UNIVERSIDADE FEDERAL DE OURO PRETO
INSTITUTO DE CIÊNCIAS EXATAS E BIOLÓGICAS
DEPARTAMENTO DE COMPUTAÇÃO

MATEUS PEVIDOR REIS
Orientador: Guillermo Cámara Chávez

RECONHECIMENTO DE PESSOAS ATRAVÉS DA BIOMETRIA DA FACE E DA ORELHA

Ouro Preto, MG
2023

UNIVERSIDADE FEDERAL DE OURO PRETO
INSTITUTO DE CIÊNCIAS EXATAS E BIOLÓGICAS
DEPARTAMENTO DE COMPUTAÇÃO

MATEUS PEVIDOR REIS

RECONHECIMENTO DE PESSOAS ATRAVÉS DA BIOMETRIA DA FACE E DA ORELHA

Monografia apresentada ao Curso de Ciência da Computação da Universidade Federal de Ouro Preto como parte dos requisitos necessários para a obtenção do grau de Bacharel em Ciência da Computação.

Orientador: Guillermo Cámara Chávez

Ouro Preto, MG
2023

SISBIN - SISTEMA DE BIBLIOTECAS E INFORMAÇÃO

R375r Reis, Mateus Pevidor.
Reconhecimento de Pessoas Através da Biometria da Face e da Orelha. [manuscrito] / Mateus Pevidor Reis. - 2023.
64 f.: il.: color., gráf., tab..

Orientador: Prof. Dr. Guillermo Cámara Chávez.
Monografia (Bacharelado). Universidade Federal de Ouro Preto. Instituto de Ciências Exatas e Biológicas. Graduação em Ciência da Computação .

1. Biometria. 2. Identificação. 3. Redes neurais (computação). 4. Reconhecimento facial. I. Cámara Chávez, Guillermo. II. Universidade Federal de Ouro Preto. III. Título.

CDU 004

Bibliotecário(a) Responsável: Luciana De Oliveira - SIAPE: 1.937.800

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
INSTITUTO DE CIENCIAS EXATAS E BIOLOGICAS
DEPARTAMENTO DE COMPUTACAO

FOLHA DE APROVAÇÃO

Mateus Pevidor Reis

Reconhecimento de pessoas através da biometria da face e da orelha

Monografia apresentada ao Curso de Ciência da Computação da Universidade Federal
de Ouro Preto como requisito parcial para obtenção do título de Bacharel em Ciência da Computação

Aprovada em 23 de Março de 2023.

Membros da banca

Guillermo Cámara Chávez (Orientador) - Doutor - Universidade Federal de Ouro Preto
Rafael Alves Bonfim de Queiroz (Examinador) - Doutor - Universidade Federal de Ouro Preto
Pedro Henrique Lopes Silva (Examinador) - Doutor - Universidade Federal de Ouro Preto

Guillermo Cámara Chávez, Orientador do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 23/03/2023.

Documento assinado eletronicamente por **Guillermo Camara Chavez**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 27/03/2023, às 13:33, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **0493047** e o código CRC **C5244DAC**.

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35402-163
Telefone: 3135591692 - www.ufop.br

Agradecimentos

Primeiramente, gostaria de agradecer a Deus, por me guiar e me dar força e perseverança para concluir este trabalho. Sem Sua graça, eu não teria conseguido alcançar este marco em minha jornada acadêmica. Através dos altos e baixos, Deus permaneceu fiel, e por isso sou profundamente grato.

Eu gostaria de estender meus agradecimentos aos meus pais, que sempre foram uma fonte de amor e apoio incondicional. Foram meus pilares de sustentação desde o início de minha jornada acadêmica. Eles fizeram inúmeros sacrifícios para me fornecer um ambiente propício para o aprendizado. Seu encorajamento inabalável, sacrifícios e orações foram fundamentais para moldar a pessoa que sou hoje. Sou eternamente grato por tudo que fizeram por mim.

Agradeço também ao meu irmão pelo constante apoio e motivação ao longo do meu percurso acadêmico. Sua crença em minhas habilidades tem sido uma fonte de inspiração para eu buscar a excelência.

Gostaria de agradecer aos meus amigos e colegas, que sempre foram uma fonte constante de apoio e incentivo. Sua camaradagem e vontade de compartilhar seus conhecimentos e experiências foram inestimáveis para enriquecer minha pesquisa.

Sou profundamente grato ao meu orientador, Guillermo Cámara Chávez, por sua inestimável orientação, apoio e mentoria ao longo da duração deste projeto. Seu amplo conhecimento, percepções críticas e feedback construtivo foram fundamentais para moldar a direção de minha pesquisa e me ajudar a atingir meus objetivos.

Por fim, gostaria de expressar minha gratidão a todos os meus professores, que desempenharam um papel significativo no meu desenvolvimento acadêmico e pessoal. Sua paixão pelo ensino, sua dedicação às suas respectivas áreas e sua vontade de ir além para ajudar seus alunos têm sido uma fonte constante de inspiração para mim.

Resumo

A biometria refere-se aos métodos automatizados de identificação de indivíduos com base em suas características físicas ou comportamentais únicas. Essas características podem incluir impressões digitais, características faciais, padrões de íris, impressões de voz e até traços comportamentais, como ritmo de digitação ou a forma de andar. Os sistemas de identificação biométrica funcionam capturando essas características únicas e comparando-as com um banco de dados pré-existente de características conhecidas para autenticar a identidade do indivíduo. Os métodos biométricos tornaram-se cada vez mais populares nos últimos anos devido à sua capacidade de identificar indivíduos com precisão e segurança com base nessas características. Diferentes métodos de extração de características são utilizados nessa área, tais como Histograma de Gradientes Orientados - *Histogram of Oriented Gradients* (HOG), *Support Vector Machine* (SVM) e Rede Neural Convolucional - *Convolutional Neural Network* (CNN), como também o avanço recente na visão computacional, os *Vision Transformer* (ViT). Portanto, neste trabalho é explorado o uso de *deep learning*, utilizando como extratores de características CNNs e ViTs com o objetivo de realizar o reconhecimento de indivíduos a partir de imagens de sua face e orelha. São propostos dois modelos que fazem o uso de três classificadores distintos, sendo um para classificar o indivíduo a partir da face, outro para classificá-lo a partir da orelha, e um último, este sendo um Multi-layer Perceptron (MLP), para combinar os resultados dos anteriores em uma única classificação. Um dos modelos, utilizando CNNs como extrator de características, e o outro utilizando ViTs. Para testar a eficácia das composição dos modelos, foram conduzidos experimentos computacionais, utilizando diferentes conjuntos de dados que foram obtidos tanto em ambientes controlados quanto em ambientes não controlados. Os conjuntos de dados obtidos em ambiente controlado usados nesses experimentos incluem os conjuntos de dados AMI Ear e FEI Face, enquanto os conjuntos de dados obtidos em ambiente não controlado incluem os conjuntos de dados VGGFace-Ear e LFW Face. As amostras dos conjuntos de dados de faces e de orelhas foram pareados um a um, gerando uma base de dados sintética, de modo que pares de faces e orelhas componham um indivíduo. Quando treinado com o conjunto de dados controlados, o modelo baseado em CNNs obteve a acurácia de 100%, e também de 100% quando treinado com o conjunto de dados não controlados. O modelo baseado em ViTs, treinado apenas no conjunto de dados não controlados, obteve acurácia de 100% no conjunto de dados controlados. Diante dos resultados obtidos, foi possível concluir que a combinação dos resultados de múltiplos modelos por meio de uma rede neural é uma prática válida, além de que o uso de ViTs pode trazer melhores resultados em relação ao uso de CNNs.

Palavras-chave: Biometria. Identificação. Reconhecimento facial. Reconhecimento da orelha. Deep learning. Redes neurais. Vision Transformers.

Abstract

Biometrics refers to automated methods for individual identification based on their unique physical or behavioral characteristics. These characteristics may include fingerprints, facial characteristics, iris patterns, voice impressions and even behavioral traits, such as typing rhythm or walking manner. Identification systems work by capturing these unique characteristics and comparing them with a pre existing database of known characteristics to authenticate the individual's identity. Biometric methods have become more popular on the past years due to their capacity of identify individuals with accuracy and security based on those characteristics. Different feature extraction methods are used in this area, such as Histograma de Gradientes Orientados - *Histogram of Oriented Gradients* (HOG), *Support Vector Machine* (SVM), and Rede Neural Convolucional - *Convolutional Neural Network* (CNN), aswell as the recent computational vision advance, *Vision Transformer* (ViT). Therefore, this work explores the use of *deep learning*, using CNNs e ViTs as feature extractors, in order to perform individual recognition based on their face and ear. Two models are proposed that make use of three distinct classifiers, one for facial recognition, other for ear recognition, and the last, as a Multi-layer Perceptron (MLP), for combining the results of the later two into a unique classification. One of the models, using CNNs as a feature extractor, and the other using ViTs. In order to test the efficiency of the models' composition, computational experiments were conducted using different datasets with data that were obtained in both constrained and unconstrained environments. The constrained datasets used in the experiments include the AMI Ear dataset and the FEI Face dataset, while the unconstrained datasets include the VGGFace-Ear and the LFW datasets. The samples of the ear and face datasets were paired one by one, generating a synthetic database, so that ear and face pairs make up a subject. When trained with the constrained datasets, the CNN-based model achieved 100% accuracy, and also 100% accuracy when trained with the unconstrained datasets. The ViT-based model, trained only on the unconstrained dataset, achieved 100% accuracy. In view of the obtained results, it was possible to conclude that the combination of results of multiple models by using a neural network is a valid practice, in addition to the fact that the use of ViTs can bring better results in comparison to the use of CNNs.

Keywords: Biometrics, Identification, Facial recognition. Ear recognition. Deep learning. Neural Networks. Vision Transformers.

Lista de Ilustrações

Lista de Tabelas

Lista de Acrônimos

ANN Rede Neural Artificial - *Artificial Neural Network*. 3, 9, 10, 23

CLAHE Contrast Limited Adaptative Histogram Equalization. 6
CNN Rede Neural Convolucional - *Convolutional Neural Network*. iii, iv, 3, 8–10, 12, 13, 19–24, 37, 43, 44, 46, 47

DL Aprendizado Profundo - *Deep Learning*. 8, 10, 26

HOG Histograma de Gradientes Orientados - *Histogram of Oriented Gradients*. iii, iv, 1, 8, 18

ICP *Iterative Closest Point*. 17, 18

KNN K-Nearest Neighbors. 47

LFW Labeled Faces in the Wild. v, 29, 31
LPE-ViT Learnable Position Embeddings. 21

MLP Multi-layer Perceptron. iii, iv, 13, 24
MST Masked Self-Supervised Transformer. 21

PCA *Principal Component Analysis*. 7, 8, 17–19

ReLU *Rectified Linear Units*. 10

SGDM Stochastic Gradient Descent with Momentum. 20
SIFT *Scale-Invariant Feature Transform*. 6
SURF *Speeded Up Robust Features*. 8
SVM *Support Vector Machine*. iii, iv, 47

T2T-ViT Tokens-to-Token ViT. 21

ViT *Vision Transformer*. iii–v, 3, 8, 13, 14, 21, 24, 37, 44–47

Sumário

1 Introdução

A biometria é o processo de identificação de características e comportamentos de seres humanos que nos possibilita distinguirmos-nos uns dos outros. Somos capazes de perceber uma série de traços físicos de outras pessoas, remetendo-nos à identidade do outro. Realizar esse mesmo processo de forma automática utilizando sistemas computacionais de forma precisa é um desafio que vem sendo enfrentado durante as últimas décadas (JAIN et al., 2004).

Realizar a identificação de um indivíduo de forma automática se torna útil em diversas situações. É possível notar o uso desse recurso em controladores de acesso, como no desbloqueio de dispositivos eletrônicos (PATEL; HAN; JAIN, 2016) ou na liberação de acesso a um espaço físico, monitoramento de pessoas em câmeras de vigilância (XU, 2021), identificação de suspeitos na área criminal, autenticação de documentos (WAYMAN et al., 2005), entre outros.

Nessas diversas aplicações, diferentes formas de biometria podem ser utilizadas. A escolha de qual método utilizar depende de alguns fatores (SHARIF et al., 2019):

- A quantidade de dados disponíveis tanto para o cadastro do indivíduo quanto para o processo de reconhecimento é um fator que pode ser determinante para a confiabilidade do algoritmo. Quando poucas imagens são alimentadas ao algoritmo na fase de treinamento ou na fase de reconhecimento, pode ser esperado que a confiança do algoritmo ao reconhecer uma pessoa seja baixa.

- A má qualidade das imagens utilizadas, tanto no quesito de iluminação e resolução quanto no quesito da disposição do indivíduo, como pose ou obstrução das características pelo uso de acessórios, impacta negativamente no desempenho do algoritmo.

- A forma de obtenção dos dados a serem utilizados varia de acordo com o método de biometria a ser utilizado. Alguns métodos demandam que a obtenção seja de forma mais intrusiva, fazendo com que o contato físico do indivíduo seja necessário, como ao cadastrar uma digital ou ao assinar algum documento com o próprio punho. Já em outros métodos, a obtenção se dá de forma não intrusiva, como acontece na captura de uma imagem por meio de uma câmera. Nota-se, então, que os métodos intrusivos necessitam que o indivíduo colabore com a obtenção dos dados, o que nem sempre é possível.

Neste trabalho será estudado o método de biometria por meio da combinação do reconhecimento através da face e da orelha. Esses dois métodos são estudados individualmente há décadas, de forma que muitas técnicas de reconhecimento já foram experimentadas, como o uso de modelos 3D da orelha (YAN; BOWYER, 2005) e o uso do HOG, por (DAMER; FÜHRER, 2012). Ambos permitem a obtenção de dados de forma não intrusiva. Entretanto, nem sempre é

possível obter imagens de boa qualidade, além de apresentarem aspectos que podem dificultar o reconhecimento.

O reconhecimento facial é possível graças ao fato de que a face, construída biologicamente, apresenta características únicas nas regiões dos olhos, boca e nariz, tornando a face de cada pessoa praticamente única. Porém, alguns fatores podem dificultar o processo de reconhecimento por esconderem algumas dessas características (CÁRDENAS; BELTRÁN; GUTIÉRREZ, 2019; DADI; PILLUTLA; MAKKENA, 2018). A simples obstrução de alguma região da face pode atrapalhar no processo de reconhecimento, como o uso de acessórios tais como óculos, máscaras ou bonés. Essas características também podem ser modificadas por conta da iluminação do ambiente, expressão facial do indivíduo, a simples mudança de posição na imagem ou, principalmente, o envelhecimento. Dessa forma, Singh et al. (2020) mencionam que o desempenho de algoritmos de reconhecimento facial são medidos, normalmente, por meio da validação utilizando imagens que contém essas anomalias, estimando, assim, sua robustez.

A orelha, similarmente à face, possui características que são únicas de pessoa para pessoa. A obstrução dessas características ainda ocorre com o uso de acessórios, além, também, das mudanças de iluminação e posição no momento da captura da imagem. Por outro lado, a orelha não sofre tantas alterações como a face, em que não conseguimos alterar o seu formato voluntariamente. Um ponto importante apontado por Iannarelli (1989) em seu livro é que a orelha sofre alterações em ritmo linear apenas nos primeiros quatro meses de vida e, dos oito anos aos setenta, permanece constante. Dessa forma, a orelha acaba não sendo significativamente influenciada pelo fator envelhecimento como ocorre com a face.

Apesar do reconhecimento da orelha parecer mais promissor por não apresentar tantos empecilhos em seu processo em relação à face, as características únicas de cada orelha não são facilmente identificadas pelos algoritmos (MAHOOR; CADAVID; ABDEL-MOTTALEB, 2009), de forma que os resultados apresentados em outros estudos sejam menos precisos do que os de reconhecimento facial. Sabendo disso, este trabalho abordará sobre cada um dos métodos de biometria individualmente e, depois, realizará a fusão dos resultados de cada um dos modelos em um único, esperando, assim, resultados preferíveis.

Na abordagem desse tema, serão utilizadas técnicas de aprendizado em profundidade, contrapondo às técnicas de *handcrafted features* (GEORGESCU; IONESCU; POPESCU, 2019). Trabalhos da última década que optaram por essa decisão apresentaram melhores resultados, como é mostrado no capítulo 3, motivando essa escolha. É apresentada no capítulo de metodologia uma proposta de *pipeline* de processamento para realizar a união das tarefas de reconhecimento facial e reconhecimento da orelha para que se obtenha um classificador para o reconhecimento de pessoas.

1.1 Justificativa

O reconhecimento facial é uma área de estudo que recebe muita atenção devido a suas diversas utilidades. O ser humano, capaz de reconhecer padrões em todas as coisas, consegue identificar a face de uma pessoa com facilidade. Transferir esse tipo de inteligência para uma máquina é uma tarefa que tem sido estudada por muitos anos, e que tem apresentado bons resultados, como é dito na pesquisa de (FU et al., 1976), publicada em 1976, que a pesquisa e o desenvolvimento da área de reconhecimento de padrões e processamento de imagens tem tomado lugar há 20 anos. Os melhores algoritmos para o reconhecimento de faces se baseiam, principalmente, em características presentes na região dos olhos, nariz e boca.

A pandemia do vírus Covid-19 causou inúmeros distúrbios na sociedade, mudando a forma que vivemos, sejam nos nossos costumes, nossas relações com outras pessoas ou nossa forma de enxergar o mundo. Uma das medidas preventivas para a disseminação do vírus foi adotar o uso de máscara em locais públicos, cobrindo uma grande parcela do rosto, incluindo o nariz e a boca. A partir disso, Ge et al. (2017) apontam que até mesmo algoritmos considerados estado da arte para a detecção de faces perdem precisão devido ao uso de máscaras. Com isso, algoritmos de reconhecimento de pessoas enfrentam grandes dificuldades ao identificar pessoas de máscara, já que boa parte das características que eles utilizam no processo ficam obstruídas.

A partir disso, fazer o uso de outras características corporais no reconhecimento se torna necessário. A orelha, portanto, é um ótimo candidato, já que suas características não são obstruídas pelo uso de máscara e a obtenção de imagens da orelha ocorre de forma não intrusiva, assim como da face. Além disso, a colaboração do indivíduo para a obtenção da imagem da orelha não é tão necessária quanto a da face, já que seres humanos não têm um grande controle sobre os músculos da orelha.

1.2 Objetivos

Este trabalho propõe o uso de técnicas recentes de reconhecimento de padrões para sintetizar uma solução para o reconhecimento de pessoas através da face e da orelha. Para atingir esse objetivo, será utilizada a técnica de *deep learning* sobre um *dataset* com imagens da face e da orelha, fazendo uso de redes neurais para realizar o reconhecimento das pessoas. Dessa forma, os objetivos específicos a serem atingidos são:

1. Implementar uma CNN para o reconhecimento a partir da face.
2. Implementar outra CNN para o reconhecimento a partir da orelha.
3. Implementar uma Rede Neural Artificial - *Artificial Neural Network* (ANN) para a fusão dos resultados das outras duas redes.
4. Substituir as duas CNNs por redes baseadas em ViTs.

1.3 Organização do Trabalho

Os seguintes capítulos deste trabalho se encontram na seguinte ordem: o Capítulo 2 expõe conceitos relevantes sobre a área que são necessários para a compreensão deste trabalho. O Capítulo 3 apresenta a revisão de literatura sobre o tema abordado, exibindo as estratégias e técnicas aplicadas em outros estudos. O Capítulo 4 apresenta a metodologia utilizada de forma detalhada, exibindo cada uma das etapas realizadas até a implementação final do algoritmo, juntamente de explicações sobre cada uma das decisões tomadas ao longo do processo. Os experimentos realizados juntamente dos resultados obtidos são apresentados e discutidos no Capítulo 5, seguidos pelas conclusões obtidas, no Capítulo 6.

2 Fundamentação teórica

As técnicas para o reconhecimento de pessoas a partir de suas características corporais mudaram consideravelmente nos últimos 30 anos. Em 1993, o Departamento de Defesa dos Estados Unidos deu origem ao programa FERET (RAUSS et al., 1997), que fomentou a pesquisa no assunto de reconhecimento de faces para auxiliar na segurança por meio do reconhecimento de indivíduos a partir de fotografias.

2.1 Pré-processamento

Dado que as diversas fotografias utilizadas para reconhecimento são retiradas em diferentes condições, como em diferentes ambientes, ângulos e direções, é necessário que haja uma etapa de pré-processamento das imagens para que elas se tornem adequadas para a fase de processamento. Esta etapa pode ser compreendida por procedimentos 100% automáticos ou podem contar também com procedimentos manuais (VICTOR; BOWYER; SARKAR, 2002). Pode ser necessário que sejam realizados tratamentos na imagem, como ajuste de brilho e contraste ou centralização e mascaramento da região de interesse.

Dharavath, Talukdar e Laskar (2014) demonstraram quatro diferentes passos a serem realizados no pré-processamento de imagens que serão utilizadas no reconhecimento facial, visando melhorar os resultados:

- O recorte da face é importante para que apenas características necessárias estejam presentes na imagem. Dessa forma, podem ser utilizadas técnicas de detecção de face para determinar quais são os limites do rosto para que o recorte seja feito.
- O redimensionamento da imagem utilizada também é importante para que o tamanho das imagens de entrada seja homogêneo. Durante o redimensionamento, foi utilizada a interpolação pelo vizinho mais próximo.
- A iluminação, sendo um fator de grande importância durante o reconhecimento, deve manter também um padrão entre as imagens para se obter melhores resultados. Assim, é necessária a normalização da distribuição dos níveis de intensidade da imagem. Para isso, foi utilizada a técnica de equalização de histograma.
- A eliminação de ruídos por meio de filtros contribui para eliminar artefatos presentes da imagem. Os autores utilizam um filtro passa-baixa nessa etapa de filtragem.

Esses procedimentos do pré-processamento também podem ser aplicados em imagens no reconhecimento de orelhas. Nigam e Gupta (2014) desenvolveram um modelo que faz uso

de técnicas de processamento de imagens, como o algoritmo de detecção de bordas proposto por Canny (1986), e do ajuste de contraste Contrast Limited Adaptative Histogram Equalization (CLAHE) (ZUIDERVELD, 1994), com o objetivo de realizar a detecção da orelha de forma automática.

2.2 Extração de características

O processo de extração de características de uma imagem é uma etapa que ocorre durante o reconhecimento, que tem o objetivo de separar componentes importantes para a tarefa a ser realizada, reduzindo a dimensão do problema em questão. No contexto de reconhecimento facial, as características a serem extraídas são, geralmente, olhos, nariz e boca. Já no reconhecimento da orelha, as características que diferenciam umas das outras são as posições e os formatos de suas diferentes partes.

Dentre as diversas técnicas de extração de características, muitas delas são utilizadas tanto para o reconhecimento facial quanto para o reconhecimento da orelha. Ghoualmi, Draa e Chikhi (2016), em seu estudo, fazem uso do algoritmo de extração de características *Scale-Invariant Feature Transform* (SIFT) para extrair as características mais discriminantes da orelha. Da mesma forma, Geng e Jiang (2009) fizeram o uso do mesmo algoritmo, porém para extrair características da face.

2.2.1 Extração de características da orelha

Apesar de ser possível intercalar algumas técnicas entre o reconhecimento facial e o reconhecimento da orelha, muitos estudos já foram realizados com o objetivo de extrair características especificamente da orelha. Ghoualmi, Draa e Chikhi (2016) elencam diversos trabalhos com esse propósito, que fazem uso de imagens 2D ou de modelos 3D da orelha. Em cenários onde são utilizadas imagens 2D, as técnicas de extração podem ser divididas em duas categorias: abordagem geométrica e abordagem global.

A abordagem geométrica trabalha com detecção de bordas para rastreamento e detecção do contorno, pontos de referência para resolver o problema de imagens rotacionadas ou estabelecimento de marcos para imagens com variações de posição e iluminação. Como é possível perceber na Figura 2.1, as diferentes regiões da orelha são facilmente distinguíveis quando levamos em consideração suas bordas:

Figura 2.1 – Anatomia da orelha humana com identificação das suas partes. (UOL, 2015)

Já nas abordagens globais, algumas técnicas foram citadas pelos autores estudados no capítulo 3, como a de PCA, *Active shape models* (ASMs) e *Force field transform* (FFT). O PCA é uma técnica que tem como objetivo diminuir a dimensão do problema trabalhado (VICTOR; BOWYER; SARKAR, 2002). Ou seja, apenas informações relevantes são mantidas para a etapa de classificação. A técnica ASMs foi utilizada por Lu et al. (2006), em que ele é descrito como um método estatístico poderoso, mas que sofre com variações de iluminação na definição dos *landmarks*. O método de extração de características FFT foi desenvolvido e demonstrado por Hurley, Nixon e Carter (2000), se baseando na ideia de várias partículas que são colocadas na imagem e são atraídas por regiões específicas, como pelo contorno da orelha. Dessa forma, a trajetória formada pelas partículas são tratadas como canais de fluxo, e o pontos de concentração são tratados como poços de energia. A partir dessas informações é possível estabelecer marcadores para realizar a extração.

Em 2006, Sana, Gupta e Purkait (2007) introduziram uma nova técnica para realizar a extração de características da orelha, fazendo uso da transformada de Haar. Essa técnica faz uso da transformada para obter os coeficientes que são agrupados em coeficientes horizontais, verticais, diagonais e de aproximação. Com isso, a imagem é decomposta, possibilitando que redundâncias sejam detectadas e removidas. Esse procedimento é executado em quatro níveis, onde se nota que os coeficientes do quarto e do primeiro níveis são iguais, em que, por fim, apenas os do quarto nível são utilizados para representar a orelha.

Se tratando de modelos 3D da orelha, alguns trabalhos são citados por (GHOUALMI; DRAA; CHIKHI, 2016). Dentre estes, técnicas como *Iterative Closest Point* (ICP) e o detector *AdaBoost* são utilizados, às vezes em conjunto, para realizar a extração de características. Alguns autores optaram por utilizar algoritmos próprios para trabalhar com os modelos que, de forma geral, funcionam através de casamento de padrão de características geométricas. É também dito

que os modelos 3D eliminam alguns problemas presentes no processo de extração em imagens, e que o fato de que o modelo pode ser transformado geometricamente, além de ter uma terceira dimensão, pode aprimorar a precisão do sistema de reconhecimento.

Estudos mais recentes fazem uso de métodos de Aprendizado Profundo - *Deep Learning* (DL), dada a capacidade de realizar o reconhecimento de padrões em dados brutos. Priyadharshini, Arivazhagan e Arun (2021) utilizam, em seu estudo, CNN para realizar tanto o reconhecimento de padrões na imagem quanto a classificação das imagens de orelhas e Han et al. (2021) propuseram uma nova arquitetura de *Vision Transformers* (ViTs) para o reconhecimento facial. Por meio dessas técnicas, foram obtidos resultados superiores a outros métodos, principalmente de *Handcrafted Features*, citados em diferentes bases de imagens experimentadas, como no trabalho de (DAMER; FÜHRER, 2012), que fazem uso de HOG, e de (XU; MU; YUAN, 2007), que fazem uso do PCA.

2.2.2 Extração de características da face

As técnicas extração de características da face são amplamente estudadas, dado que de todas as partes do nosso corpo que podem ser capturadas por câmeras, a face é a parte que melhor nos identifica. Shoba e Sam (2020) fazem uso, em seu trabalho, de três técnicas diferentes em conjunto para realizar a extração de características da face. Seu objetivo ao juntar técnicas distintas é de aplicá-las a regiões específicas da face. O algoritmo de extração de características *Speeded Up Robust Features* (SURF) é utilizado para obter aspectos da face como um todo. São obtidas informações da região compreendida pelos olhos, ponte nasal, nariz e boca (ver Figura 2.2). Com isso, mudanças na posição da cabeça se tornam menos impactantes no momento de classificação. O algoritmo HOG é utilizado na região da ponte nasal, já que ele trabalha com o gradiente da região da imagem. Essa região da face apresenta poucas alterações com o envelhecimento do indivíduo, se mostrando confiável para o reconhecimento (IANNARELLI, 1989). O terceiro algoritmo utilizado é o *Maximally Stable Extremal Regions* (MSER), que é aplicado nas regiões dos olhos, nariz e boca, que apresentam grandes variações dependendo da expressão demonstrada na imagem. Dessa forma, se torna necessário obter as informações que são estáveis nessas regiões.

Outras técnicas de extração de características podem ser utilizadas no mesmo contexto. As próprias características de Haar, mencionada anteriormente, foram utilizadas por Tathe, Narote e Narote (2016), juntamente de outras técnicas, como as características Eigenface e Gabor. Em sua pesquisa consta que as características de Haar apresentam bons resultados quando o indivíduo se posiciona de frente para a câmera. Também consta que o tempo de detecção das características de Gabor podem ser consideradas muito longo quando se trata de realizar o processamento de uma imagem para realizar a detecção. Dessa forma, dependendo da aplicação, como em detecções em tempo real, esse método pode não ser adequado.

Assim como na extração de características da orelha, a técnica de DL pode ser utilizada também para realizar essa tarefa no contexto de faces. Xu (2021) realizou testes de performance

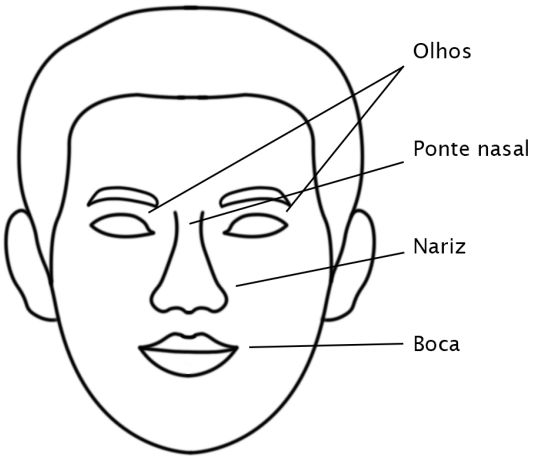

Figura 2.2 – Diagrama com as regiões da face importantes para o reconhecimento facial.

em seus experimentos com diferentes redes neurais convolucionais, mostrando que a precisão, muitas vezes, é obtido a custo do aumento do tempo de processamento.

2.3 Redes neurais convolucionais

Redes Neurais Convolucionais - *Convolutional Neural Networks* (CNNs) são um tipo de ANN. Sua característica é fazer uso de operações de convolução sobre dados utilizando filtros, extraindo características da imagem de entrada. A arquitetura de CNNs são baseadas em camadas, em que são realizadas diferentes operações, como convolução, subamostragem, preenchimento, entre outras. Ao fim são utilizadas camadas totalmente conectadas, permitindo a interpretação e classificação da entrada.

Dessa forma, CNNs são arquitetadas de modo a seguir um modelo de camadas geral, alterando apenas os hiperparâmetros de cada camada. Esse modelo é ilustrado na Figura 2.3. Ele é composto por vários grupos de camadas de convolução, em que é executada a convolução, a função de ativação e a subamostragem, nessa ordem. Ao fim, é realizada a vetorização e a função de ativação *softmax* para estabelecer a probabilidade de cada classe. Mais detalhes sobre cada uma das etapas serão discutidos adiante.

2.3.1 Redes neurais artificiais

As Redes Neurais Artificiais - *Artificial Neural Networks* (ANNs) tradicionais são formadas por camadas de neurônios que assumem valores em função do estímulo recebido pela entrada ou pelas camadas anteriores. A primeira camada é chamada de camada de entrada (*input layer*), as seguintes de camadas escondidas (*hidden layer*), e a última é a camada de saída (*output layer*). As ligações entre os neurônios das diferentes camadas têm pesos que são levados em

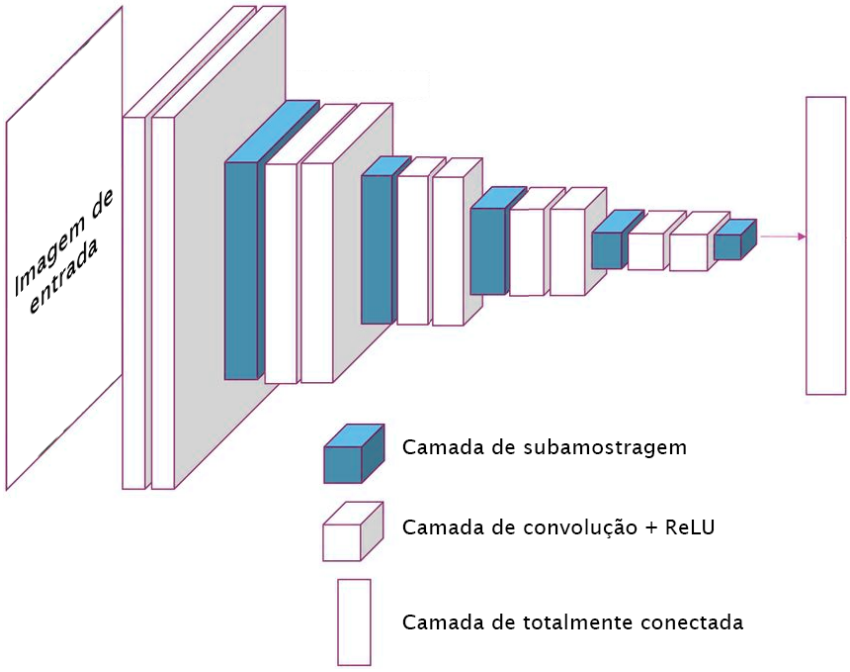

Figura 2.3 – Visualização dos componentes de uma rede neural convolucional.

consideração no cálculo do valor de cada um deles. Além dos pesos, também existe um viés que permite a melhor generalização da rede.

O valor de cada neurônio é calculado pela multiplicação do estímulo de entrada com o peso, somado ao viés. Esse resultado é processado por uma função de ativação, que quando esta é uma função não linear, permite que a rede molde problemas não linearmente separáveis. Uma das funções mais utilizadas é a *Rectified Linear Units* (ReLU), que geralmente faz com que o modelo convirja mais rapidamente.

Em DL, principalmente com base de dados pequenos, o sobreajuste (*overfitting*) pode ser um problema. De acordo com Srivastava et al. (2014), a melhor maneira de evitar esse problema é treinar vários modelos e obter a média das predições deles. Porém, computacionalmente isso é muito custoso e, consequentemente, impraticável. Dessa forma, para regularizar uma rede, pode ser utilizada a estratégia de *dropout*, que é desligar alguns neurônios aleatoriamente durante o processo de treinamento. Isso resulta na melhor capacidade de generalização da rede.

2.3.2 Camada totalmente conectada

As CNNs que realizam o processo de classificação têm ao fim de sua arquitetura camadas densas de neurônios totalmente conectados. Seu funcionamento é equivalente às ANNs tradicionais, com a entrada igual ao número de características extraídas na camada anterior. Dessa forma, é necessário realizar a conversão delas em um vetor de características (*Flatten*). A última camada totalmente conectada usualmente tem o número de neurônios igual ao número de classes do problema, em que o sinal de cada neurônio equivale à probabilidade da imagem pertencer a

aquela classe. Para obter esse resultado, é utilizada a função de ativação *Softmax*.

2.3.3 Camada de convolução

A camada de convolução realiza a operação de convolução sobre a imagem de entrada utilizando um filtro. Seu objetivo é evidenciar características presentes na imagem, como na detecção de padrões e bordas. A convolução é o processo de gerar uma matriz de valores baseado na matriz de valores da imagem original. Para cada conjunto de pixels do tamanho do filtro, os pixels da imagem são multiplicados pelos valores do filtro, e o elemento na posição central do filtro assume o valor igual à soma dos resultados das multiplicações. Esse processo é exemplificado na Figura 2.4.

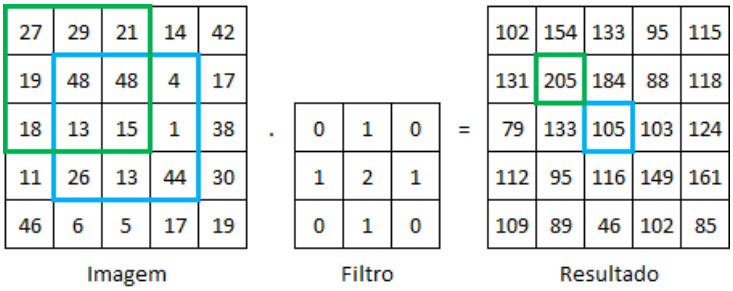

Figura 2.4 – Exemplo de aplicação da operação de convolução em uma matriz.

Pode ser possível perceber que o filtro não é totalmente aplicado ao calcular as bordas da matriz resultante. Nessa situação, é comum realizar o preenchimento das bordas da imagem original, adicionando um contorno de zeros de acordo com o tamanho do filtro, ver Figura 2.5.

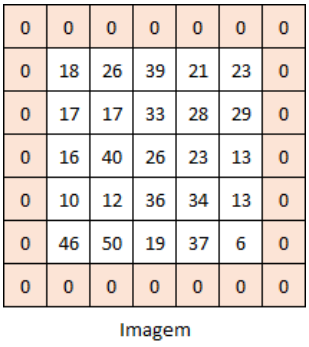

Figura 2.5 – Demonstração visual da operação de preenchimento em uma matriz.

Se nota que a matriz resultante tem as mesmas dimensões da imagem de entrada. Isso acontece porque a sobreposição do filtro sobre a imagem ocorre com deslocamentos de um elemento por operação. É possível aumentar esse deslocamento, tanto vertical quanto horizontalmente. Essa alteração fará com que a matriz resultante tenha dimensões menores do que a

imagem original. A definição desse deslocamento é chamado de *stride*, que pode ser aplicado tanto nas camadas de convolução quanto nas camadas de subamostragem.

Também é importante observar que os exemplos apresentados trabalham com matrizes bidimensionais, mas, tratando-se de imagens coloridas, são utilizadas matrizes tridimensionais, em que a terceira dimensão é o espaço de cores RGB. Nesses casos, filtros 3D também podem ser aplicados. Além disso, em uma única camada de convolução, vários filtros são aplicados na imagem, gerando múltiplas matrizes resultantes referentes a cada um deles. Dessa forma, em uma única camada de convolução, várias características são extraídas da imagem simultaneamente.

2.3.4 Camada de subamostragem

A camada de subamostragem (*pooling*) é utilizada para reduzir a dimensionalidade do problema. Normalmente, a rede precisa identificar se determinada característica está presente em uma imagem e em qual região ela se encontra. Dessa forma, não é necessário manter todas as informações dessa característica dentro da rede. Realizar essa operação faz com que o poder de generalização da rede aumente.

O processo de subamostragem ocorre também por meio de uma convolução, mas em vez do resultado ser igual à multiplicação e soma dos valores da região, ele é a média dos valores (*average pooling*), o máximo valor encontrado (*max pooling*) ou a raiz da soma dos quadrados dos valores (*l2 pooling*). Da mesma forma que ocorre na camada de convolução, é possível alterar o tamanho do deslocamento do filtro. Só é importante notar que a maioria das arquiteturas mais famosas de CNNs utilizam regiões que se sobrepõem.

A Figura 2.6 ilustra a execução de subamostragem com regiões de tamanho 3×3 e stride de 3×3, utilizando as três formas de agrupamento de valores mencionadas.

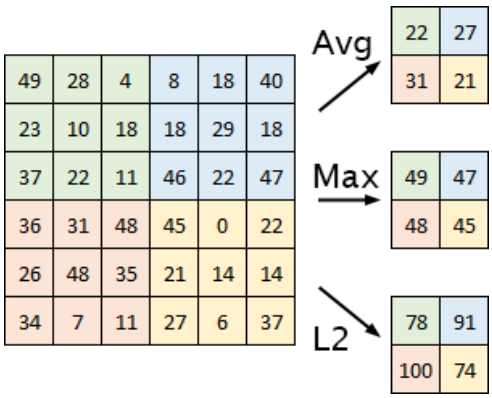

Figura 2.6 – Operação de *pooling* em uma matriz de valores.

2.4 Vision Transformers

Os ViTs são um tipo de arquitetura de aprendizado profundo que ganhou muita atenção recentemente por seu notável desempenho em tarefas de classificação de imagens. Ao contrário das tradicionais CNNs, os ViTs usam um mecanismo de *self-attention* que lhes permite capturar informações globais da imagem de entrada. *Self-attention* é um mecanismo usado em ViTs que permite que o modelo analise diferentes partes da entrada, capturando características locais e globais. No contexto de ViTs, a entrada é tipicamente uma sequência de *patches* da imagem.

A arquitetura ViT começa dividindo a imagem de entrada em uma sequência de *patches* de tamanho fixo. Cada *patch* é então transformado em uma representação vetorial e passado por um camada de projeção linear para obter um *embedding* que pode ser aprendido. Esses *embeddings* são então alimentados a um codificador chamado de *Transformer*, que é uma série de camadas de *self-attention*, utilizando o algoritmo de *feedforward*. As camadas de *feedforward* processam a saída das camadas de *self-attention* para produzir um *embedding* final para a determinada sequência de entrada, no caso, a imagem. Depois que o *embedding* final é obtido, ele é passada por um MLP para produzir uma lista de probabilidades da imagem pertencer a cada classe.

Uma das principais vantagens dos ViTs é que eles podem processar imagens de tamanho arbitrário sem exigir nenhuma operação de processamento de imagens, como distorções. Isso os torna mais flexíveis do que as CNNs, que exigem que as imagens de entrada tenham um tamanho fixo. ViTs também têm menos parâmetros do que CNNs de tamanho semelhante, o que os torna mais eficientes computacionalmente. No geral, os vit mostraram resultados impressionantes em vários benchmarks de classificação de imagens, demonstrando que são uma alternativa promissora às CNNs tradicionais.

A Figura 2.7 demonstra os elementos, descritos anteriormente, de um ViT, sendo possível identificar a segmentação da imagem de entrada em *patches*, a vetorização dos dados, a extração de *embeddings* que são enviados ao *transformer*, e o MLP que faz a classificação:

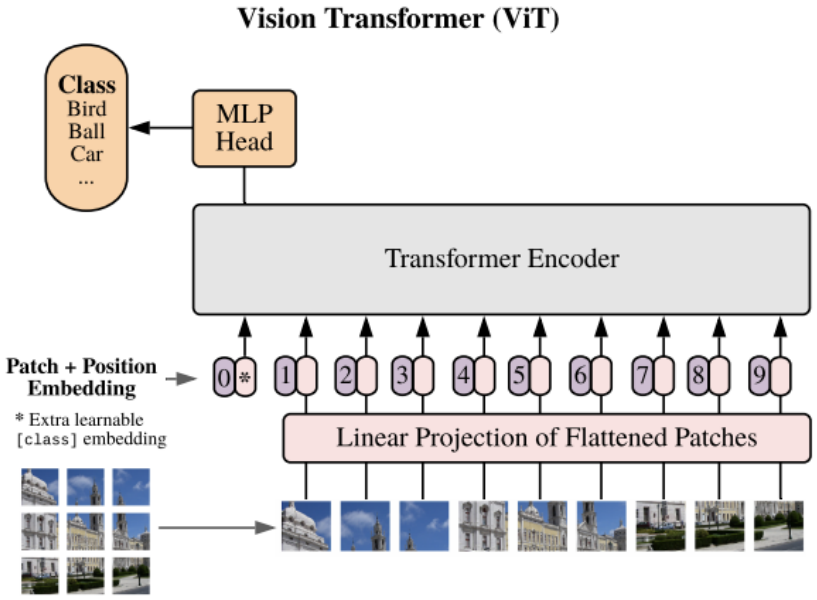

Figura 2.7 – Estrutura de um ViT. (DOSOVITSKIY et al., 2020)

2.5 Resultados de múltiplos modelos

A biometria em questão se refere à biometria facial e à biometria da orelha. Apesar de serem duas áreas muito estudadas, a execução delas em um mesmo modelo é algo incomum. Dessa forma, unir o reconhecimento facial e o reconhecimento da orelha se torna uma tarefa de agrupar resultados de dois modelos distintos.

O agrupamento dos resultados pode ser feito de diferentes formas (MAHOOR; CADAVID; ABDEL-MOTTALEB, 2009). Dentre estas, são citados os modelos de soma ponderada e combinação direta por meio da normalização pela tangente hiperbólica:

- A soma ponderada consiste em atribuir pesos para cada um dos componentes da soma e realizar a soma dos produtos dos valores pelos seus respectivos pesos. No estudo em questão, foram atribuídos pesos para o resultado do reconhecimento facial e do reconhecimento da orelha de forma empírica.
- Na combinação direta, os resultados de cada um dos modelos podem ser incompatíveis, ou seja, podem assumir valores pertencentes a domínios diferentes. Dessa forma, necessita-se de uma maneira de transformar os valores obtidos em valores que pertençam a um mesmo domínio. Para isso, é utilizada a função da tangente hiperbólica para normalizar os valores, permitindo, assim, a sua combinação.

Apesar das alternativas à soma dos resultados obtidos, Bowyer et al. (2006) apontam que em vários estudos que abordam a biometria com vários modelos consideram diferentes formas de unir os resultados de cada um, mas apesar disso, estatisticamente, não há uma melhoria significativa ao adotar outras técnicas além da soma. Em seu estudo, são apresentados alguns casos em que a normalização Min-Max apresentou melhores resultados do que a soma. Já nos experimentos de (YAN; BOWYER, 2005), essa normalização foi superada pela soma simples em todos os casos.

2.6 Métricas de avaliação

A avaliação de modelos de inteligência artificial ocorre por meio de diferentes métricas, cada uma com diferentes significados. Elas são baseadas no que o modelo apontou como verdadeiro o que é verdadeiro, o que apontou como falso o que é falso, e o que ele errou. A partir disso, é possível montar uma matriz que sumariza o comportamento do modelo, representada pela Figura 2.8:

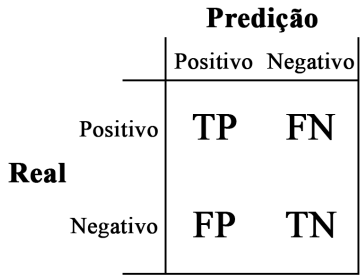

Figura 2.8 – Possíveis valores em uma matriz de confusão

- *True positive* (TP): ocorre quando o modelo prediz o valor positivo e o valor correto é positivo.
- *True negative* (TN): ocorre quando o modelo prediz o valor negativo e o valor correto também é negativo.
- *False positive* (FP): ocorre quando o modelo prediz o valor positivo mas o valor correto é negativo.
- *False negative* (FN): ocorre quando o modelo prediz o valor negativo mas o valor correto é positivo.

A partir dessa contagem de erros e acertos, é possível estabelecer razões numéricas que definem a performance do modelo, como a acurácia, precisão e revocação.

- A acurácia mede a proporção de acertos em relação a todas as predições realizadas:

$$\textit{Acurácia} = \frac{TP + TN}{TP + TN + FP + FN}$$

- A precisão mede, de todos os que foram detectados como positivos, quantos destes foram corretamente preditos:

$$\textit{Precisão} = \frac{TP}{TP + FP}$$

- A revocação mede a proporção de acertos em relação a todos os que deveriam ser positivos:

$$\textit{Revocação} = \frac{TP}{TP + FN}$$

Esse método de cálculo pode ser utilizado para medir a qualidade de modelos que tratam de classificações binárias, onde existem apenas duas classes. No entanto, neste trabalho, a classificação de pessoas é compreendido como um problema de múltiplas classes. Dessa forma, não é possível estabelecer apenas quatro possíveis resultados para as predições em *n* classes.

Para resolver esse problema, é possível utilizar duas abordagens: uma delas é de calcular a métrica em questão para cada uma das classes individualmente e depois obter a média da métrica em cada uma delas. A outra é de somar todos os TP, TN, FP e FN para depois aplicar a fórmula da métrica. A primeira é chamada de abordagem *macro* e a segunda de abordagem *micro*.

3 Revisão de Literatura

Este capítulo conta com a exposição de diversos estudos sobre o tema deste trabalho. São analisados estudos sobre o reconhecimento facial, sobre o reconhecimento da orelha e sobre a união de ambos em um único modelo. São discutidas as técnicas e estratégias aplicadas, datasets utilizados e resultados obtidos.

3.1 Trabalhos Relacionados

Nesta seção, será apresentada uma revisão de trabalhos presentes na literatura, apresentando métodos realizados no processo da área de extração de características para classificação de imagens. Especificamente, esta seção será dividida em duas subseções: *Handcrafted Features* e *Deep Learning*. A Subseção 3.1.1 se concentrará nos métodos tradicionais de extração de características, que dependem do design manual de características com base no conhecimento do domínio. A subseção 3.1.2 apresentará alguns trabalhos que realizam o processo de extração de características baseada em *deep learning*, que mostram resultados notáveis em várias tarefas de classificação de imagens.

3.1.1 Handcrafted Features

Yan e Bowyer (2005) experimentou diferentes técnicas para o reconhecimento facial e o reconhecimento da orelha. Foi utilizado uma base de dados com imagens obtidas da Universidade de Notre Dame, contendo imagens 2D e imagens 3D capturadas por um *scanner*. Foi utilizado o anteriormente mencionado *Principal Component Analysis* (PCA) tanto para as imagens 2D quanto para as 3D, e o algoritmo *Iterative Closest Point* (ICP) para a síntese de modelos em 3D. Dessa forma, foi realizada a combinação dos resultados obtidos por cada modelo através, primeiramente, da normalização Min-Max e, posteriormente, da obtenção do valor mínimo obtido, soma e soma ponderada. O método de fusão de resultados que apresentou melhor precisão foi o da soma ponderada, com peso maior para o reconhecimento do modelo 3D da orelha em relação ao reconhecimento da imagem 2D, que se deu por volta de 80% e 20%, respectivamente. Outros experimentos são realizados elencando múltiplos modelos com diferentes algoritmos, apresentando diferentes combinações do PCA e do ICP. Os resultados com as melhores configurações encontradas apresentaram precisão de aproximadamente 90%, utilizando o PCA para o reconhecimento nas imagens 2D e o ICP nas imagens 3D.

De forma similar ao trabalho mencionado anteriormente, Mahoor, Cadavid e Abdel-Mottaleb (2009) também trabalhou com imagens 2D e 3D, porém as imagens 2D são da face do indivíduo e o modelo 3D é da orelha do indivíduo que é obtido através de um videoclipe. É

utilizado o algoritmo *Shape From Shading* (SFS) para construir o modelo 3D da orelha a partir de cada quadro do clipe, que são então alinhados por meio do ICP. A partir da construção do modelo 3D, ele é alinhado com os modelos da base de dados e é feito um cálculo da distância Euclideana entre os pontos dos dois modelos, em que o modelo que tiver maior similaridade é aquele que pertence ao indivíduo. Para o reconhecimento facial, a extração de características ocorre por meio de filtros de Gabor, mencionado anteriormente e, por fim, a partir do resultado obtido da extração, os pontos são extraídos utilizando o *Active Shape Model*. Para realizar a fusão dos resultados dos dois modelos, é utilizada a soma ponderada, em que os pesos para cada um dos termos da soma foram encontrados de forma empírica. Após a combinação dos resultados por meio da soma ponderada, os resultados obtidos foram de 100% de precisão na base com 402 indivíduos que foi utilizada.

O algoritmo PCA também foi utilizado pelos autores (CHANG et al., 2003). Seu experimento foi realizado sobre um base de dados fornecida pela *University of South Florida* (USF), contendo quase 200 indivíduos. O trabalho foi realizado em imagens da face e da orelha, realizando o reconhecimento para desvendar questões que, no geral, apontariam que utilizar as mesmas técnicas de reconhecimento para a face a para a orelha traria resultados similares. Foi notado que a precisão do reconhecimento utilizando faces produz melhores resultados, apesar de não muito distantes. Para realizar a união dos resultados obtidos, foi utilizada uma abordagem de manipulação de pixels em vez de uma abordagem métrica. As imagens já pré-processadas da face e da orelha de um mesmo indivíduo foram concatenadas, formando uma única imagem. Essa manipulação levou a resultados melhores, demonstrando que realizar a união de dois tipos diferentes de biometria pode ser vantajoso.

Atualmente, redes neurais são amplamente utilizadas na área de processamento digital de imagens. É esperado, dessa forma, que vários estudos no tópico de biometria surjam utilizando essa abordagem. (PRIYADHARSHINI; ARIVAZHAGAN; ARUN, 2021) produziu um modelo de rede neural convolucional para realizar tanto a extração de características quanto a classificação das imagens. Foram utilizados duas bases de dados diferentes: o IITD-I e IITD-II, com um total de 346 indivíduos, e o AMI, com 100 indivíduos. A escolha de qual função de ativação utilizar na rede neural ocorreu por meio de experimentos com cada uma, em que a tangente hiperbólica foi a que apresentou melhores resultados. Com essa configuração, o modelo atingiu 97% de precisão no reconhecimento de orelhas.

Utilizando o Histograma de Gradientes Orientados - *Histogram of Oriented Gradients* (HOG), Damer e Führer (2012) explorou o reconhecimento de orelhas fazendo uso da base de imagens IITD-II. Para a extração de características, utilizou o HOG, que mostrou robustez quanto a variações na iluminação, e para reduzir a redundância dos descritores, utilizou o PCA. Para realizar a classificação das imagens, utilizou o *K-Nearest Neighbor* (k-NN) com distância Euclideana. Os resultados obtidos se mostraram na média de 90% precisão, que é 5% abaixo de outros métodos comparados no estudo.

Xu, Mu e Yuan (2007) mostrou em seu estudo sobre reconhecimento facial e reconhecimento da orelha o uso de técnicas de fusão a nível de características. Foi utilizada a base de dados da *University of Science and Technology Beijing* (USTB), que contém aproximadamente 80 indivíduos. As imagens foram manipuladas para gerar mais dados de treinamento, mais especificamente rotacionadas em intervalos de 5 graus. A extração de características é baseada no algoritmo *Fisher Discriminant Analysis* (FDA). Ele funciona por meio da segmentação do espaço de informações, encontrando quais são as informações discriminantes, evitando que essas sejam descartadas durante o uso do PCA para redução da dimensionalidade do problema. O modelo proposto utiliza, de fato, o *Kernel Fisher Discriminant Analysis* (KFDA) que, diferentemente do FDA que opera no espaço linear, opera no espaço não linear, permitindo extrair características não-lineares. Esse algoritmo é utilizado tanto para as imagens da face quanto para as imagens da orelha, gerando matrizes do kernel de ambas as regiões. Essas matrizes são então combinadas por um conjunto de funções e de uma regra de fusão. As regras de fusão testadas foram do produto, da média e da soma ponderada. Assim como em outros estudos, a união dos dois modelos melhorou os resultados e, dentre as diferentes regras de fusão, a soma ponderada apresentou maiores ganhos, atingindo 96.84% de precisão.

3.1.2 Deep Learning

(ALSHAZLY et al., 2019) fizeram o uso de *ensembles* para produzir um classificador capaz de reconhecer pessoas a partir das orelhas. Também fizeram o uso de *transfer learning* utilizando as redes VGG como extratores de características. Os experimentos realizados pelos autores foram em cima de diferentes datasets com imagens de orelhas que foram obtidas em ambientes controlados. Dentre eles, está o dataset AMI, que contém imagens de 100 indivíduos com 7 amostras de cada, totalizando, assim, 700 amostras. Em algumas imagens desse dataset contém outras partes do corpo como o pescoço ou partes do cabelo. Em função disso, eles criaram um novo dataset com as mesmas imagens, porém cortadas, eliminando esses artefatos. Também foi utilizado o dataset WPUT, que contém 3348 imagens de 474 indivíduos. Trabalhar com esse dataset apresenta um desafio maior porque muitas das imagens estão muito obstruída pelo cabelo ou por outros acessórios. Foi realizado o processo de *data augmentation* para aumentar o número de amostras, incluindo transformações como ajuste de cor, mudança de escala, rotação, recorte e aplicação de *blur*. Para a criação dos modelos, foram testados o treinamento do zero em redes separadas, treinamento utilizando *transfer learning* em redes separadas, uso apenas da extração de características e uso de ensembles, todos com as redes VGG-11, VGG-13, VGG-16 e VGG-19. Os modelos que apresentaram melhores resultados nos datasets foram os que fizeram uso de ensembles, atingindo de 93% a 99% de precisão.

Tratando de imagens obtidas em ambientes não controlados, (EMERŠIČ et al., 2019) propuseram um *pipeline* de processamento com modelos para realizar tanto a detecção da orelha quanto o reconhecimento da pessoa. Todo o *pipeline* se baseia em CNN's, utilizando redes como a

RefineNet para a detecção e a ResNet-152 para o reconhecimento, de forma que novos indivíduos sejam identificados sem a necessidade de que a rede seja passada pelo processo de treinamento novamente. Nos experimentos realizados, foram utilizados os datasets *Annotated Web Ears* (AWE) e *Unconstrained Ear Recognition Challenge* (UERC). No total, constam 4.004 imagens de 336 indivíduos, com números variáveis de imagens por indivíduo. Essas amostras apresentam diferentes formas de iluminação, gênero, idade, raça, e outros fatores. Para o treinamento do modelo que realiza a detecção das orelhas, 1.000 imagens de 100 indivíduos foram manualmente marcadas para servirem de dados de treinamento e teste. Para avaliar os modelos criados, foi levado em consideração a acurácia, precisão e revocação. Os melhores resultados obtidos constam 99.8% de acurácia, 91.7% de precisão e 91.6% de revocação.

(TIAN; MU, 2016), em seu trabalho, propõem uma arquitetura de CNN com três camadas convolucionais, três camadas de subamostragem e duas camadas totalmente conectadas para realizar a identificação de pessoas por meio da orelha. Toda a rede é treinada do zero a partir do dataset USTB III, que contém imagens obtidas em ambiente controlado de 79 indivíduos com variações de rotação, com 10 imagens de cada um dos indivíduos. Foi realizado um pré-processamento para aumentar a quantidade de dados. Nessa etapa, as imagens originais do dataset foram condicionadas a operações de rotação, escala, ajuste de contraste e normalização de média. Foi realizado um experimento relacionado à obstrução da orelha nas imagens. Foram inseridos quadrados pretos em regiões aleatórias da imagem, simulando uma obstrução, permitindo que a CNN tenha a capacidade de lidar com imagens de orelhas obstruídas. Por meio das técnicas de aumento de dados, os experimentos realizados contaram com 237.000 imagens para treinamento e 1.975 para testes. A precisão média obtida foi de 98.27% para imagens sem obstrução. Com obstrução, a taxa de reconhecimento caiu para aproximadamente 25% com 50% da imagem obstruída. Foi concluído que, por mais que os resultados obtidos foram bons, o reconhecimento de pessoas através da orelha em ambientes não controlados apresenta com problemas a serem enfrentados, principalmente referente à oclusão da orelha. É dito que realizar o reconhecimento com grandes porcentagens de oclusão com precisão não é realista.

(ALMISREB; JAMIL; DIN, 2018) propuseram a criação de uma CNN baseada na *AlexNet*, aplicando a técnica de *transfer learning*. O processo utilizado para construir a rede consiste em substituir as camadas totalmente conectadas e a camada de classificação por novas, e, no processo de treinamento, apenas as novas camadas serão treinadas, enquanto os parâmetros das camadas convolucionais permanecem os mesmos. O dataset utilizado foi criado pelos autores em um ambiente controlado, compreendendo 30 imagens por indivíduo, de um total de 10 indivíduos. Dessa forma, do total de 300 imagens, 250 são utilizadas para treinamento e 50 para testes. Outras formas de pré-processamento não foram aplicadas. Na experimentação foi utilizado o Stochastic Gradient Descent with Momentum (SGDM) com *momentum* em 0,9. O modelo atingiu 100% de acurácia na etapa de testes.

Uma técnica recente para realizar a classificação de imagens vem sendo amplamente

estudada devido ao seu potencial de criar modelos com desempenho igual ou superior ao uso de CNNs. O artigo original do *Vision Transformer* (ViT), de (DOSOVITSKIY et al., 2020), demonstra que o ViT alcança bons resultados em vários *datasets* de referência, incluindo *ImageNet*, com significativamente menos parâmetros em comparação a outros modelos. Os autores também mostraram que o ViT pode ser treinado de forma eficiente em grandes conjuntos de dados usando técnicas modernas de treinamento em paralelo.

Desde então, o ViT tem sido objeto de vários estudos de acompanhamento explorando diferentes aspectos da arquitetura e seu desempenho. Por exemplo, um artigo de (YUAN et al., 2021) introduziu uma modificação na arquitetura ViT original, chamada Tokens-to-Token ViT (T2T-ViT), que substitui os *patches embeddings* usados no modelo original por uma sequência de *tokens* derivados diretamente dos pixels da imagem. Os autores mostraram que o T2T-ViT atinge um desempenho ainda melhor do que o ViT original em vários *benchmarks* de classificação de imagens.

Outro estudo de 2021, de (HAN et al., 2021), exploraram o papel das *position embeddings* no ViT. Os autores propuseram um novo tipo de *posistion embeddings* que é aprendida durante o treinamento, em vez de ser fixa, como na arquitetura original do ViT. O modelo resultante, chamado Learnable Position Embeddings (LPE-ViT), alcançou ótimos resultados no *ImageNet* e em vários outros *benchmarks* de classificação de imagens.

De acordo com (LI et al., 2021), a estrutura de *self-supervised* baseada no Masked Self-Supervised Transformer (MST) permite um treinamento eficiente em conjuntos de dados de grande escala. Eles demonstraram que, ao pré-treinar o modelo MST em um grande conjunto de dados, como *ImageNet*, e ajustá-lo a um conjunto de dados menor, o desempenho de ponta pode ser alcançado com apenas uma fração dos dados de treinamento normalmente necessários por modelos supervisionados.

(ALEJO, 2021) propõe uma nova abordagem para o reconhecimento a partir da orelha usando ViT. O método proposto utiliza a arquitetura do ViT para processar imagens da orelha representadas como sequências de *patches*, produzindo uma representação final da orelha para reconhecimento. O método supera os métodos de reconhecimento de orelha considerados estado da arte em dois *datasets* públicos de reconhecimento de orelha, demonstrando sua eficácia para o reconhecimento da orelha em ambientes não controlados. Além disso, as representações da orelha aprendida são altamente discriminativas, sugerindo a adequação do método proposto para outras tarefas relacionadas à orelha, como estimativa de sexo e idade. A capacidade do método proposto de lidar com imagens de orelhas em ambientes não controlados o torna mais prático para aplicações do mundo real, como autenticação biométrica e vigilância.

(WANG et al., 2022) propõem um novo modelo de reconhecimento facial de pessoas que fazem uso de máscaras, que combina a tecnologia de reconhecimento facial 3D com Vision Transformers para melhorar o desempenho do reconhecimento. O modelo é treinado usando imagens faciais de pessoas sem máscaras, localizando e segmentando a região facial completa

e a região facial não obstruída por máscaras das nuvens de pontos faciais, isso porque, em um curto período de tempo, é difícil encontrar um grande conjunto de dados com pessoas fazendo uso de máscara. O modelo proposto é comparado com modelos existentes de reconhecimento facial de pessoas com máscaras e foi obtida a melhora na precisão do reconhecimento em até 34,81% em testes em diferentes conjuntos de dados, demonstrando sua eficácia e estabilidade.

O artigo de (SU et al., 2023) discute as limitações das CNNs em reconhecer rostos humanos devido à sua incapacidade de entender dependências de características globais e locais em imagens. Para resolver esse problema, os autores propõem um módulo chamado de *Hybrid token Transformer* (HOTformer) que pode aproveitar as regiões faciais centrais e contextuais para gerar tokens híbridos discriminativos para reconhecimento facial. O módulo HOTformer é um módulo *plug-and-play* que pode ser inserido em CNNs pequenas, como MobileFaceNets. Os autores conduziram experimentos em *benchmarks* de reconhecimento facial amplamente utilizados e descobriram que o HOTformer superou os métodos recentes considerados estado-da-arte em até 2.7% de performance.

4 Metodologia

Como apresentado anteriormente, vários estudos foram realizados para realizar o reconhecimento de pessoas através da face e da orelha. Dentre eles, diferentes técnicas foram abordadas para atingir seus objetivos. A partir desses estudos, este capítulo tem como finalidade apresentar detalhadamente todos os processos realizados durante o desenvolvimento do trabalho, como também a explicação de cada decisão tomada para a composição da solução.

4.1 Algoritmo proposto

Fazendo valer os objetivos específicos propostos no Capítulo 1, é proposto a utilização de ANNs para realizar o reconhecimento de pessoas. Partindo dos dados de entrada, estes serão alimentados às duas CNNs, que são responsáveis por fazer, uma, o reconhecimento da face e, outra, o reconhecimento da orelha. Os resultados de cada uma destas são então enviados a uma outra ANN que, por sua vez, realiza a classificação da pessoa.

A Figura 4.3 apresenta, em linhas gerais, o procedimento adotado para a confecção do algoritmo de reconhecimento proposto:

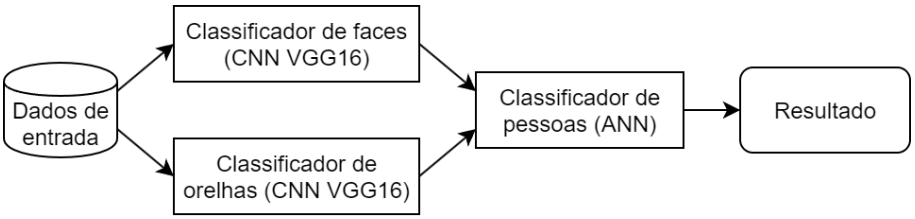

Figura 4.1 – Etapas que compõem o algoritmo de reconhecimento proposto.

Os classificadores de faces e de orelhas são redes neurais convolucionais, que são montadas de acordo com a arquitetura VGG-16. O classificador de pessoas é uma ANN que, a partir dos resultados das duas CNNs, realiza a classificação da pessoa.

4.1.1 Arquitetura VGG-16

A técnica de transferência de aprendizado é empregada para utilizar modelos pré-treinados na rede. Duas instâncias da rede VGG-16 são utilizadas para realizar o reconhecimento facial e o reconhecimento da orelha dos indivíduos. Sua arquitetura é ilustrada na Figura 4.2.

Alguns pontos sobre essa arquitetura são importantes: ela aceita como entrada imagens de resolução 224×224 pixels, com três canais de cores, ou seja, imagens coloridas. Suas camadas convolucionais são descritas na Figura 4.2, juntamente das camadas totalmente conectadas. A última camada é a responsável por realizar a classificação da imagem, contendo 1.000 objetos de

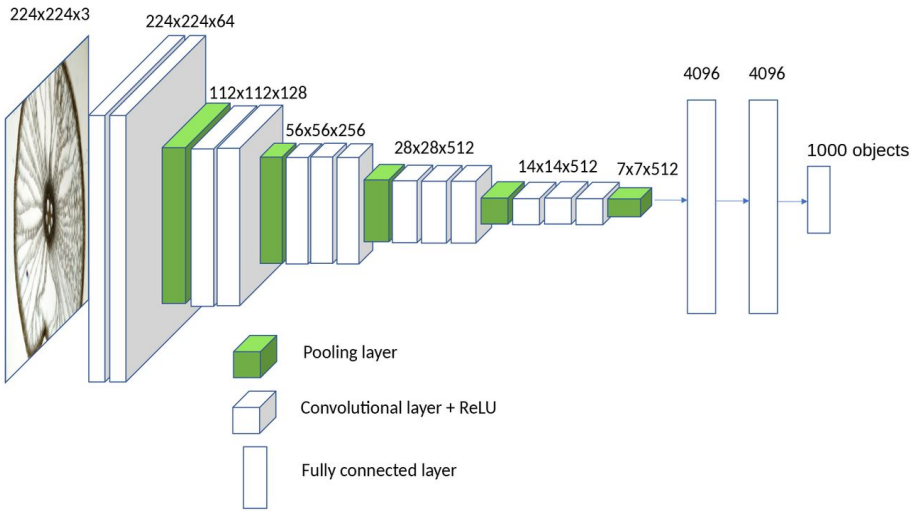

Figura 4.2 – Arquitetura da rede VGG-16. (STACKEXCHANGE, 2022)

saída, referentes à probabilidade da imagem pertencer a aquela classe. Essa última camada deve ser descartada e, em seu lugar, adicionar uma nova camada que será treinada, contendo o número de saídas igual ao número de classes da base de dados utilizada.

Sumarizando, a arquitetura da rede VGG-16 recebe como entrada imagens coloridas de 224×224 pixels, que é processada por duas camadas de convolução com 64 filtros, que são subamostradas com *stride* de 2×2, fazendo o *downsample* para 112×112 pixels. O processo se repete mais uma vez, mas com 128 filtros e *pooling* de 2×2. Nos próximos conjuntos de camadas são três camadas de convolução com 256 filtros juntamente do *max-pooling*, e, na seguinte, 512 filtros. É realizado um *pooling* final, resultando em 512 características de 7×7. Essas características são vetorizadas e passadas por duas camadas densas de 4096 neurônios que, por fim, passam pela função de ativação *softmax* para a classificação de 1000 classes.

4.1.2 Arquitetura ViT-B/32

Também fazendo o uso de *transfer learning*, é proposta a substituição dos modelos pré-treinados de arquitetura VGG-16 por modelos que fazem uso da arquitetura ViT-B/32 como extrator de características. Essa variante do ViT possui 32 blocos *Transformer* e 12 nós de *self-attention* por bloco. Os dados de entrada são imagens coloridas de 224×224 pixels, e a saída é passada por um MLP com camadas densas totalmente conectadas e, por fim, uma camada *softmax* para fazer a classificação.

O uso dessa arquitetura será feito para o classificador de orelhas e para o classificador de faces, substituindo as CNNs para efeito de comparação de desempenho. Após a substituição, o classificador de pessoas será treinado novamente para se adaptar à saída produzida pelos novos classificadores ViT.

4.2 Classificadores

O algoritmo de classificação proposto é composto por classificadores que atuam em conjunto para realizar o reconhecimento a partir da face, da orelha, e a união dos resultados obtidos pelas duas anteriores. As etapas executadas para a criação destes classificadores são exibidas na Figura 4.3. É apresentado, posteriormente, o detalhamento de cada um dos componentes.

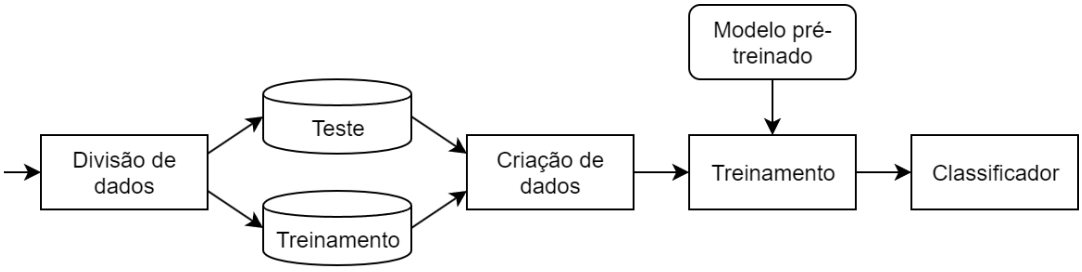

Figura 4.3 – Etapas para a criação dos classificadores.

4.2.1 Divisão de dados

No processo de criação de um modelo de aprendizado de máquina, é necessário que os dados utilizados para teste sejam diferentes dos dados utilizados para treinamento. Isso é importante para medir o desempenho do modelo na etapa de classificação, impedindo que o modelo aprenda características específicas do conjunto de treinamento, prejudicando sua capacidade de generalização.

A divisão dos conjuntos de treinamento e de teste pode ser realizada de diversas maneiras. A mais trivial é realizar a separação em porcentagens, como 80% dos dados para treinamento e 20% para teste, ou com outras proporções, como 70% e 30%. Essa técnica, porém, pode não apresentar resultados satisfatórios. É possível que, com essa separação, por acaso, um dos subconjuntos obtidos tenha uma característica específica. Isso afetará a capacidade de generalização do modelo negativamente. Diante disso, foi utilizada outra técnica.

O uso de validação cruzada com K-Folds se baseia no conceito de utilizar todo o conjunto de dados tanto para treinamento quanto para teste. Os dados são subdivididos em K conjuntos, e em cada momento um desses conjuntos é selecionado para teste e o restante para treinamento. Dessa forma, o modelo é treinado e testado várias vezes, e sua precisão final é obtida pela média das precisões de todas as rodadas de treinamento e teste, como exemplificado na Figura 4.4.

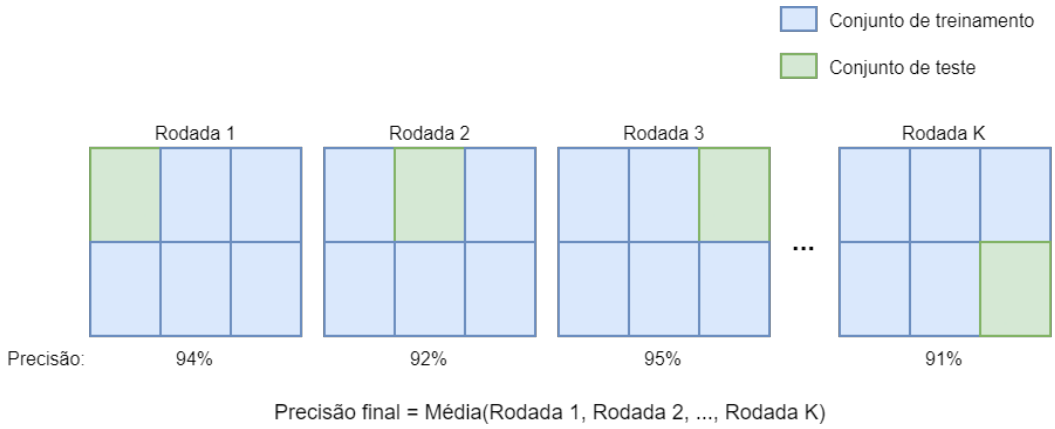

Figura 4.4 – Ilustração da aplicação da validação cruzada em um conjunto de dados.

Fazendo uso da validação cruzada, o problema mencionado na separação em porcentagens não ocorre, dado que mesmo que em alguns subconjuntos existam características específicas, em outras rodadas esse subconjunto será alterado, eliminando o problema.

4.2.2 Criação de dados

O DL é uma forma de aprendizado de máquina que apresenta melhores resultados quando grandes volumes de dados são utilizados em seu treinamento. Isso ocorre porque redes profundas constam milhares senão milhões de parâmetros a serem ajustados em seu interior, e esse ajuste ocorre na fase de treinamento em função dos dados fornecidos. Dessa forma, se poucos dados são fornecidos, esses parâmetros acabam não sendo ajustados apropriadamente.

Para resolver esse problema tendo apenas um pequeno conjunto de dados, é possível utilizar uma estratégia chamada *data augmentation*. Ela consiste em fazer transformações geométricas nas imagens do conjunto de dados de forma a produzir novas imagens. Dentre essas transformações, é possível: inverter a imagem horizontal ou verticalmente, rotacioná-la, mover o objeto de interesse, recortá-la, aumentar ou diminuir seu tamanho, inserir ruído, entre outras.

Além de transformações geométricas, tratando-se de faces, é também possível adicionar acessórios como óculos, ou mudar o tipo de cabelo do indivíduo, como é mostrado por (LV et al., 2017)

A Figuras 4.5 e 4.6 exibem os resultados da execução dessa estratégia em imagens de exemplo, tanto em uma orelha quanto em uma face:

4.2.3 Treinamento

A fase de treinamento do modelo é o momento em que os parâmetros da rede são ajustados conforme os dados de entrada são fornecidos. Como explicado na fundamentação teórica, as camadas convolucionais atuam como extratores de características. Dessa forma, é possível utilizar

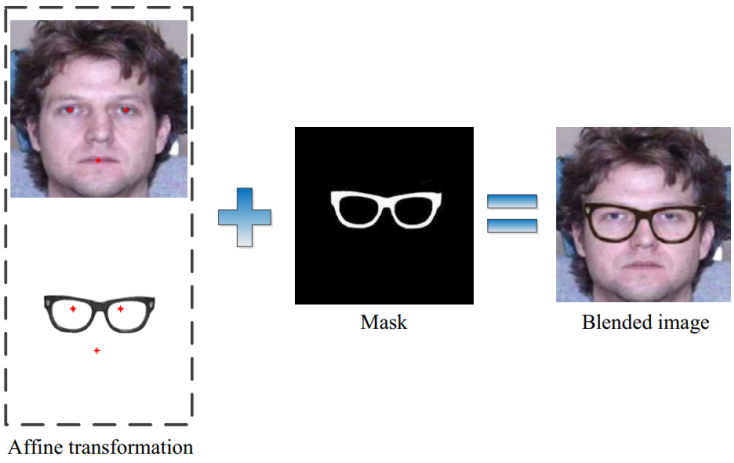

Figura 4.5 – *Data augmentation* da face (LV et al., 2017)

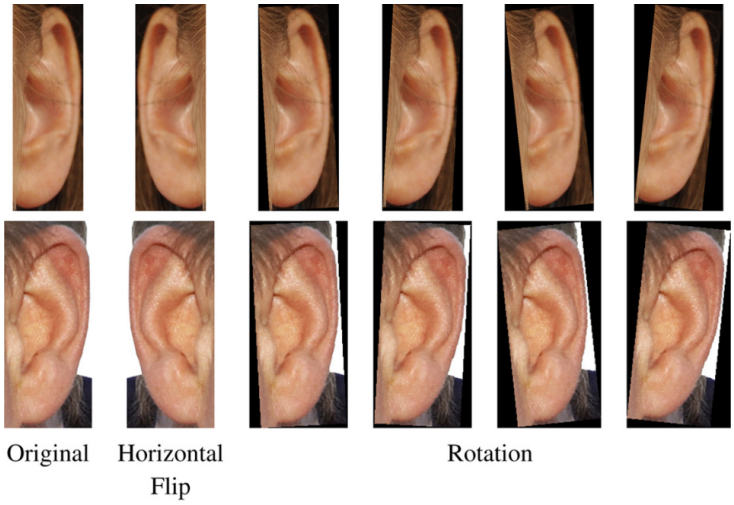

Figura 4.6 – *Data augmentation* da orelha (DODGE; MOUNSEF; KARAM, 2018)

redes que já possuem esses extratores treinados, necessitando do treinamento apenas das camadas totalmente conectadas. Esse procedimento de transferência de conhecimento da rede é chamado de *transfer learning*. O modelo já treinado utilizado foi o VGG-16, tanto para o reconhecimento facial quanto para o reconhecimento da orelha.

Para realizar o treinamento, é necessário definir algumas configurações de execução. Por se atuarem na etapa de treinamento de forma externa, sem interagir com a arquitetura da rede ou com os dados, esses são chamados de hiperparâmetros. Os hiperparâmetros a serem definidos são:

- A taxa de aprendizado, que dita o quão rápido o modelo aprende, mas que não deve assumir valores muito altos para evitar que a solução nunca chegue a um ponto ótimo. É comum utilizar um *scheduler*, que permite que diferentes taxas de aprendizado sejam utilizadas conforme o processo de treinamento avança.

- O número de épocas (*epochs*), que são quantos ciclos de treinamento são realizados até o fim da fase de treinamento.
- O otimizador, que determina qual algoritmo utilizar para minimizar a função de perda, ajudando a encontrar os melhores valores de parâmetros para o modelo.
- O tamanho das amostras de *mini-batch*, que define a quantidade de elementos nos conjuntos de treinamento para o otimizador de descida do gradiente.

4.2.4 Classificação

A classificação é o momento em que serão comparadas as representações das faces e orelhas conhecidas e a representação da face ou orelha detectada. Dessa forma, a imagem do indivíduo conhecida é interpretada pela rede e sua representação é salva, para que quando uma nova imagem for interpretada, a representação desta seja comparada com as representações salvas.

Essa comparação entre as duas representações pode ser realizada por meio de métodos matemáticos, como a distância Euclideana ou a similaridade por cosseno. A partir de um limiar definido previamente, se o resultado dessa comparação for menor, pode-se concluir que a face foi reconhecida.

5 Experimentos computacionais

No presente capítulo são apresentados os experimentos realizados como forma de validação da metodologia proposta no trabalho. Todos os experimentos foram conduzidos em um computador com processador AMD Ryzen 5 5600x 4.7GHz, 32GB de memória RAM DDR4 3200MHz, placa gráfica NVIDIA RTX 3060 Ti 8Gb, em ambiente Windows 10 Education 64 bits. Foi utilizada a versão 3.8.10 da linguagem de programação Python, fazendo uso das bibliotecas TensorFlow e Keras, ambas na versão 2.11.0, para a construção dos modelos. Também foram utilizadas as bibliotecas scikit-learn (v1.0.2), para o carregamento de dados e métricas de performance, NumPy (v1.22.4) para a manipulação de dados, e Matplotlib (v3.5.3) para a construção de gráficos.

Foi realizada a experimentação com os *datasets* AMI (GONZALEZ; MAZORRA, 2012), que conta com imagens de orelhas obtidas em ambiente controlado, que é mais profundamente detalhado na Seção 5.1.1, e o FEI Face database (THOMAS, 2006), que contém imagens da face retiradas em um ambiente controlado, mais detalhado na Seção 5.1.2. Para a experimentação com *datasets* não controlados, foi utilizado o Labeled Faces in the Wild (LFW) (HUANG et al., 2007), que conta com imagens de faces encontradas na Internet, e rotuladas com o nome da pessoa, detalhado na Seção 5.1.3. Nesse mesmo contexto, foi utilizado o *dataset* VGGFace-Ear (RAMOS-COOPER; GOMEZ-NIETO; CAMARA-CHAVEZ, 2022), detalhado na Seção 5.1.4, que conta com imagens de orelhas extraídas do *dataset* VGGFace2. A Seção 5.3 trata dos modelos utilizados para realizar os experimentos, juntamente das configurações de hiperparâmetros definidas. Por fim, a Seção 5.4 apresenta os resultados obtidos nos experimentos realizados, relacionados ao reconhecimento de pessoas através da face, da orelha, e da combinação dos dois anteriores.

O código-fonte criado para a realização dos experimentos se encontra publicamente neste repositório.

5.1 Bases de imagens

Na condução dos experimentos, foram escolhidas duas bases de imagens distintas para realizar o treinamento dos classificadores de faces e orelhas. Foi criada também uma base de imagens a partir da mesclagem das duas bases escolhidas, combinando as faces de uma base com as orelhas da outra. Os fatores que influenciaram na decisão de quais *datasets* utilizar foram a quantidade de amostras, qualidade das imagens em respeito à resolução e presença de cores, uso na literatura e

5.1.1 AMI Ear

A base de imagens utilizada para realizar os experimentos relacionados à biometria da orelha foi a *AMI Ear*, que conta com 7 amostras da orelha de cada um dos 100 indivíduos, todas obtidas em ambiente controlado, totalizando 700 amostras. Das 7 amostras de cada indivíduo, uma é da orelha esquerda, e as demais são da orelha direita em diferentes posições e ângulos. Todas as imagens são coloridas e têm a resolução de 492×702 pixels. As 7 amostras de um dos indivíduos são mostradas na Figura 5.1.

Figura 5.1 – Amostras de um indivíduo da base *AMI Ear*

5.1.2 FEI Face

A base de imagens utilizada para realizar os experimentos relacionados à biometria facial foi a *FEI face*, que contém com 14 amostras da região do pescoço para cima de 200 indivíduos, todas obtidas em ambiente controlado, totalizando 2600 amostras. 10 amostras de cada indivíduo o apresenta em diferentes rotações com expressão neutra, e as outras 4 de frente para a câmera, três delas com níveis de iluminação diferentes, e uma com o indivíduo sorrindo. Todas as imagens são coloridas e têm a resolução de 640×480 pixels. As 14 amostras de um dos indivíduos são mostradas na Figura 5.2.

Figura 5.2 – Amostras de um indivíduo da base *FEI face* (THOMAS, 2006)

Para a realização dos experimentos, entretanto, foram selecionadas 7 amostras de cada indivíduo, em que eles estão direcionados à câmera. Essas são as de índice 3, 4, 5, 6, 7, 8 e 11, que estão destacadas com um asterisco na Figura 5.2. Além disso, dado que a base de dados AMI conta com 100 indivíduos, foram escolhidos apenas os primeiros 100 indivíduos da FEI Face, seguindo a ordem alfabética das rotulações dadas a cada um deles, de forma que as duas bases tenham a mesma quantidade de indivíduos. Dessa forma, o número de amostras foi reduzido de 2600 para 700.

5.1.3 LFW Face

O conjunto de dados LFW consiste em imagens faciais de 5.749 pessoas, cada uma com uma ou mais imagens de amostra de 250×250 pixels. Essas imagens são coloridas e foram passadas pelo processo de *deep funneling*, que tem como objetivo padronizar a posição e a orientação da face, de forma que se crie uma consistência no alinhamento e posição do rosto da pessoa. A Figura 5.3 ilustra algumas amostras presentes no *dataset*. É possível perceber que há bordas pretas, que são resultado do processo mencionado:

Figura 5.3 – Algumas amostras presentes no *dataset* LFW

Para garantir que o conjunto de dados contenha amostras suficientes para cada pessoa, foi realizada uma filtragem para remover todas as pessoas em que existem menos de 7 amostras de sua face. Isso garante que o conjunto de dados tenha amostras suficientes para representar o rosto de cada pessoa com precisão e reduz o risco de sobreajuste durante o treinamento dos modelos.

Após a filtragem, com o objetivo de seguir a mesma quantidade de indivíduos em todos os *datasets*, foram escolhidos os primeiros 100 indivíduos quando ordenados em ordem alfabética pelo seu nome.

5.1.4 VGGFace-Ear

O *dataset* VGGFace-Ear conta com centenas de amostras de cada um dos 600 indivíduos presentes na base, sendo estas de tamanho variável e coloridas. Na tentativa de normalizar esse aspecto, foram escolhidas as 7 amostras de maior resolução para cada um dos indivíduos, e depois foram redimensionadas para atenderem à resolução de 224×224 pixels. As amostras exibem a orelha dos indivíduos de formas variadas, como rotacionadas, com níveis de iluminação diferentes, e nem sempre são da mesma orelha.

Assim como foi feito nas outras bases, foram selecionados, dos 600 indivíduos, os 100 primeiros seguindo a ordem alfabética da rotulação utilizada. A Figura 5.4 demonstra 10 amostras escolhidas aleatoriamente após o processo de seleção:

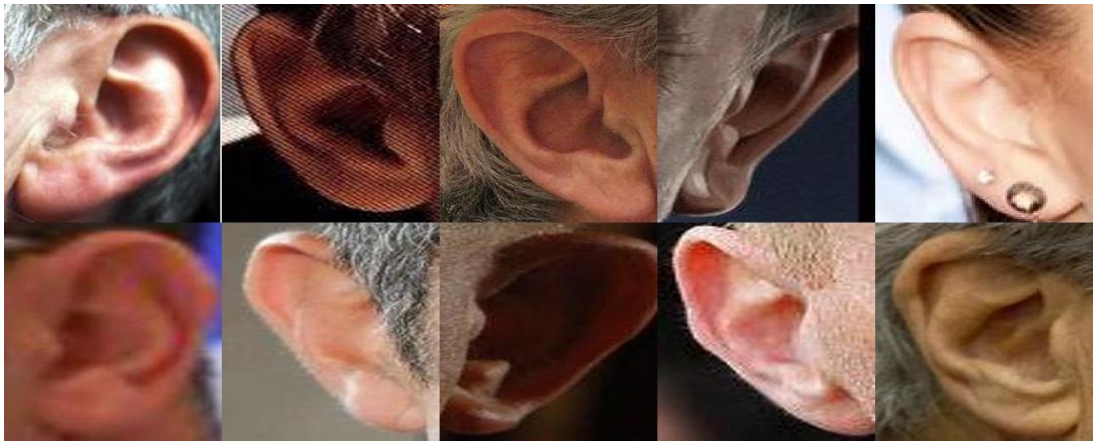

Figura 5.4 – Algumas amostras presentes no *dataset* VGGFace-Ear

5.1.5 Combinação de faces e orelhas

Para realizar o treinamento do classificador de pessoas, foram definidos dois pares entre as 4 bases, sendo um par compreendido pelos *datasets* AMI e o FEI Face, e o outro pelos *datasets* LFW e o VGGFace-Ear. Foi criado um relacionamento entre os dois elementos dos pares de bases de imagens mencionadas. Dado que cada indivíduo em cada base conta com 7 amostras, foi feito o relacionamento direto baseado nos índices de cada indivíduo e de cada amostra pertencente a eles, de modo que a amostra de índice i da face do indivíduo de índice j do *dataset* de faces seja pareada com a amostra de índice i da orelha do indivíduo j do *dataset* de orelhas.

5.2 Aumento de dados

A técnica de *data augmentation* foi utilizada para ampliar o número de dados utilizados no processo de experimentação, evitando que aconteça o sobreajuste na rede, melhorando sua capacidade de generalização. As transformações realizadas nas imagens foram as de rotação entre 0 e 20 graus, tanto no sentido horário quanto no sentido anti-horário e translação entre 0% e 10%, tanto no eixo horizontal quanto no vertical. O uso dessa técnica proporcionou o aumento das bases em 15 vezes, dado que para cada imagem foram aplicadas 5 rotações diferentes, e para cada uma foi aplicado 3 níveis de translação, passando de 700 amostras para 11200 amostras em cada um dos *datasets* escolhidos.

O aumento de dados foi realizado apenas para o treinamento dos classificadores de faces e de orelhas. O classificador de pessoas foi treinado com a saída dos classificadores referentes às amostras originais.

5.3 Modelos utilizados

Foram definidas duas arquiteturas diferentes para serem criados os três classificadores. Uma delas faz o uso de *transfer learning*, utilizando uma rede já treinada como base, adicionando apenas uma capa de classificação para se adequar à base de dados. Essa arquitetura foi utilizada para construir o classificador de faces e o classificador de orelhas, que são explicados na Subseção 5.3.1. A outra arquitetura é um *multi-layer perceptron* que recebe como entrada a saída de probabilidades dos dois classificadores concatenadas, e gera como saída a probabilidade das amostras a pertencerem a uma determinada classe. Seu detalhamento ocorre na Subseção 5.3.3.

5.3.1 Classificadores de faces e de orelhas

A arquitetura utilizada para realizar a classificação das imagens dos datasets é o modelo pré-treinado com arquitetura VGG-16, mencionada na Seção 4.1.1. Fazendo o uso de *transfer learning*, os pesos dos parâmetros internos foram carregados do modelo VGGFace (PARKHI; VEDALDI; ZISSERMAN, 2015), que fez uso de 2,6 milhões de imagens, contemplando 2,6 mil pessoas em seu treinamento.

Apesar de ser um modelo treinado para o reconhecimento de faces, ele também é utilizado para realizar o reconhecimento de orelhas como um extrator de características, que é possível por meio do treinamento apenas da camada totalmente conectada, que foi reestruturada de acordo com a base de imagens escolhida. Dado que a entrada da rede VGG-16 é de imagens coloridas com resolução de 224×224 pixels, as imagens dos *datasets* foram redimensionadas para atender a esse requisito.

A Figura 5.5 apresenta a estrutura definida para a realização dos experimentos. A Tabela 5.1 apresenta a descrição das camadas VGGFace, e a Tabela 5.2 compara as diferentes camadas entre o modelo VGGFace e o modelo construído através do *fine-tunning*. Tanto o classificador de faces quanto o classificador de orelhas fazem uso da mesma arquitetura proposta, diferindo apenas na camada *softmax*, que possui o número de nós referente ao número de indivíduos do *dataset*, ou seja, 100 nós no classificador de orelhas e 200 nós no classificador de faces.

5.3.2 Seleção de hiperparâmetros

A preparação do modelo para treinamento depende da definição de alguns parâmetros, como a taxa de aprendizado, função de otimização, tamanho de amostras e quantidade de épocas, como mencionado na Seção 4.2.3. Através de tentativa e erro, os melhores hiperparâmetros que foram encontrados são descritos na Tabela 5.3

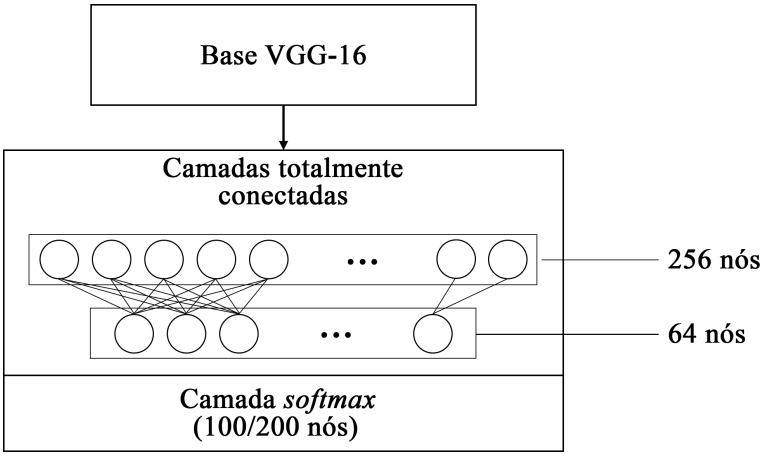

Figura 5.5 – Ilustração do modelo construído para experimentação

Camada	Nome da camada	Tipo da camada	Função de ativação	Detalhes da camada
1	"input"	Imagem de entrada	—	Imagem 224×224
2	"conv1"	Convolução	ReLU	64 convoluções 3×3
3	"conv2"	Convolução	ReLU	64 convoluções 3×3
4	"pool1"	Subamostragem	—	Subamostragem 2×2 com stride 2×2
5	"conv3"	Convolução	ReLU	128 convoluções 3×3
6	"conv4"	Convolução	ReLU	128 convoluções 3×3
7	"pool2"	Subamostragem	—	Subamostragem 2×2 com stride 2×2
8	"conv5"	Convolução	ReLU	256 convoluções 3×3
9	"conv6"	Convolução	ReLU	256 convoluções 3×3
10	"conv7"	Convolução	ReLU	256 convoluções 3×3
11	"pool3"	Subamostragem	—	Subamostragem 2×2 com stride 2×2
12	"conv8"	Convolução	ReLU	512 convoluções 3×3
13	"conv9"	Convolução	ReLU	512 convoluções 3×3
14	"conv10"	Convolução	ReLU	512 convoluções 3×3
15	"pool4"	Subamostragem	—	Subamostragem 2×2 com stride 2×2
16	"conv11"	Convolução	ReLU	512 convoluções 3×3
17	"conv12"	Convolução	ReLU	512 convoluções 3×3
18	"conv13"	Convolução	ReLU	512 convoluções 3×3
19	"pool5"	Subamostragem	—	Subamostragem 2×2 com stride 2×2
20	"conv14"	Convolução	ReLU	4096 convoluções 7×7
21	"conv15"	Convolução	ReLU	4096 convoluções 1×1
22	"conv16"	Convolução	—	2622 convoluções 1×1
23	"prob"	Softmax	—	Softmax com 2622 classes

Tabela 5.1 – Detalhamento das camadas da arquitetura VGGFace

VGGFace					**Modelo construído**				
Camada	**Nome**	**Tipo**	**Ativação**	**Detalhes**	**Camada**	**Nome**	**Tipo**	**Ativação**	**Detalhes**
23	"prob"	Softmax	—	Softmax com 2622 classes	23	"fc1"	Totalmente conectada	ReLU	FC com 256 nós
					24	"fc2"	Totalmente conectada	ReLU	FC com 64 nós
					25	"prob"	Softmax	—	Softmax com 100 nós

Tabela 5.2 – Comparação das capas de classificação entre VGGFace e o modelo construído

Hiperparâmetro	Valor
Taxa de aprendizado	0.01
Nº epochs	40
Tamanho da amostra	8
Otimizador	SGD
Momentum	0.9

Tabela 5.3 – Hiperparâmetros adotados para a fase de treinamento

Tratando de explicações sobre alguns desses parâmetros, o número de épocas foi definido em 40 porque, de acordo com as métricas de performance, o modelo já havia convergido até chegar nesse limite. O tamanho da amostra (*batch*) foi definido em 8 em função do uso de memória durante o treinamento e do sobreajuste que ocorria ao utilizar valores maiores. Valores menores para a taxa de aprendizado faziam com que o modelo demorasse mais para convergir. Apenas utilizar o valor de 0.01 fez com que ele melhorasse a performance mais rapidamente, mas não convergia. Para resolver isso, foi definido um *scheduler* para a taxa de aprendizado, em que ela assume o valor de 0.01 até a época 10, e depois assume o valor igual ao valor atual $\times\ e^{-0,1}$. Dessa forma, o modelo converge mais rápido e, quando se aproxima de um ponto ótimo, ele não o ultrapassa:

$$scheduler = \begin{cases} lr & \text{se } epoch < 10 \\ lr * e^{-0,1} & \text{se } epoch \geq 10 \end{cases}$$

5.3.3 Classificador de pessoas

A arquitetura utilizada para realizar a classificação de pessoas foi um *multi-layer perceptron* com 4 camadas. A camada de entrada recebe um vetor de 200 valores numéricos, que diz respeito ao resultado da concatenação das saídas dos dois classificadores previamente explicados. As duas camadas internas são de 256 e 128 nós respectivamente, e totalmente conectadas. A camada de saída é uma *Softmax* com 100 nós, que são referentes à probabilidade das amostras a pertencerem a cada uma das classes.

A Figura 5.6 ilustra o funcionamento do classificador de pessoas, esclarecendo a sua estrutura interna e como é a entrada de dados na rede. A Tabela 5.4 descreve detalhadamente cada camada que compõe a rede.

Camada	Nome da camada	Tipo da camada	Função de ativação	Detalhes da camada
1	“input”	Entrada	—	200 valores numéricos
2	“fc1”	Totalmente conectada	ReLU	256 nós conectados
3	“fc2”	Totalmente conectada	ReLU	128 nós conectados
4	“prob”	Softmax	—	Softmax com 100 classes

Tabela 5.4 – Detalhamento da arquitetura do classificador de pessoas

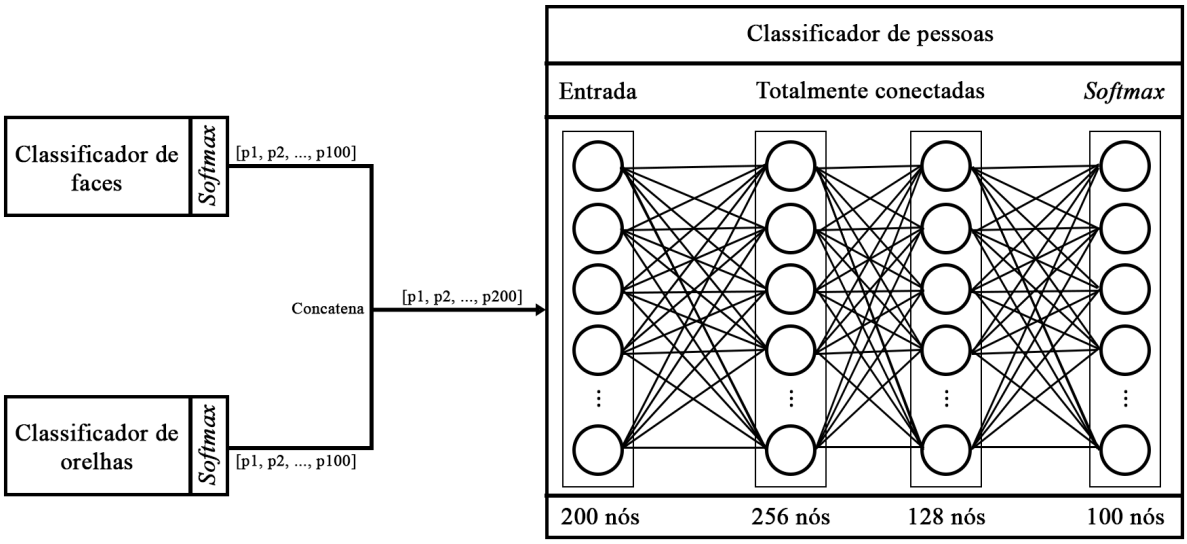

Figura 5.6 – Ilustração do funcionamento do classificador de pessoas

5.3.4 Seleção de hiperparâmetros

Os parâmetros selecionados para esse classificador foram baseados nos utilizados nos outros dois classificadores, com algumas modificações que foram realizadas buscando melhores resultados. Os melhores parâmetros encontrados são descritos na Tabela 5.5.

Hiperparâmetro	**Valor**
Taxa de aprendizado	0.1
Nº epochs	100
Tamanho da amostra	32
Otimizador	SGD
Momentum	0.9

Tabela 5.5 – Hiperparâmetros adotados para a fase de treinamento

O número de épocas foi definido em 100 por proporcionar tempo de treinamento suficiente para o modelo atingir melhores resultados. A taxa de aprendizado foi definida em 0.1 para acelerar a convergência do modelo, e o tamanho da amostra foi definido em 32, por não apresentar sinais de sobreajuste. O otimizador e o momentum foram mantidos por apresentarem resultados superiores aos conseguidos com outros hiperparâmetros durante os experimentos.

Assim como feito anteriormente, foi definido um *scheduler* para a taxa de aprendizado, em que ela assume o valor de 0.1 até a época 100, e depois assume o valor igual ao valor atual $\times\, e^{-0,1}$, como é especificado a seguir:

$$scheduler = \begin{cases} lr & \text{se } epoch < 100 \\ lr * e^{-0,1} & \text{se } epoch \geq 100 \end{cases}$$

5.4 Resultados

Os resultados apresentados nessa seção são referentes a três momentos distintos da fase de experimentação. Em um primeiro momento, foram utilizados os *datasets* obtidos em ambiente controlado para o treinamento dos classificadores baseados em CNNs. Em um segundo momento, foi realizado novamente o treinamento dos classificadores, porém com os *datasets* obtidos em ambiente não controlado. Por fim, foram treinados os classificadores baseados em ViTs com os *datasets* obtidos em ambiente não controlado. No início da experimentação, foi identificada a necessidade de realizar o aumento da base de dados devido aos indícios de sobreajuste no classificador. A partir disso, nos classificadores que ainda seriam treinados, a etapa de aumento de dados já foi realizada logo no ponto de partida.

5.4.1 Experimentos com datasets controlados

Na primeira etapa dos experimentos, foi realizado o treinamento dos classificadores utilizando uma CNN como extrator de características, em que os conjuntos de treinamento e de teste foram divididos de forma que uma divisão aproximada de 85% das amostras fossem destinadas a treinamento, 7,5% a teste e 7,5% a validação.

5.4.1.1 Treinamento com a base AMI Ear original

Inicialmente, nenhum tipo de pré-processamento foi realizado nas amostras, exceto o redimensionamento para que elas atendessem o tamanho da entrada. As acurácias obtidas durante o processo de treinamento, para os conjuntos de treinamento e de validação, são ilustradas na Figura 5.7

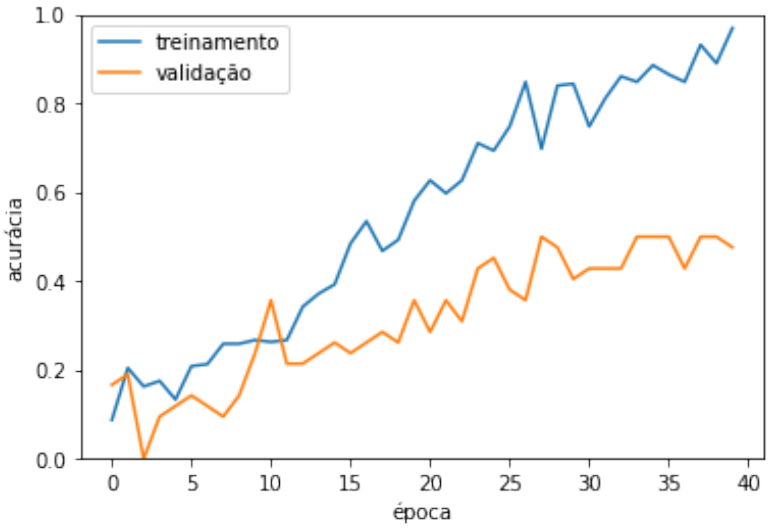

Figura 5.7 – Acurácia do modelo durante o experimento com a base original

É possível notar que a acurácia no conjunto de treinamento quase atingiu a marca de 100%, enquanto a de validação se situou muito abaixo. Isso é um indicativo de que a capacidade de generalização do modelo está baixa. Geralmente, uma forma de resolver esse problema é reduzir o tamanho da amostra nos hiperparâmetros, mas o valor de 8 já é o que traz melhores resultados. Outra forma de resolver é a partir do uso de mais dados para o treinamento, de modo a expor o modelo a diferentes situações para reconhecimento. A partir disso, foi experimentado o treinamento com o aumento de dados, seguindo o que foi descrito na Seção 5.2.

5.4.1.2 Experimentos com a base aumentada

Fazendo uso de *data augmentation*, o modelo é treinado de acordo com imagens da orelha em que ela se encontra em diferentes orientações e posições, permitindo que o modelo tenha uma capacidade melhor de generalização. Assim como no experimento anterior, foi mantida a proporção de 85% e 7,5% e 7,5% para os conjuntos de treinamento, validação e teste, respectivamente. Dessa forma, cada indivíduo do *dataset* conta com 95 amostras de treinamento, 8 amostras de validação e 8 amostras de teste. A Figura 5.8 descreve o progresso de treinamento em relação à acurácia nos conjuntos de treinamento e de validação.

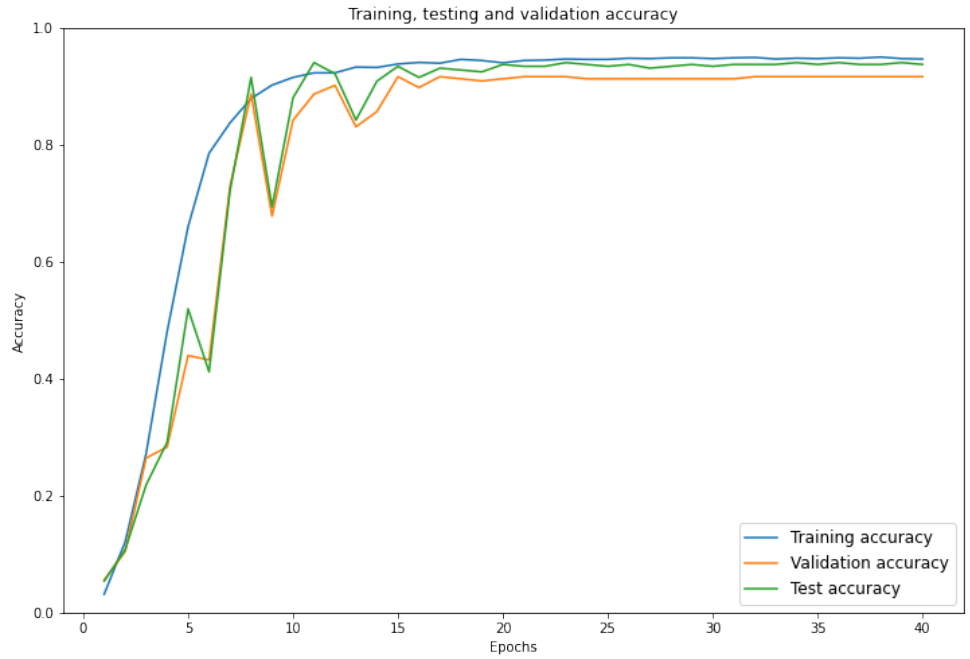

Figura 5.8 – Acurácia do modelo durante o experimento com a base aumentada

Percebe-se que, agora, a acurácia de validação atinge valores desejáveis, aproximando-se de 96% no conjunto de teste. Dessa forma, é possível concluir que a capacidade de generalização do modelo melhorou consideravelmente, não mais havendo sobreajuste. A Tabela 5.6 mostra o comparativo dos resultados dos dois experimentos em termos das métricas apresentadas na Seção 2.6.

5.4.2 Experimentos com o classificador de faces

A experimentação com o classificador de faces se deu diretamente com a base aumentada, já prevendo a ocorrência de sobreajuste como ocorreu com o classificador de orelhas. A divisão de conjuntos de treinamento, validação e teste foi feita seguindo a proporção de 85%, 7,5% e 7,5%, respectivamente, resultando em 95 amostras de treinamento, 8 de validação e 8 de teste por indivíduo. A Figura 5.9 mostra a acurácia do modelo conforme o processo de treinamento progredia:

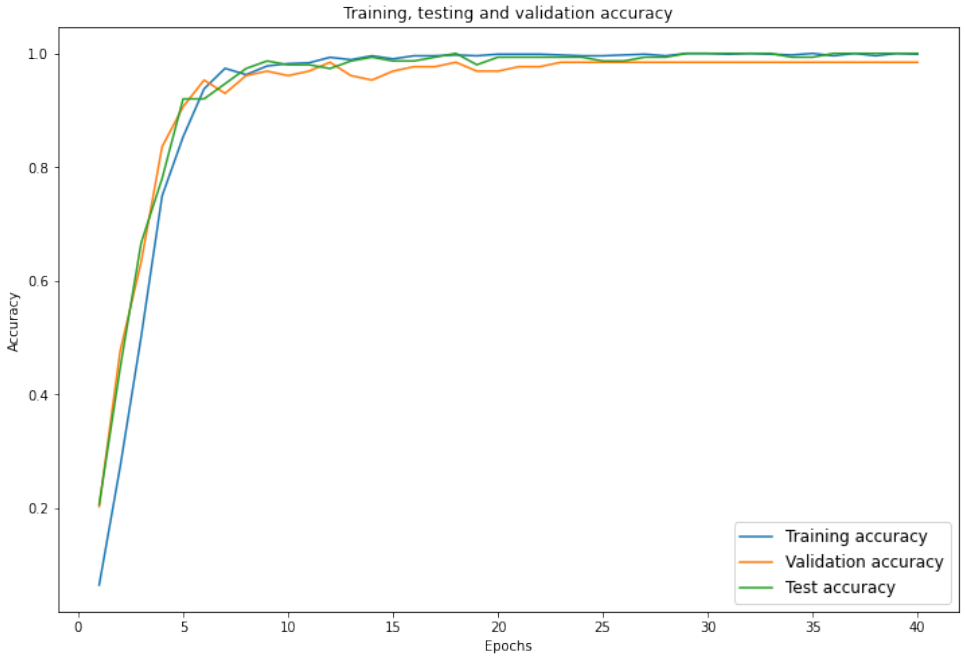

Figura 5.9 – Acurácia do classificador de faces com *dataset* controlado

Através da alta acurácia tanto no conjunto de teste quanto no de treinamento, atingindo valores próximos a 100%, se conclui que não está ocorrendo sobreajuste e o modelo está alcançando bons resultados.

5.4.3 Experimentos com o classificador de pessoas

O treinamento utilizando o dataset artificial foi feito utilizando a saída direta dos outros dois classificadores, sem nenhum pré-processamento nos dados de entrada. Foi feita a divisão das amostras seguindo as mesmas proporções utilizadas nos outros modelos. A Figura 5.10 mostra o progresso do modelo em relação à acurácia:

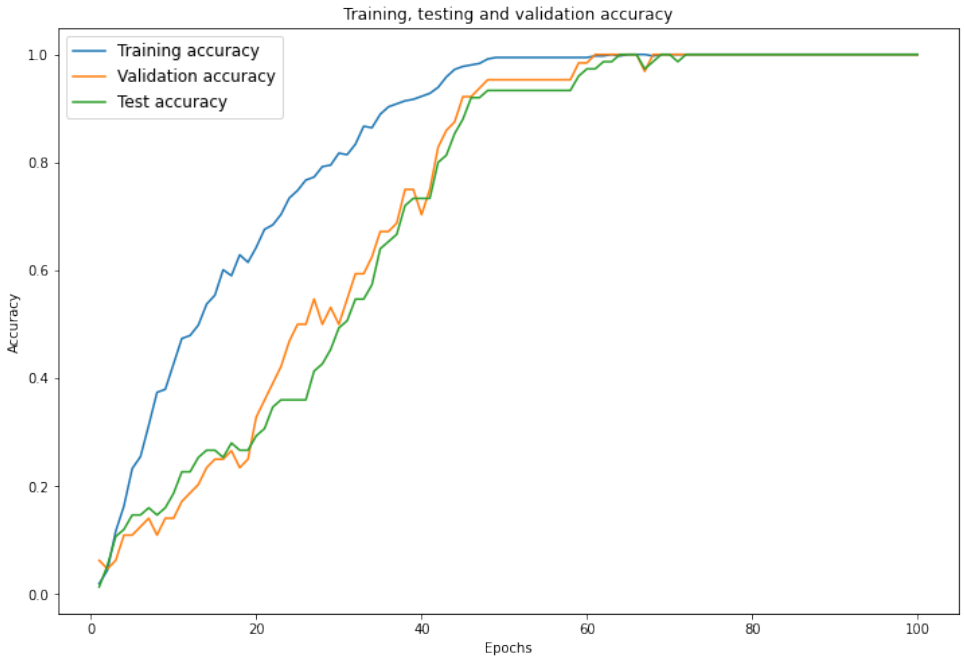

Figura 5.10 – Acurácia do classificador de pessoas com *dataset* controlado

É possível notar que o modelo, após 72 épocas, convergiu para 100% de acurácia em todos os conjuntos.

A Tabela 5.6 mostra o comparativo entre as diferentes etapas do processo de experimentação em relação às métricas de acurácia, precisão e revocação:

Métrica	**Orelha (aumentada)**	**Face**	**Pessoas**
Acurácia	95,68%	98,66%	100%
Precisão	95,73%	98,97%	100%
Revocação	96,22%	99,07%	100%

Tabela 5.6 – Comparativo de métricas entre os experimentos realizados

5.4.4 Experimentos com datasets não controlados

Nesta etapa, a princípio, nenhum hiperparâmetro foi modificado para trabalhar com os novos *datasets*. No entanto, a partir da análise da acurácia dos três conjuntos de dados, foi notado que estava ocorrendo o sobreajuste do modelo. Diferentes estratégias foram tomadas para melhorar a capacidade de generalização do modelo, como a alteração no tamanho da amostra, da taxa de aprendizado, na arquitetura do modelo, como alterar, adicionar e remover camadas densas, incluir camadas de *drop-out* e criar mais amostras com *data augmentation*. De todas as tentativas, o sobreajuste foi parcialmente resolvido, mas não completamente. Os melhores resultados foram obtidos ao manter a arquitetura já utilizada, porém adotar novos hiperparâmetros, exibidos na Tabela 5.7:

Hiperparâmetro	**Valor**
Taxa de aprendizado	0.1
Nº epochs	100
Tamanho da amostra	32
Otimizador	SGD
Momentum	0.9

Tabela 5.7 – Novos hiperparâmetros adotados para treinamento

A taxa de aprendizado foi aumentada de 0.01 para 0.1, o número de *epochs* foi aumentado de 40 para 100, e o tamanho da amostra foi aumentado de 8 para 32. Essa configuração foi utilizada para o treinamento dos classificadores de faces e de orelhas. Já no classificador de pessoas, as configurações de hiperparâmetros foi mantida.

5.4.4.1 Classificador de orelhas

Como esperado ao trabalhar com *datasets* não controlados em comparação aos controlados, a acurácia diminuiu consideravelmente, com o conjunto de teste atingindo aproximadamente 63%, e o de treinamento próximo a 69%, como é exibido na Figura 5.11

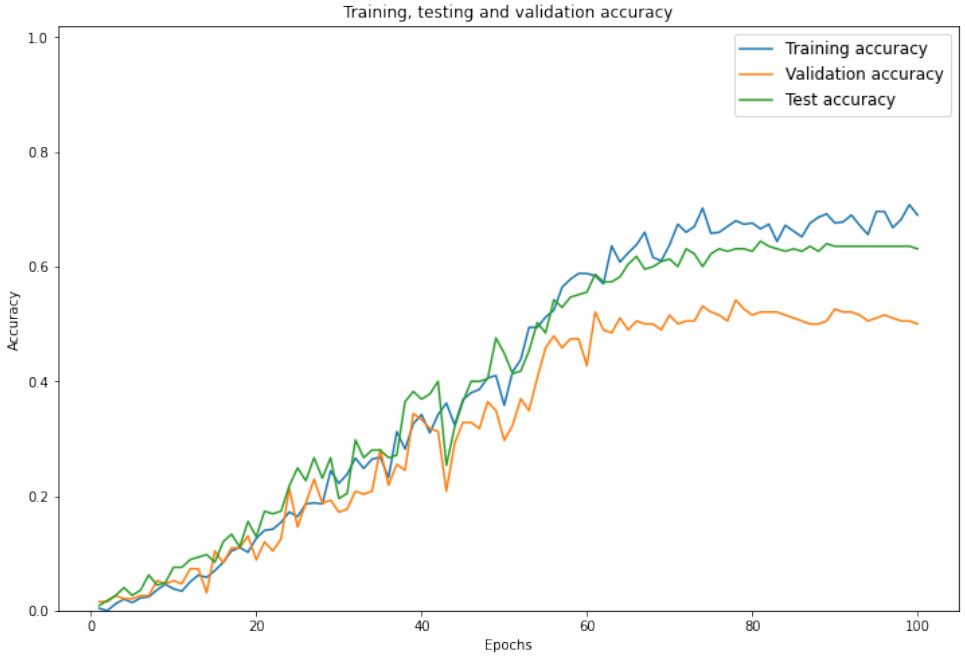

Figura 5.11 – Acurácia do classificador de orelhas com *dataset* não controlado

É possível notar que por volta da época 70, o modelo atingiu um ponto em que melhorias significativas não eram mais presentes.

5.4.4.2 Classificador de faces

Como mencionado, o sobreajuste não foi completamente eliminado pelos ajustes nos hiperparâmetros dos modelos. Isso pode ser notado também no classificador de faces, que atingiu quase 100% de acurácia no conjunto de treinamento, e estabilizou próximo a 90% nos conjuntos de validação e de teste, como é mostrado na Figura 5.12

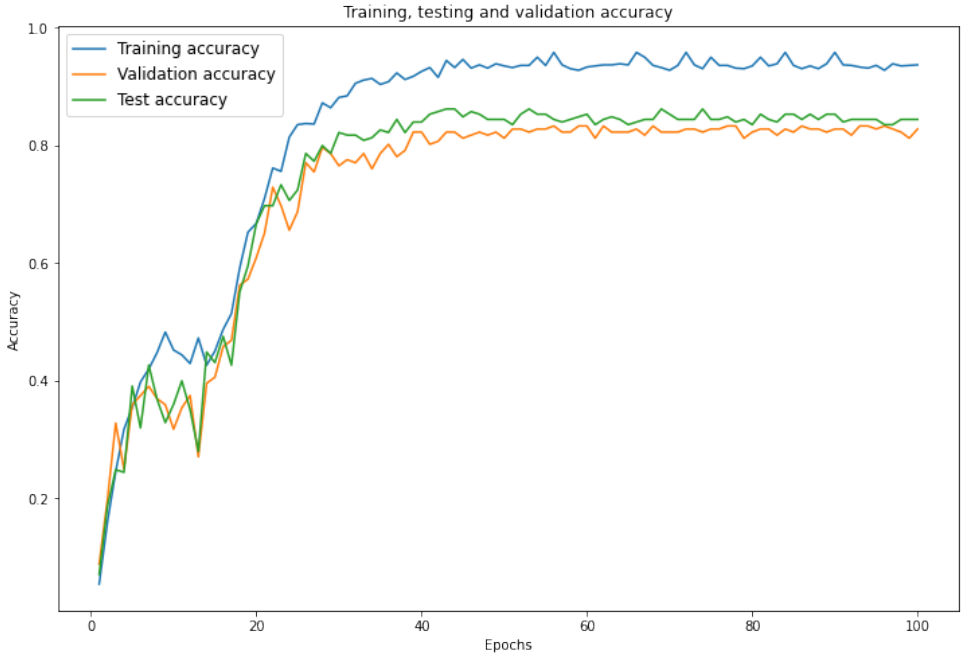

Figura 5.12 – Acurácia do classificador de faces com *dataset* não controlado

É notável que a convergência do modelo ocorreu de forma mais abrupta em relação ao classificador de orelhas.

5.4.4.3 Classificador de pessoas

Apesar dos resultados adversos nos dois classificadores, o classificador de pessoas foi capaz de elevar a performance do modelo como um todo, atingindo 100% de acurácia nos conjuntos de treinamento e de teste, como a Figura 5.13 ilustra:

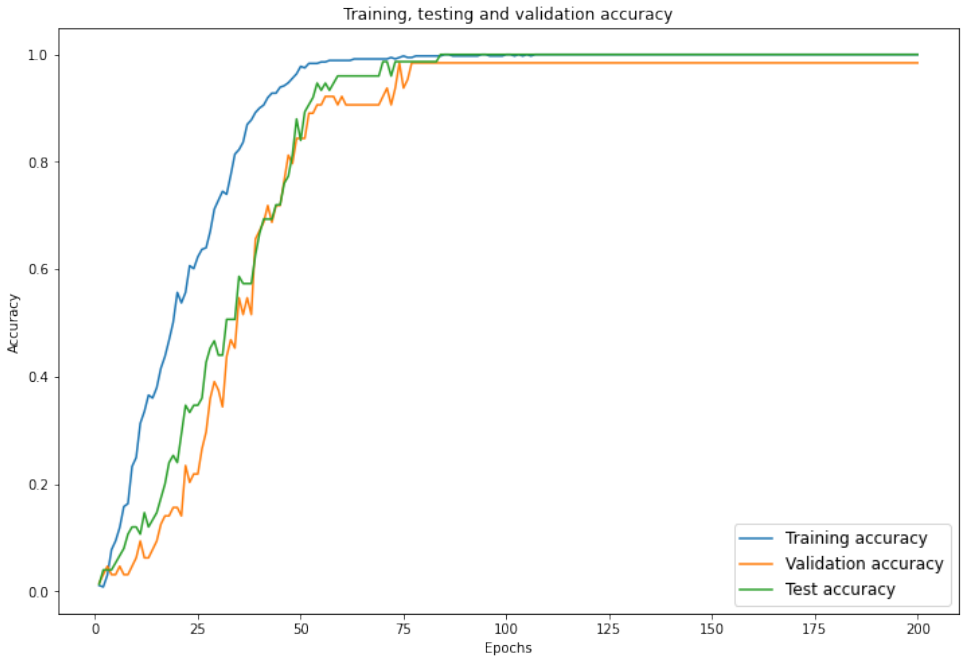

Figura 5.13 – Acurácia do classificador de pessoas com *dataset* não controlado

Embora o modelo tenha demorado cerca de 100 épocas para atingir um platô, a combinação dos dois classificadores utilizando uma CNN produziu resultados satisfatórios. A Figura 5.14 mostra a acurácia atingida nos conjuntos de teste por cada um dos classificadores durante o processo de treinamento:

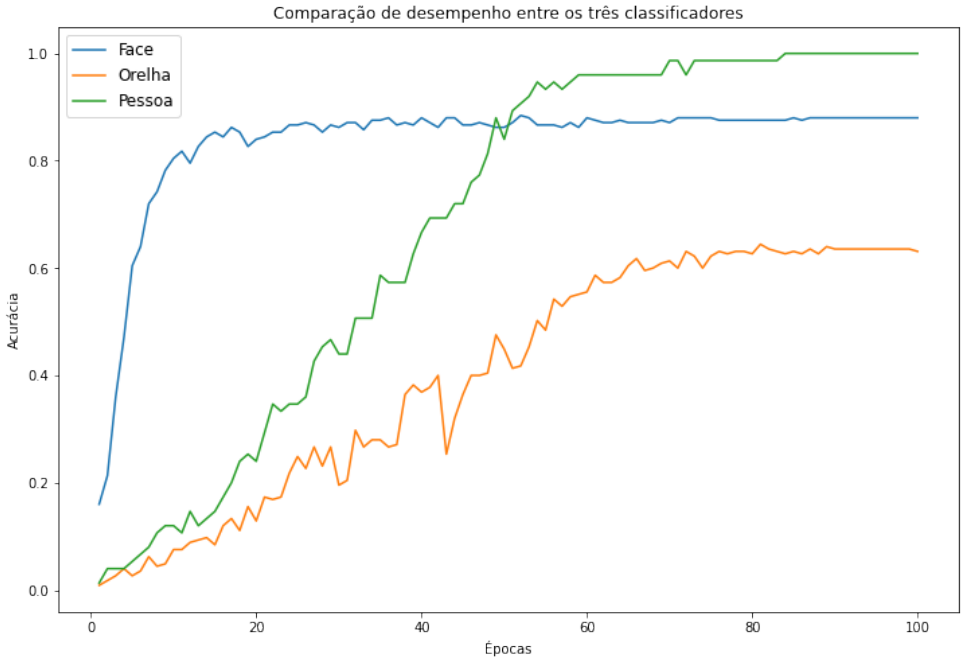

Figura 5.14 – Comparação da acurácia dos três classificadores em *datasets* não controlados

5.4.5 Experimentos com Vision Transformers

A experimentação com os ViTs ocorreu de modo que apenas as camadas do modelo pré-treinado VGGFace fossem removidas e, em seu lugar, fossem colocadas as camadas do modelo pré-treinado ViT-B/32. Assim como nos modelos baseados em CNN, os modelos estavam sobreajustado aos dados de treinamento, o que levou à alteração dos hiperparâmetros definidos. A taxa de aprendizado foi reduzida de 0.1 para 0.05, o tamanho da amostra foi reduzido de 32 para 8, e nas camadas totalmente conectadas, foram adicionadas uma camada de *drop-out* com 30% de taxa de abandono e uma camada densa de 128 nós. Essa configuração foi adotada tanto para o classificador de faces quanto para o classificador de orelhas. Os hiperparâmetros do classificador de pessoas foram mantidos.

As figuras a seguir mostram o progresso do treinamento dos três classificadores. É importante mencionar que todos atingem 100% de acurácia no conjunto de treinamento, sendo que nos classificadores de faces e de orelhas, demonstrados nas Figuras 5.15 e 5.16, respectivamente, é notável que ocorre o sobreajuste, dado que há um intervalo considerável entre a acurácia do conjunto de treinamento em relação aos conjuntos de validação e teste:

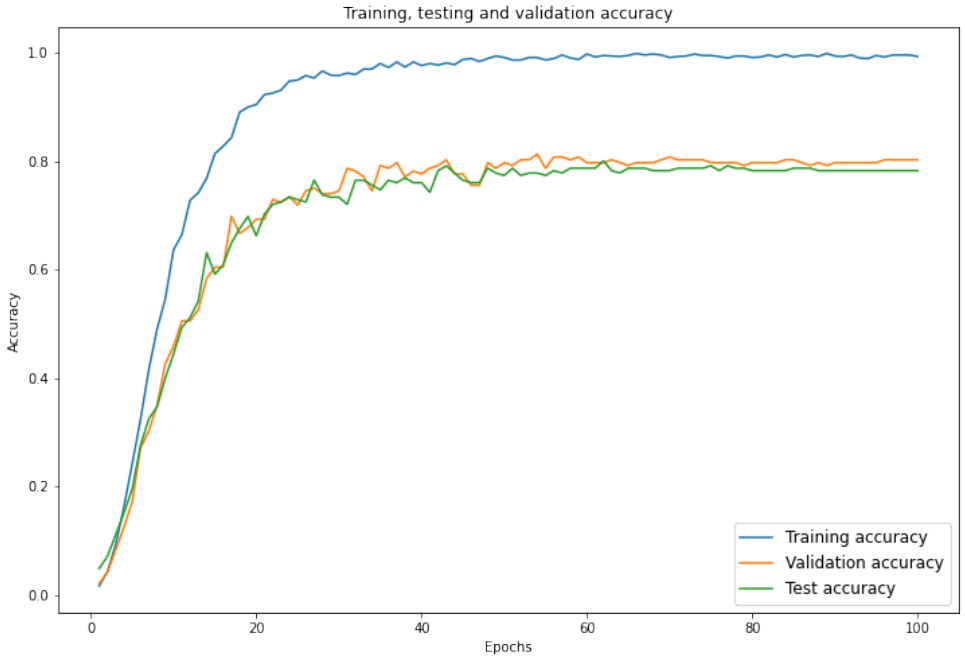

Figura 5.15 – Acurácia do classificador de orelhas baseado em ViT

Também se nota que o classificador de orelhas demonstra um desempenho substancialmente superior ao classificador de orelhas baseado em CNN.

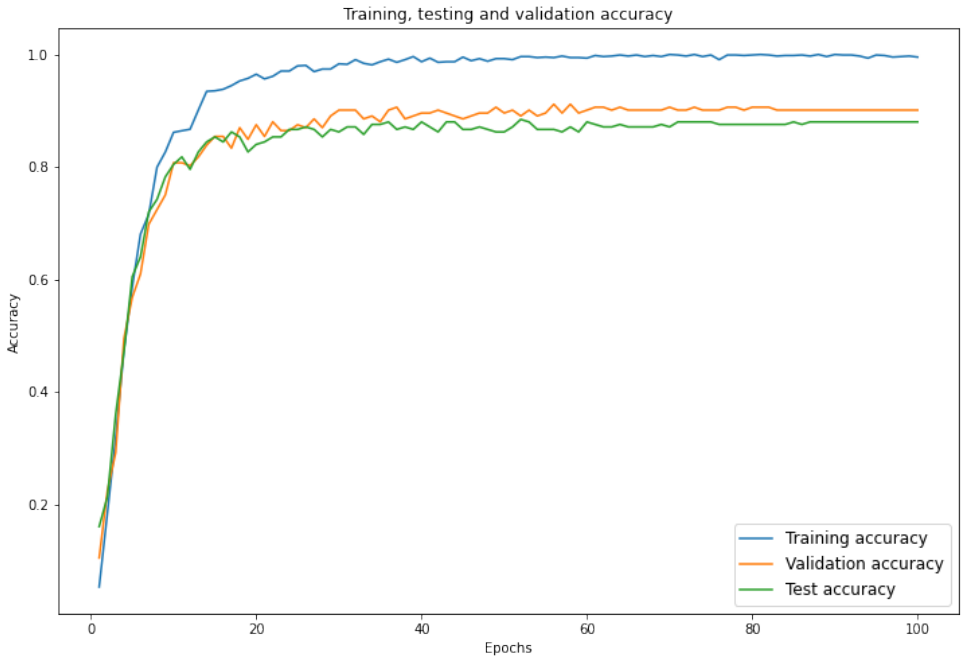

Figura 5.16 – Acurácia do classificador de faces baseado em ViT

No geral, os modelos convergiram mais rapidamente. Por fim, nota-se que o classificador de pessoas atingiu a marca de 100% de acurácia nos conjuntos de treinamento, validação e teste, como é exibido na Figura 5.17.

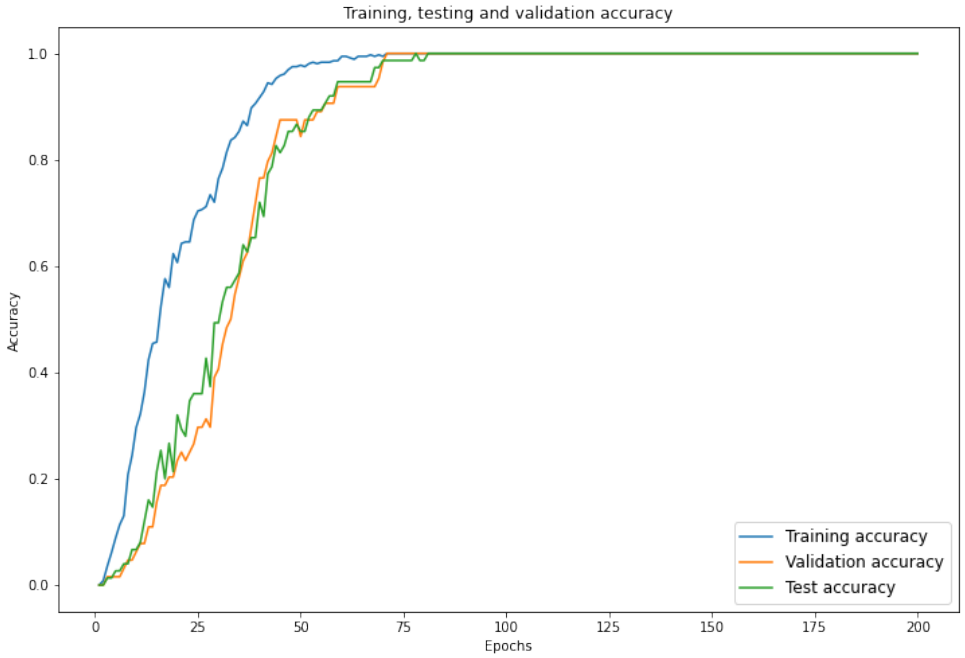

Figura 5.17 – Acurácia do classificador de pessoas baseado em ViT

Por fim, a Figura 5.18 sumariza o desempenho dos diferentes classificadores tomando como referência a acurácia dos conjuntos de teste:

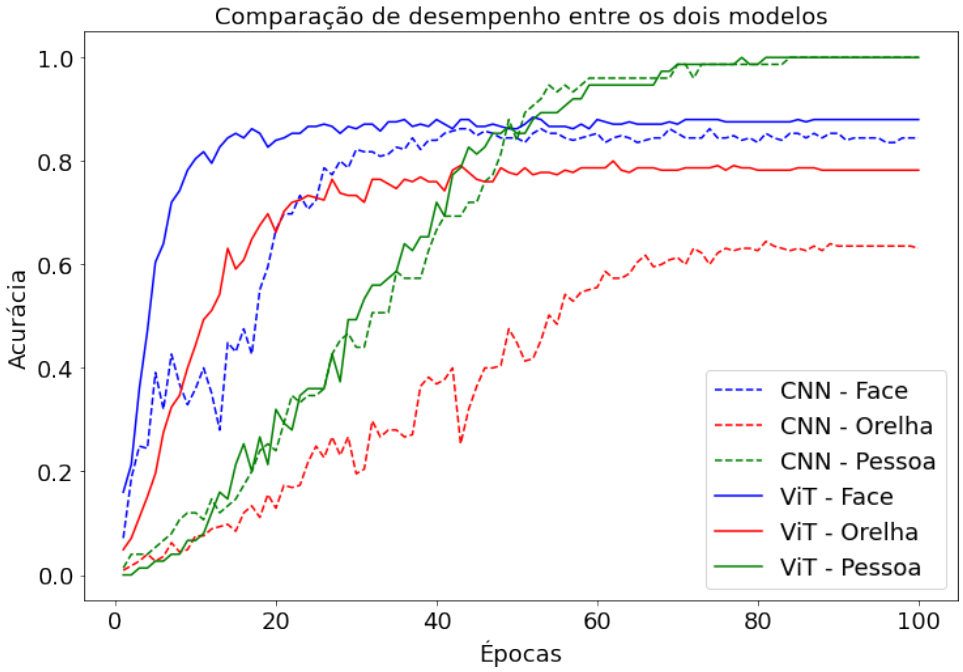

Figura 5.18 – Comparação entre os dois modelos treinados com os *datasets* não controlados

Finalmente, a Tabela 5.8 mostra a acurácia, precisão e revocação de cada um dos classificadores dos dois modelos criados.

Métrica	**CNN**			**ViT**		
	Orelha	**Face**	**Pessoas**	**Orelha**	**Face**	**Pessoas**
Acurácia	63,10%	85,33%	100,00%	78,22%	88,00%	100,00%
Precisão	63,27%	85,31%	100,00%	78,77%	88,14%	100,00%
Revocação	63,90%	86,04%	100,00%	78,16%	89,40%	100,00%

Tabela 5.8 – Métricas de performance do dois modelos

Além da convergência mais rápida dos modelos baseados em ViT em relação aos baseados em CNN, por se tratarem de arquiteturas com número de parâmetros a serem treinados significativamente menores, o tempo de processamento para realizar o treinamento do modelo foi consideravelmente menor tanto para o classificador de orelhas quanto para o classificador de faces. Como ambos os classificadores fazem uso da mesma arquitetura, seja baseada em CNN ou em ViT, o tempo de treinamento de cada classificador foi o mesmo, em função da base. O tempo médio gasto por época ao treinar os classificadores baseados em CNN foi de aproximadamente 11,56 segundos, em contraponto a 5,68 segundos no treinamento dos classificadores baseados em ViT, configurando uma aceleração de 103,52% no processo de treinamento.

6 Conclusão

Neste trabalho foi realizado um estudo sobre as diferentes técnicas utilizadas no contexto de biometria da face e biometria da orelha. Esse estudo foi motivado pelo fato de que a orelha se mostra um bom candidato para o processo biométrico, dado que ela possui características que são únicas de pessoa para pessoa, além de que a obtenção de imagens da orelha ocorre de forma não intrusiva. Ainda, a pandemia da Covid-19 resultou no uso de máscaras pela população, dificultando o reconhecimento apenas pela face. A partir disso, unir o reconhecimento facial com o reconhecimento da orelha pode trazer melhores resultados.

Foi proposta uma arquitetura para um modelo de inteligência artificial para realizar o reconhecimento de pessoas através da face e da orelha utilizando redes neurais. Neste trabalho, foi possível construir dois modelos de inteligência artificial, cada um composto por três classificadores capazes de reconhecer quem é um indivíduo baseado apenas na sua orelha, na sua face, ou em ambos, com alta precisão, pelos datasets controlados AMI Ear e FEI Face, e pelos datasets não controlados LFW Face e VGGFace-Ear. O classificador baseado em CNN que reconhece através da orelha atingiu 95% de acurácia, o que reconhece através da face atingiu 98%, e o que une ambos atingiu 100% de acurácia, todos sobre os datasets controlados. Nos datasets não controlados, o que reconhece pela orelha atingiu 63% de acurácia, o que reconhece pela face atingiu 85%, e o que combina os resultados de ambos atingiu 100%. Já o modelo que tem seus classificadores baseados em ViTs, o treinamento foi realizado apenas com datasets não controlados, em que o classificador de orelhas obteve 78% de acurácia, o de faces obteve 88% e o de pessoas obteve 100%. Foram criados dois modelos, um baseado em CNNs para extrair as características da imagem, baseadas na rede VGGFace utilizando *transfer learning*, e outro baseado em ViTs, baseados no modelo ViT-B/32, em que ambos agem como classificadores de pessoas a partir da orelha e da face, que estão de acordo com dois dos objetivos específicos expostos no início desse texto. Foi também criado um *multi-layer perceptron* para combinar os resultados das duas CNNs e uma única rede e realizar o reconhecimento, o que configura a conclusão do terceiro objetivo específico proposto. A criação do modelo baseado em ViTs contempla o quatro e último objetivo específico colocado.

A partir dos resultados alcançados, é possível concluir que o uso de uma rede neural para a combinação de resultados de múltiplos modelos é uma técnica viável e alternativa ao uso de classificadores como SVM e K-Nearest Neighbors (KNN), dado que a performance alcançada por ela supera a performance dos modelos individuais. Também é notável que os *Vision Transformers* trazem excelentes resultados, dado que, nesse trabalho, superou o desempenho de modelos pré-treinados baseados em CNNs.

O código desenvolvido e os experimentos realizados, juntamente do estudo e da funda-

mentação teórica agregada neste documento selam o propósito desse trabalho.

6.1 Trabalhos Futuros

Visando a continuidade do desenvolvimento desse trabalho, espera-se que os modelos sejam treinados em conjuntos maiores de dados, propiciando um ambiente mais desafiador a ser enfrentado. Também é esperado que outras estratégias para a melhoria dos modelos sejam experimentadas, dado que alguns classificadores apresentam sobreajuste sobre os conjuntos de dados utilizados.

Desejando que este trabalho tenha maior utilidade em cenários do mundo real, a portabilização dos modelos criados para o funcionamento em tempo real pode permitir que ele seja utilizado em imagens de vídeo, ampliando ainda mais os seus casos de uso.

Por fim, modelo pode ser adaptado para funcionar com um conjunto aberto de indivíduos, permitindo que novas classes sejam adicionadas ao modelo sem a necessidade de realizar o processo de treinamento novamente. Nesse caso, pode ser importante que diferentes métricas sejam utilizadas para avaliar o desempenho do modelo.

Referências

ALEJO, M. B. Unconstrained ear recognition using transformers. *Jordanian Journal of Computers and Information Technology*, Scientific Research Support Fund of Jordan Princess Sumaya University for . . . , v. 7, n. 4, 2021.

ALMISREB, A. A.; JAMIL, N.; DIN, N. M. Utilizing alexnet deep transfer learning for ear recognition. In: IEEE. *2018 Fourth International Conference on Information Retrieval and Knowledge Management (CAMP)*. [S.l.], 2018. p. 1–5.

ALSHAZLY, H.; LINSE, C.; BARTH, E.; MARTINETZ, T. Ensembles of deep learning models and transfer learning for ear recognition. *Sensors*, MDPI, v. 19, n. 19, p. 4139, 2019.

BOWYER, K. W.; CHANG, K. I.; YAN, P.; FLYNN, P. J.; HANSLEY, E.; SARKAR, S. Multi-modal biometrics: an overview. In: *Second workshop on multi-modal user authentication*. [S.l.: s.n.], 2006. v. 105.

CANNY, J. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, Ieee, n. 6, p. 679–698, 1986.

CÁRDENAS, R. J.; BELTRÁN, C. A.; GUTIÉRREZ, J. C. Small face detection using deep learning on surveillance videos. *environment*, v. 2, n. 5, p. 14, 2019.

CHANG, K.; BOWYER, K. W.; SARKAR, S.; VICTOR, B. Comparison and combination of ear and face images in appearance-based biometrics. *IEEE Transactions on pattern analysis and machine intelligence*, IEEE, v. 25, n. 9, p. 1160–1165, 2003.

DADI, H. S.; PILLUTLA, G. K. M.; MAKKENA, M. L. Face recognition and human tracking using gmm, hog and svm in surveillance videos. *Annals of Data Science*, Springer, v. 5, n. 2, p. 157–179, 2018.

DAMER, N.; FÜHRER, B. Ear recognition using multi-scale histogram of oriented gradients. In: IEEE. *2012 Eighth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*. [S.l.], 2012. p. 21–24.

DHARAVATH, K.; TALUKDAR, F. A.; LASKAR, R. H. Improving face recognition rate with image preprocessing. *Indian Journal of Science and Technology*, Citeseer, v. 7, n. 8, p. 1170–1175, 2014.

DODGE, S.; MOUNSEF, J.; KARAM, L. Unconstrained ear recognition using deep neural networks. *IET Biometrics*, Wiley Online Library, v. 7, n. 3, p. 207–214, 2018.

DOSOVITSKIY, A.; BEYER, L.; KOLESNIKOV, A.; WEISSENBORN, D.; ZHAI, X.; UNTERTHINER, T.; DEHGHANI, M.; MINDERER, M.; HEIGOLD, G.; GELLY, S. et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

EMERŠIČ, Ž.; KRIŽAJ, J.; ŠTRUC, V.; PEER, P. Deep ear recognition pipeline. In: *Recent Advances in Computer Vision*. [S.l.]: Springer, 2019. p. 333–362.

FU, K.-S. et al. Pattern recognition and image processing. *IEEE transactions on computers*, IEEE, v. 100, n. 12, p. 1336–1346, 1976.

GE, S.; LI, J.; YE, Q.; LUO, Z. Detecting masked faces in the wild with lle-cnns. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2017. p. 2682–2690.

GENG, C.; JIANG, X. Face recognition using sift features. In: IEEE. *2009 16th IEEE international conference on image processing (ICIP)*. [S.l.], 2009. p. 3313–3316.

GEORGESCU, M.-I.; IONESCU, R. T.; POPESCU, M. Local learning with deep and handcrafted features for facial expression recognition. *IEEE Access*, IEEE, v. 7, p. 64827–64836, 2019.

GHOUALMI, L.; DRAA, A.; CHIKHI, S. An ear biometric system based on artificial bees and the scale invariant feature transform. *Expert Systems with Applications*, Elsevier, v. 57, p. 49–61, 2016.

GONZALEZ, L. A. E.; MAZORRA, L. *AMI Ear Database*. 2012. Disponível em: <https://ctim.ulpgc.es/research_works/ami_ear_database/>.

HAN, K.; XIAO, A.; WU, E.; GUO, J.; XU, C.; WANG, Y. Transformer in transformer. *Advances in Neural Information Processing Systems*, v. 34, p. 15908–15919, 2021.

HUANG, G. B.; RAMESH, M.; BERG, T.; LEARNED-MILLER, E. *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments*. [S.l.], 2007.

HURLEY, D. J.; NIXON, M. S.; CARTER, J. N. A new force field transform for ear and face recognition. In: IEEE. *Proceedings 2000 International Conference on Image Processing (Cat. No. 00CH37101)*. [S.l.], 2000. v. 1, p. 25–28.

IANNARELLI, A. *Ear Identification*. Fremont, California: Paramont Publishing Company, 1989.

JAIN, A. K.; PANKANTI, S.; PRABHAKAR, S.; HONG, L.; ROSS, A. Biometrics: a grand challenge. In: IEEE. *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.* [S.l.], 2004. v. 2, p. 935–942.

LI, Z.; CHEN, Z.; YANG, F.; LI, W.; ZHU, Y.; ZHAO, C.; DENG, R.; WU, L.; ZHAO, R.; TANG, M. et al. Mst: Masked self-supervised transformer for visual representation. *Advances in Neural Information Processing Systems*, v. 34, p. 13165–13176, 2021.

LU, L.; ZHANG, X.; ZHAO, Y.; JIA, Y. Ear recognition based on statistical shape model. In: IEEE. *First International Conference on Innovative Computing, Information and Control-Volume I (ICICIC'06)*. [S.l.], 2006. v. 3, p. 353–356.

LV, J.-J.; SHAO, X.-H.; HUANG, J.-S.; ZHOU, X.-D.; ZHOU, X. Data augmentation for face recognition. *Neurocomputing*, Elsevier, v. 230, p. 184–196, 2017.

MAHOOR, M. H.; CADAVID, S.; ABDEL-MOTTALEB, M. Multi-modal ear and face modeling and recognition. In: IEEE. *2009 16th IEEE International Conference on Image Processing (ICIP)*. [S.l.], 2009. p. 4137–4140.

NIGAM, A.; GUPTA, P. Robust ear recognition using gradient ordinal relationship pattern. In: SPRINGER. *Asian conference on computer vision*. [S.l.], 2014. p. 617–632.

PARKHI, O. M.; VEDALDI, A.; ZISSERMAN, A. Deep face recognition. British Machine Vision Association, 2015.

PATEL, K.; HAN, H.; JAIN, A. K. Secure face unlock: Spoof detection on smartphones. *IEEE transactions on information forensics and security*, IEEE, v. 11, n. 10, p. 2268–2283, 2016.

PRIYADHARSHINI, R. A.; ARIVAZHAGAN, S.; ARUN, M. A deep learning approach for person identification using ear biometrics. *Applied intelligence*, Springer, v. 51, n. 4, p. 2161–2172, 2021.

RAMOS-COOPER, S.; GOMEZ-NIETO, E.; CAMARA-CHAVEZ, G. Vggface-ear: An extended dataset for unconstrained ear recognition. *Sensors*, v. 22, n. 5, 2022. ISSN 1424-8220. Disponível em: <https://www.mdpi.com/1424-8220/22/5/1752>.

RAUSS, P. J.; PHILLIPS, J.; MOON, H.; RIZVI, S. A.; HAMILTON, M. K.; DEPERSIA, A. T. Feret (face recognition technology) program. In: SPIE. *Surveillance and Assessment Technologies for Law Enforcement*. [S.l.], 1997. v. 2935, p. 2–11.

SANA, A.; GUPTA, P.; PURKAIT, R. Ear biometrics: A new approach. In: *Advances in pattern recognition*. [S.l.]: World Scientific, 2007. p. 46–50.

SHARIF, M.; RAZA, M.; SHAH, J. H.; YASMIN, M.; FERNANDES, S. L. An overview of biometrics methods. *Handbook of Multimedia Information Security: Techniques and Applications*, Springer, p. 15–35, 2019.

SHOBA, V.; SAM, I. S. A hybrid features extraction on face for efficient face recognition. *Multimedia Tools and Applications*, Springer, v. 79, n. 31, p. 22595–22616, 2020.

SINGH, R.; AGARWAL, A.; SINGH, M.; NAGPAL, S.; VATSA, M. On the robustness of face recognition algorithms against attacks and bias. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. [S.l.: s.n.], 2020. v. 34, n. 09, p. 13583–13589.

SRIVASTAVA, N.; HINTON, G.; KRIZHEVSKY, A.; SUTSKEVER, I.; SALAKHUTDINOV, R. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, v. 15, n. 56, p. 1929–1958, 2014. Disponível em: <http://jmlr.org/papers/v15/srivastava14a.html>.

STACKEXCHANGE. *Deep VGG-16 Net*. 2022. Disponível em: <https://ai.stackexchange.com/questions/35289/what-is-the-time-complexity-of-deep-vgg-16-net>.

SU, W.; WANG, Y.; LI, K.; GAO, P.; QIAO, Y. Hybrid token transformer for deep face recognition. *Pattern Recognition*, Elsevier, v. 139, p. 109443, 2023.

TATHE, S.; NAROTE, A.; NAROTE, S. Face detection and recognition in videos. In: IEEE. *2016 IEEE Annual India Conference (INDICON)*. [S.l.], 2016. p. 1–6.

THOMAS, C. E. *FEI Face Database*. 2006. Disponível em: <https://fei.edu.br/~cet/facedatabase.html>.

TIAN, L.; MU, Z. Ear recognition based on deep convolutional network. In: IEEE. *2016 9th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*. [S.l.], 2016. p. 437–441.

UOL. *Orelha externa*. 2015. Disponível em: <https://drauziovarella.uol.com.br/corpo-humano/orelha-externa/>.

VICTOR, B.; BOWYER, K.; SARKAR, S. An evaluation of face and ear biometrics. In: IEEE. *2002 International Conference on Pattern Recognition*. [S.l.], 2002. v. 1, p. 429–432.

WANG, Y.; YANG, Z.; ZHANG, Z.; ZANG, H.; ZHU, Q.; ZHAN, S. Learning 3d face representation with vision transformer for masked face recognition. In: IEEE. *2022 Asia Conference on Algorithms, Computing and Machine Learning (CACML)*. [S.l.], 2022. p. 505–511.

WAYMAN, J.; JAIN, A.; MALTONI, D.; MAIO, D. An introduction to biometric authentication systems. *Biometric systems: Technology, design and performance evaluation*, Springer, p. 1–20, 2005.

XU, J. A deep learning approach to building an intelligent video surveillance system. *Multimedia Tools and Applications*, Springer, v. 80, n. 4, p. 5495–5515, 2021.

XU, X.-N.; MU, Z.-C.; YUAN, L. Feature-level fusion method based on kfda for multimodal recognition fusing ear and profile face. In: IEEE. *2007 International Conference on Wavelet Analysis and Pattern Recognition*. [S.l.], 2007. v. 3, p. 1306–1310.

YAN, P.; BOWYER, K. W. Multi-biometrics 2d and 3d ear recognition. In: SPRINGER. *International Conference on Audio-and Video-Based Biometric Person Authentication*. [S.l.], 2005. p. 503–512.

YUAN, L.; CHEN, Y.; WANG, T.; YU, W.; SHI, Y.; JIANG, Z.-H.; TAY, F. E.; FENG, J.; YAN, S. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In: *Proceedings of the IEEE/CVF international conference on computer vision*. [S.l.: s.n.], 2021. p. 558–567.

ZUIDERVELD, K. Contrast limited adaptive histogram equalization. *Graphics gems*, Academic Press, p. 474–485, 1994.