197222222222222	0.8197222222222222
macro avg	0.9215897633725662	0.7780435742043735	0.8383327877535336	3600.0
weighted avg	0.8761183208067301	0.8197222222222222	0.830021745803857	3600.0

Figure – Classification report

All Features Model Performance Measurements

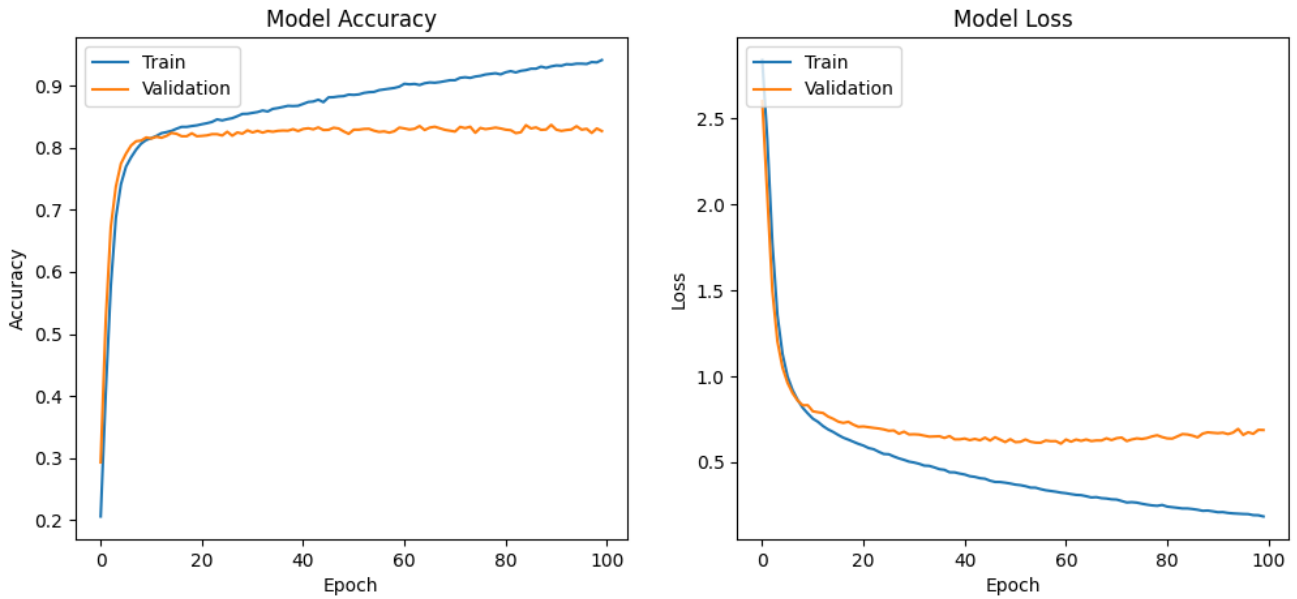

Figure – loss and accuracy

All Features Model Performance Measurements

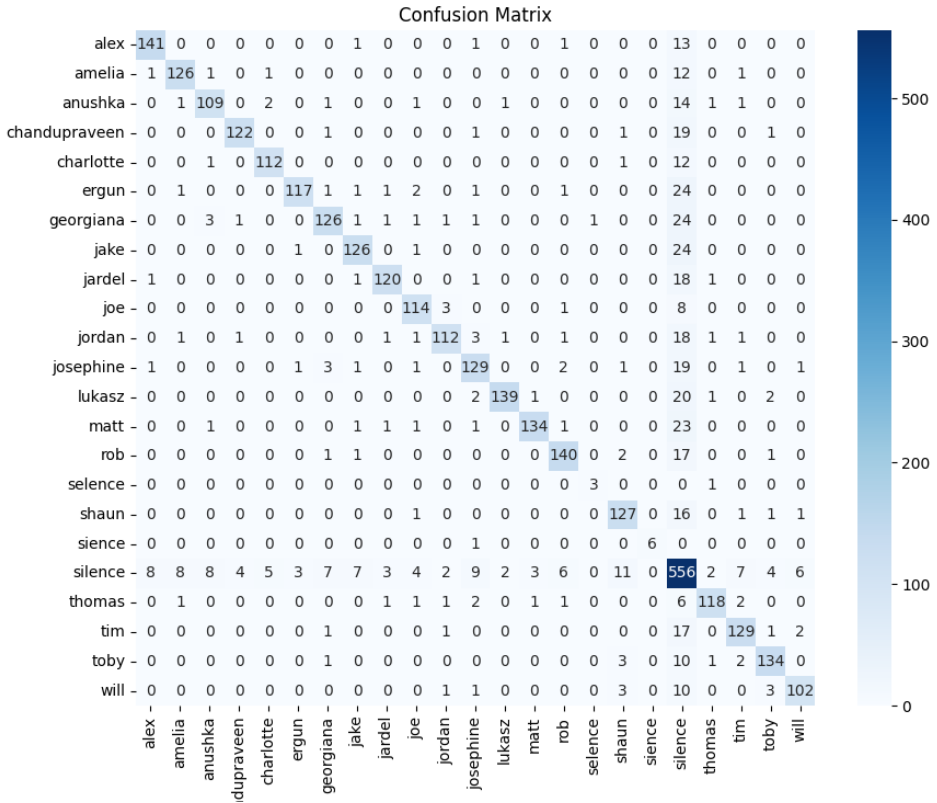

Figure – Confusion matrix

All Features Model Performance Measurements

Classification Report

	precision	recall	f1-score	support
alex	0.9276315789473685	0.8980891719745223	0.9126213592233009	157.0
amelia	0.9130434782608695	0.8873239436619719	0.9	142.0
anushka	0.8861788617886179	0.8320610687022901	0.8582677165354332	131.0
chandupraveen	0.953125	0.8413793103448276	0.8937728937728938	145.0
charlotte	0.9333333333333333	0.8888888888888888	0.9105691056910569	126.0
ergun	0.9590163934426229	0.785234899328859	0.8634686346863468	149.0
georgiana	0.8873239436619719	0.7875	0.8344370860927152	160.0
jake	0.9	0.8289473684210527	0.8630136986301371	152.0
jardel	0.9375	0.8450704225352113	0.8888888888888888	142.0
joe	0.890625	0.9047619047619048	0.8976377952755906	126.0
jordan	0.9256198347107438	0.7943262411347518	0.8549618320610687	141.0
josephine	0.8431372549019608	0.80625	0.8242811501597445	160.0
lukasz	0.972027972027972	0.8424242424242424	0.9025974025974025	165.0
matt	0.9640287769784173	0.8220858895705522	0.8874172185430463	163.0
rob	0.9090909090909091	0.8641975308641975	0.8860759493670887	162.0
selence	0.75	0.75	0.75	4.0
shaun	0.8523489932885906	0.8639455782312925	0.8581081081081081	147.0
sience	1.0	0.8571428571428571	0.923076923076923	7.0
silence	0.6318181818181818	0.8360902255639098	0.7197411003236246	665.0
thomas	0.9365079365079365	0.8805970149253731	0.9076923076923077	134.0
tim	0.8896551724137931	0.8543046357615894	0.8716216216216217	151.0
toby	0.9115646258503401	0.8874172185430463	0.8993288590604027	151.0
will	0.9107142857142857	0.85	0.8793103448275861	120.0
accuracy	0.845	0.845	0.845	0.845
macro avg	0.8993170231625179	0.8438277570774497	0.8689952172276214	3600.0
weighted avg	0.8629490728701347	0.845	0.8496839187768341	3600.0

Figure – Classification report

Experimental results comparison

Feature Type	**Accuracy**
Articulatory - Height, Width, Area and Perimeter	37.1%
Appearance - Landmark PCA	70.8%
Appearance - 2D DCT PCA	81.9%
All features - Appearance and Shape	84.5%

Table – Model Accuracy for Various Feature Type


thank you
danke
謝謝
ngiyabonga
شكراً جزيلاً
teşekkür ederim
спасибо
dank je
gracias
bedankt
dziękuje
mauruuru
mochchakkeram
go raibh maith agat
obrigado
sukriya
kop khun krap
arigato
dakujem
terima kasih
grazie
merci
감사합니다
мерси