Unconstrained Ear Recognition through State-of-the-Art Machine Learning Models: A Survey

Marwin B. Alejo

2020-20221

Disclaimer: Due to time and computing resource constraints, this topic yielded and presented only the preliminary results. Further executions are still needed for this to be a perfect paper and realized for journal publications or presentations.

Biometric Recognition and the Ears

- Advantages of Ear for Biometrics
  - Ear does not change until the age of 60.
  - Color distribution is more uniform in ear than in face, iris, retina.
  - Ear images are smaller than face images, thus more efficient when computed.
  - Ear images are not normally affected by facial accessories except for hair and earrings.

Selected Notable Works in Ear Biometrics

- Most papers utilize image processing and statistical techniques in realizing ear biometrics:
  - Canny edge algorithm, PCA, genetic algorithm
- Recent paper suggested the use of ML and CNN due to its advantage over traditional computing methods:
  - Deep unsupervised learning with GAN
  - SSD-MobileNet
  - Deep Convolutional Neural Network and handcrafted NN
  - Transfer Learning of pre-trained SOTA CNN models
- However, no existing study that present ear biometric solutions that utilizes SOTA Transformer-based and Transformer-inspired networks.
  - An open-opportunity.

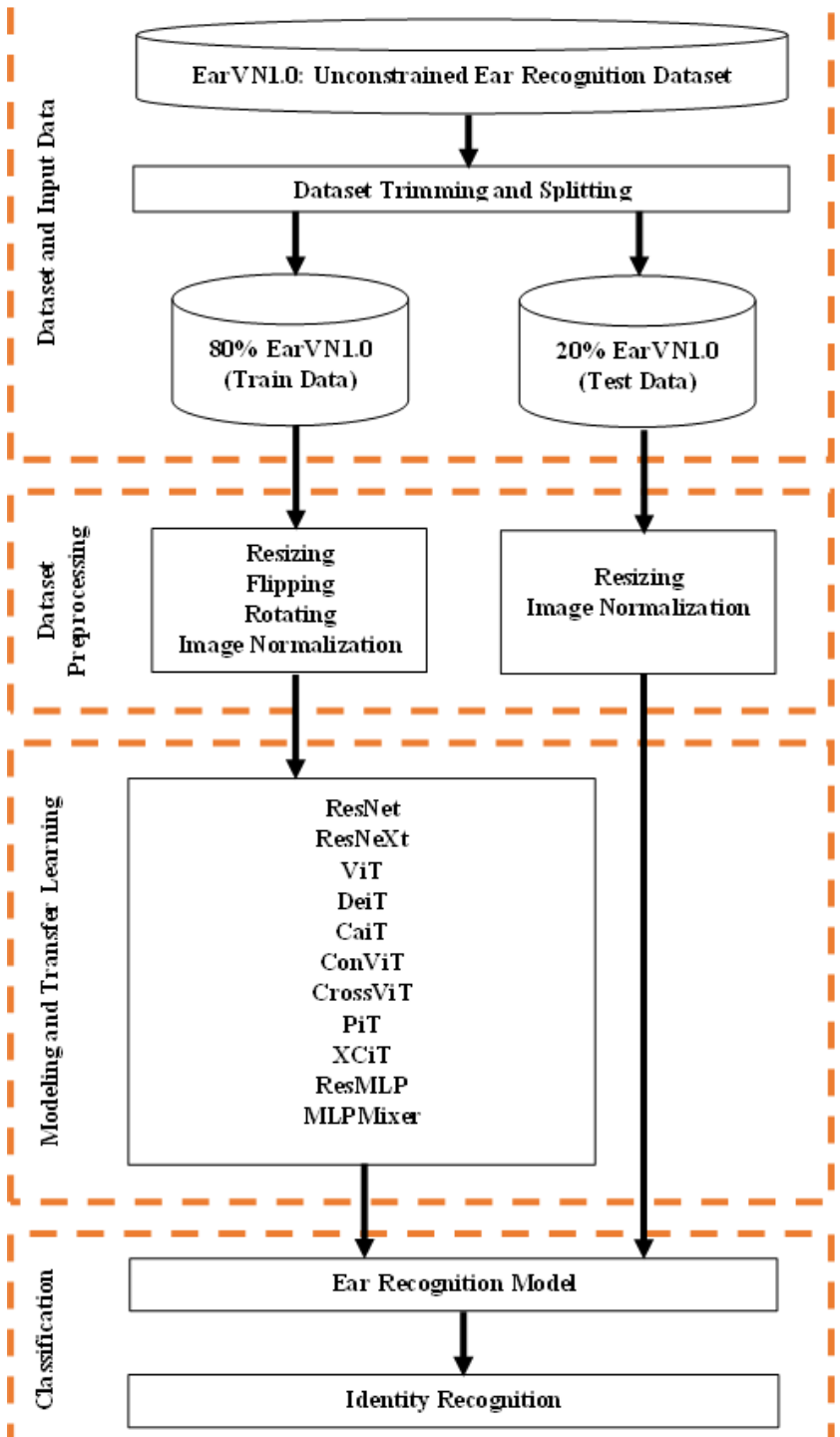
Objectives

- This study aim to:
  - Explore ear biometrics on SOTA Transformer-based and Transformer-inspired networks such as **ViT**, **DeiT**, **CaiT**, **ConViT**, **CrossViT**, **PiT, XCiT**, **Swin Transformer**, **ResMLP**, and **MLP-Mixer.**
    - Determine the performance of these models in terms of their recognition accuracy and memory utilization on ear biometrics task.
    - Provided a straightforward deep learning pipeline for ear recognition through Transfer Learning.
    - Compare the performance of these models with each other and the selected SOTA CNN models – **ResNets** and **ResNeXt**.

Note: Everyone are encouraged to read the original paper of these models to fully understand their novel features among each other.

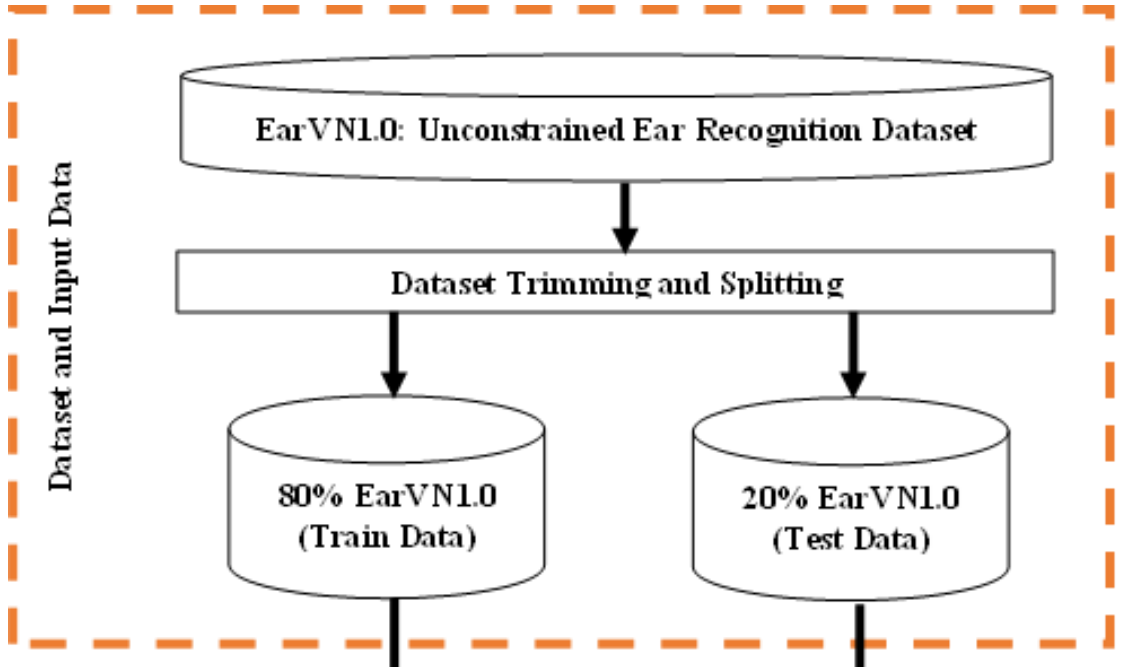
Deep Learning Pipeline

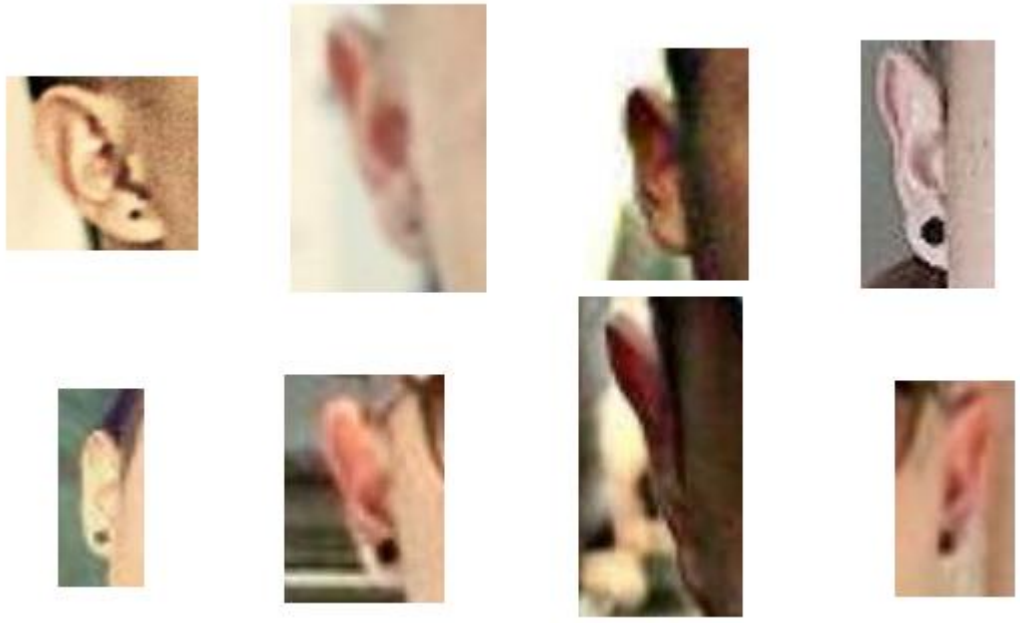
- Google Colab GPU
- PyTorch
- PyTorch Image Models (TIMM)



Deep Learning Pipeline

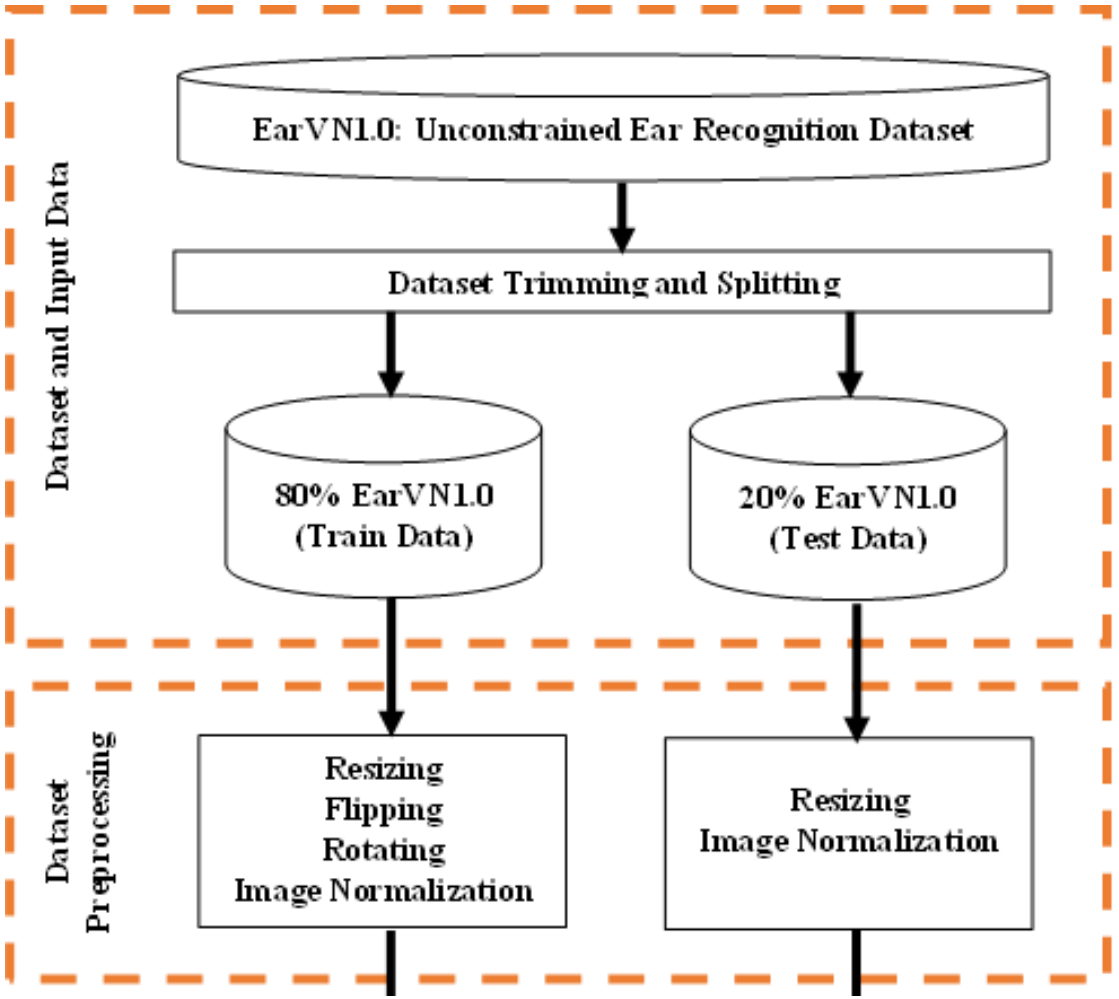
- EarVN1.0
  - World's largest collection of er images.
  - 164 individuals with each having 180 images
  - 28412 ear images.
- Data Trimming and Splitting
  - Trimmed to 20 individuals
  - Total of 4000 ear images
  - 80% Training dataset
  - 20% Testing dataset





Deep Learning Pipeline

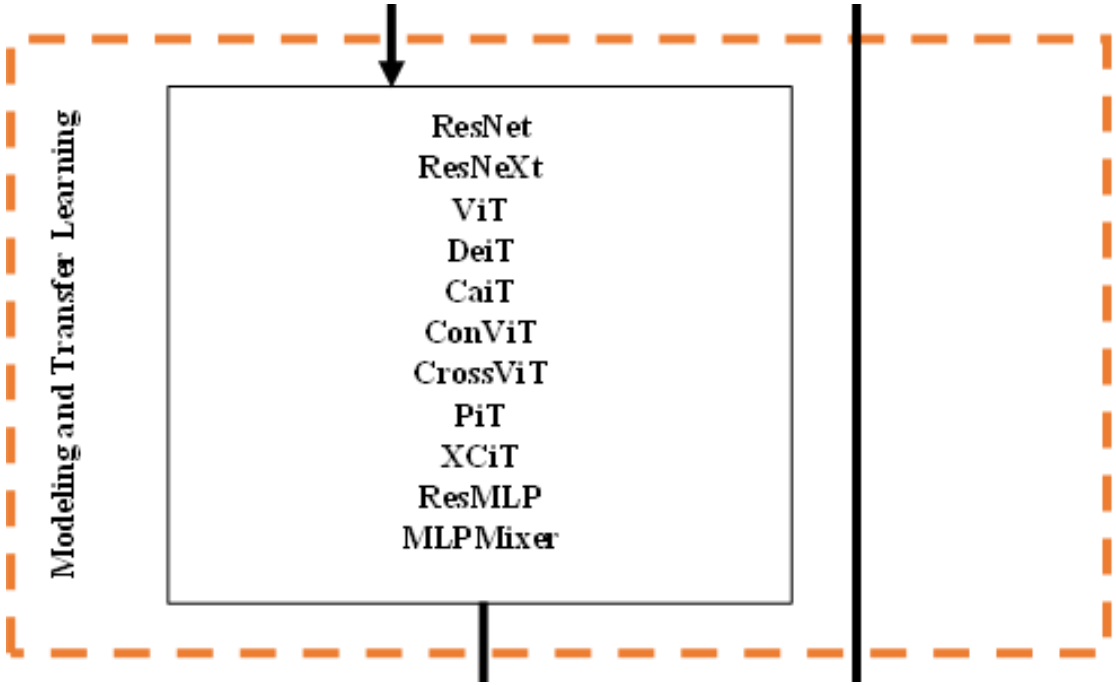
- Preprocessing
  - 3200 ear images undergone augmentation
    - Resizing
    - Flipping
    - Rotation by 30 degrees
    - Normalize using the standard ImageNet normalization values
  - 800 ear images undergone
    - Resizing
    - Image Normalization (ImageNet standard)



Deep Learning Pipeline

Table 1: Pretrained Models and Modeling Configuration

Pretrained Models	Configuration
ResNet18	Batch size: 32 Learning rate: 0.00002 Optimizer: Adam Epoch: 20
ResNet50	
ResNet152	
ResNeXt50	
ViT	
DeiT	
CaiT	
ConViT	
CrossViT	
PiT	
XCiT	
Swin Transformer	
ResMLP12	
ResMLP24	
ResMLP36	
MLP-Mixer	



Deep Learning Pipeline

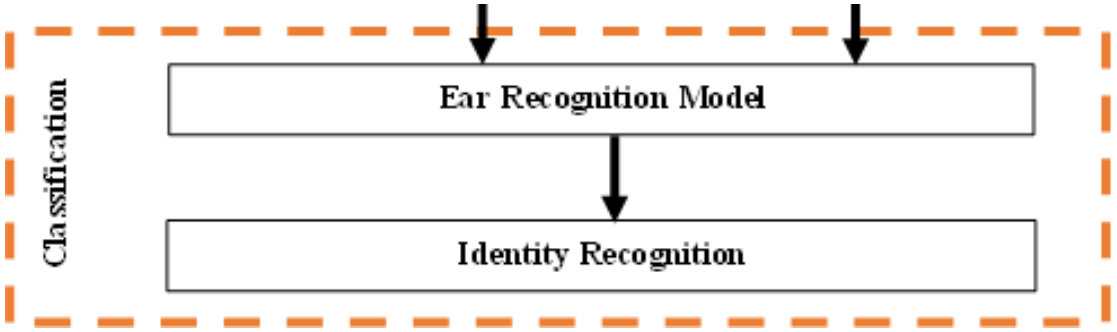


Table 2: Summary Results of ResNets

Model	Modeling Time (hours)	Recognition Accuracy (%)	Memory Utilization (GB)	Parameter Size (MB)
ResNet18	0.25	79.57	2.15	11.19
ResNet50	0.3	81	2.25	23.55
ResNet152	1.30	61.42	2.38	58.21
ResNeXt50	0.75	78.25	2.30	23.02

Table 3: Summary Results of Transformer-based Models

Model	Modeling Time (hours)	Recognition Accuracy (%)	Memory Utilization (GB)	Parameter Size (MB)
ViT	4	97.36	2.36	85.81
DeiT	3.5	95.43	2.82	85.81
ConViT	3.25	96.76	2.77	85.79
PiT	3	95.73	2.27	72.76
XCiT	0.75	75.24	2.34	2.92
CaiT	3.5	80.77	2.37	11.77
Swin Transformer	3	86.66	2.32	86.77
CrossViT	3.25	88.94	2.82	103.89

Table 4: Summary Results of Transformer-inspired Models

Model	Modeling Time (hours)	Recognition Accuracy (%)	Memory Utilization (GB)	Parameter Size (MB)
ResMLP12	3	50	1.87	14.97
ResMLP24	3	50.04	2.18	14.97
ResMLP36	3	52.16	2.24	44.31
MLP-Mixer	3	40.39	2.18	20.72

Generalization

- On the given configurations, ResNet50 performed optimally for ear biometrics.
- As ResNets become deeper, it suffers from degeneralization.
- ViT performed better than its variants regardless of modeling time and utilized memory.
- ConViT performed approximately on par with ViT.
- Despite the performance of ResMLP and MLP-Mixer with ImageNet, it suffers greatly with ear biometrics.
  - Due to the low statistics of the used dataset despite transfer learning.
  - Due to the lack of data augmentation processes.
- Overall, Transformers perform better generalization for ear biometrics over CNN and MLP.

Conclusion

- This study explored the Ear Biometrics on selected SOTA Transformer-based and Transformer-inspired models and compared each other and on CNN-based models.
  - Determined that Transformer-inspired models like ResMLP and MLP-Mixer performed lesser on Ear Biometrics over Transformer-based models, specially ViT.
  - Provided a deep learning pipeline with Transfer Learning in its core.
  - Compared the performance of these models with each other and on CNN-based models and also determined that Transformer still outperform ResNets/ResNeXt.
- For future, consider ConvNeXt for ear biometrics.

End-of-Presentation

- For inquiries:

  mbalejo.cpe@gmail.com (default)

  marwin.alejo@eee.upd.edu.ph

  mbalejo1@up.edu.ph

- Codes and other materials are in my Github:

  https://github.com/mbalejo/CS284/tree/main/MiniProject