UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO

CENTRO DE CIÊNCIAS EXATAS, NATURAIS E DA SAÚDE

COLEGIADO DO CURSO DE SISTEMAS DE INFORMAÇÃO

LAYON JOSÉ PEDROSA DOS REIS

SISTEMA AUTOMÁTICO DE CONTROLE DE ACESSO POR RECONHECIMENTO FACIAL

ALEGRE, ES

2026

LAYON JOSÉ PEDROSA DOS REIS

SISTEMA AUTOMÁTICO DE CONTROLE DE ACESSO POR RECONHECIMENTO FACIAL

Trabalho de Conclusão de Curso apresentado ao Colegiado do Curso de Sistemas de Informação do Centro de Ciências Exatas, Naturais e da Saúde da Universidade Federal do Espírito Santo, como requisito para aprovação na Disciplina Trabalho de Conclusão de Curso em Sistemas de Informação II.

Universidade Federal do Espírito Santo – UFES

Centro de Ciências Exatas, Naturais e da Saúde

Colegiado do Curso de Sistemas de Informação

Orientador: Jacson Rodrigues Correia da Silva

ALEGRE, ES

2026

Layon José Pedrosa Dos Reis

SISTEMA AUTOMÁTICO DE CONTROLE DE ACESSO POR RECONHECIMENTO FACIAL

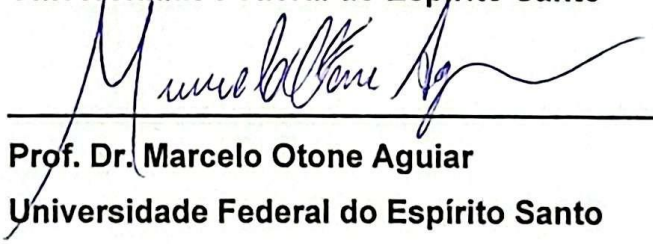
Trabalho de Conclusão de Curso apresentado ao Departamento de Computação do Centro de Ciências Exatas, Naturais e da Saúde da Universidade Federal do Espírito Santo, como requisito parcial para obtenção do grau de Bacharel em Sistemas de Informação.

Aprovado em 12 de fevereiro de 2026.

COMISSÃO EXAMINADORA

Prof. Dr. Jacson Rodrigues Correia da Silva
Universidade Federal do Espírito Santo
Orientador

Prof. Dr. Dayan de Castro Bissoli
Universidade Federal do Espírito Santo



Prof. Dr. Marcelo Otone Aguiar
Universidade Federal do Espírito Santo

Agradecimentos

Gostaria de expressar meus sinceros agradecimentos a todos que contribuíram para a realização deste trabalho.

Agradeço primeiramente à minha família, pelo apoio incondicional e incentivo ao longo de toda minha trajetória acadêmica.

Agradeço de maneira especial a minha namorada, Vítoria, pelo amor, paciência e apoio constantes ao longo de toda a realização deste trabalho. Sua compreensão nos momentos de dedicação intensa e suas palavras de incentivo foram fundamentais para que eu pudesse superar os desafios desta jornada acadêmica.

Aos professores e orientadores, em especial ao meu orientador, Jacson Rodrigues Correia da Silva, pela orientação, paciência e ensinamentos compartilhados, que foram fundamentais para o desenvolvimento deste trabalho.

Aos colegas e amigos, pelas discussões construtivas, apoio e momentos de convivência que tornaram esta jornada mais leve e agradável.

Agradeço também à instituição de ensino, Universidade Federal do Espírito Santo, pelo suporte e infraestrutura disponibilizados para a realização deste TCC.

Por fim, agradeço a todos que, de alguma forma, contribuíram direta ou indiretamente para a conclusão deste trabalho.

"We can only see a short distance ahead,
but we can see plenty there that needs to be done."

– Alan Turing, 1950

RESUMO

Situada na interseção entre Visão Computacional e Segurança da Informação, esta pesquisa aborda a limitação dos métodos tradicionais de credenciamento físico, frequentemente suscetíveis a falhas humanas e lentidão operacional. Partiu-se da hipótese de que a integração de arquiteturas modernas de Deep Learning permite a automação do controle de acesso com altos índices de confiabilidade, superando as abordagens manuais. Para validar essa premissa, desenvolveu-se um sistema biométrico automatizado utilizando o estado da arte em redes neurais. A metodologia integrou o detector YOLOv12n-face para o pré-processamento e recorte de faces em tempo real, seguido de uma análise comparativa quantitativa entre três arquiteturas de extração de características: FaceNet512, ArcFace e VGG-Face, avaliadas pela métrica de similaridade de cosseno. Os experimentos foram conduzidos com uma base de dados real coletada na Universidade Federal do Espírito Santo (UFES), submetendo o sistema a rigorosos cenários de autenticidade e tentativas de impostura. Os resultados confirmaram a hipótese inicial, demonstrando a superioridade do modelo FaceNet512, que atingiu 100% de acurácia global e um gap de separação positivo (+0.1632) entre usuários genuínos e impostores, eliminando totalmente falsos positivos e negativos. Com uma latência média de reconhecimento de 739 ms, conclui-se que o objetivo do trabalho foi plenamente atingido, validando a viabilidade técnica do sistema para o controle de acesso em tempo real com alta precisão e segurança.

Palavras-chave: Reconhecimento Facial; Controle de Acesso; YOLOv12; FaceNet512; Deep Learning.

ABSTRACT

Situated at the intersection of Computer Vision and Information Security, this research addresses the limitations of traditional physical credentialing methods, which are frequently susceptible to human error and operational slowness. It was hypothesized that the integration of modern Deep Learning architectures enables the automation of access control with high reliability, surpassing manual approaches. To validate this premise, an automated biometric system was developed using state-of-the-art neural networks. The methodology integrated the YOLOv12n-face detector for real-time preprocessing and face cropping, followed by a quantitative comparative analysis of three feature extraction architectures: FaceNet512, ArcFace, and VGG-Face, evaluated via the cosine similarity metric. Experiments were conducted using a real dataset collected at the Federal University of Espírito Santo (UFES), subjecting the system to rigorous scenarios of authenticity and imposture attempts. The results confirmed the initial hypothesis, demonstrating the superiority of the FaceNet512 model, which achieved 100% global accuracy and a positive separation gap (+0.1632) between genuine users and impostors, completely eliminating false positives and negatives. With an average recognition latency of 739 ms, it is concluded that the objective of this work was fully achieved, validating the technical viability of the system for real-time access control with high precision and security.

Keywords: Facial Recognition; Access Control; YOLOv12; FaceNet512; Deep Learning.

Lista de figuras

Figura 1 – Pipeline de Reconhecimento Facial 16
Figura 2 – O sistema YOLO 20
Figura 3 – Modelo YOLO 21
Figura 4 – Arquitetura DeepFace 27
Figura 5 – Função *Triplet Loss* 30
Figura 6 – Demonstração do Triplet Loss 31
Figura 7 – Pipeline de Treinamento com a ArcFace Loss 32
Figura 8 – Taxas de erro em sistemas biométricos 35
Figura 9 – Pipeline completo do sistema de reconhecimento facial proposto. 38

Lista de tabelas

Tabela 1 – Estatística descritiva dos scores e Gap de separabilidade entre cenários. 42
Tabela 2 – Ajuste de limiares para regime de segurança máxima. 43
Tabela 3 – Comparação final de desempenho entre os modelos. 44

Sumário

1 INTRODUÇÃO 11
1.1 O problema e sua importância 12
1.2 Objetivos 13
1.2.1 Objetivo Geral 13
1.2.2 Objetivos Específicos 13

2 REVISÃO DE LITERATURA 15
2.1 Sistemas de Reconhecimento Facial 15
2.1.1 Detecção e Pré-processamento da Face 15
2.1.2 Extração de Características e Comparação 15
2.2 Pipeline de Reconhecimento Facial 16
2.2.1 Aquisição da Imagem 16
2.2.1.1 Câmeras Dedicadas e Webcams 17
2.2.1.2 Câmeras de Celulares 17
2.2.1.3 Imagens de Vídeos 17
2.2.1.4 Outras Fontes e Sensores 18
2.2.2 Detecção da Face 18
2.2.2.1 Métodos Clássicos 19
2.2.2.1.1 Haar Cascade 19
2.2.2.1.2 Histogram of Oriented Gradients (HOG) + Support Vector Machine (SVM) 19
2.2.2.2 Métodos Baseados em Redes Neurais Convolucionais (CNNs) 20
2.2.2.2.1 R-CNN (*Regions with CNN features*) e suas variações (Fast R-CNN, Faster R-CNN) . . 20
2.2.2.2.2 YOLO 20
2.2.2.2.3 MediaPipe: 22
2.2.2.2.4 MTCNN (Multi-task Cascaded Convolutional Networks) 22
2.2.3 Pré-processamento 23
2.2.3.1 Etapas do Pré-Processamento 23
2.2.3.2 Métodos de pré-processamento 24
2.2.4 Extração de Características 25
2.2.4.1 Processo de Extração de Características Faciais 25
2.2.4.2 Métodos de Extração de Características 26
2.2.4.3 Eigenfaces (Análise de Componentes Principais - PCA) 26
2.2.4.3.1 Fisherfaces (Análise Discriminante Linear - LDA) 27
2.2.4.3.2 DeepFace 27
2.2.4.3.3 VGG-Face 28

2.2.4.3.4 FaceNet 29
2.2.4.3.5 ArcFace 31
2.2.4.3.6 Abordagens Alternativas com CNNs 33
2.3 Comparação e Métricas de Avaliação de Desempenho 33
2.3.1 Como é Realizada a Comparação 33
2.3.2 Métricas de Avaliação de Desempenho 34

3 METODOLOGIA 37
3.1 Coleta e Estruturação de Dados 37
3.2 Implementação do Sistema 38
3.3 Protocolo de Testes e Validação Automatizada 39
3.4 Métricas de Avaliação 40

4 RESULTADOS E DISCUSSÃO 42
4.1 Análise Estatística de Similaridade 42
4.2 Definição de Limiares e Calibração de Segurança 43
4.3 Resultados Quantitativos Comparativos 44
4.4 Limitações Encontradas 45

5 CONCLUSÃO 46

REFERÊNCIAS 47

1 Introdução

A segurança e a eficiência são pilares fundamentais para a gestão de ambientes que demandam controle de acesso rigoroso. Historicamente, a verificação de identidade tem evoluído de métodos manuais para sistemas baseados em credenciais físicas, como crachás e cartões. No entanto, esses métodos, embora funcionais, apresentam vulnerabilidades inerentes, como a possibilidade de perda, roubo, empréstimo indevido e a necessidade de intervenção humana, que pode gerar gargalos e custos operacionais (JAIN; ROSS; PRABHAKAR, 2004). A busca por soluções mais robustas e automatizadas tem impulsionado o avanço das tecnologias biométricas.

Nesse cenário, o reconhecimento facial emerge como uma das mais promissoras e amplamente adotadas tecnologias biométricas. Sua natureza não invasiva, a capacidade de verificação em tempo real e a crescente acurácia a tornam ideal para diversas aplicações, desde o desbloqueio de *smartphones* até sistemas de segurança de alta complexidade. O mercado global de reconhecimento facial, por exemplo, foi avaliado em bilhões de dólares em 2023 e projeta-se um crescimento exponencial nos próximos anos, impulsionado pela demanda por segurança e automação (Grand View Research, 2024). Essa expansão reflete a confiança na capacidade da tecnologia de transformar a forma como interagimos com sistemas de identificação.

Este trabalho demonstrou que é possível agilizar e tornar mais seguro o processo de controle de acesso a ambientes restritos por meio da automação da verificação de identidade, reduzindo a dependência de métodos manuais, muitas vezes lentos e suscetíveis a falhas. Com o uso de tecnologias de reconhecimento facial, foi possível oferecer uma solução moderna, eficiente e adaptável a diferentes contextos, como instituições de ensino, empresas e espaços públicos.

Para a concretização desse objetivo, é importante notar que existem diversas técnicas e arquiteturas disponíveis na literatura para lidar com visão computacional. Entre as mais proeminentes, destacam-se o YOLO (REDMON et al., 2016), o DeepFace (TAIGMAN et al., 2014), o FaceNet (SCHROFF; KALENICHENKO; PHILBIN, 2015) e o MediaPipe (LUGARESI et al., 2019). No entanto, a aplicação direta dessas ferramentas não é isenta de desafios. Muitas delas podem apresentar alto custo computacional, exigindo hardware robusto inviável para certas aplicações, ou demonstrar instabilidade frente a variações de iluminação, oclusões e mudanças de pose, o que compromete a confiabilidade em cenários reais de controle de acesso.

Diante desse cenário, este trabalho se propôs a explorar e avaliar diferentes métodos de aprendizado de máquina, envolvendo visão computacional e biometria, para analisar e

comparar as técnicas atuais mais adequadas para as etapas de detecção e reconhecimento facial, visando identificar a configuração mais eficiente para o sistema desenvolvido.

Dessa forma, o escopo deste projeto abrangeu desde a pesquisa e seleção das tecnologias mais adequadas, gerando também um conjunto de imagens próprio em conformidade com a Lei Geral de Proteção de Dados (LGPD) (BRASIL, 2018), até o desenvolvimento e validação de um protótipo funcional. A relevância deste estudo reside não apenas na contribuição tecnológica, mas também na abordagem ética e legal da coleta e tratamento de dados biométricos, um tema de crescente preocupação global (BUOLAMWINI; GEBRU, 2018).

1.1 O problema e sua importância

A motivação para este projeto surgiu da observação das ineficiências e vulnerabilidades inerentes aos sistemas de controle de acesso convencionais. A dependência de processos manuais ou de credenciais físicas, que podem ser facilmente comprometida, representa um risco significativo à segurança e à eficiência operacional (JAIN; ROSS; PRABHAKAR, 2004). Além disso, a necessidade de intervenção humana para validação de identidade gera custos contínuos e pode criar pontos de congestionamento, impactando negativamente a experiência do usuário. A busca por um método que combine segurança, agilidade e uma experiência de usuário fluida é, portanto, o principal fator motivador para a realização deste trabalho.

A justificativa para o desenvolvimento de um sistema de reconhecimento facial reside em sua superioridade técnica e prática para a finalidade de controle de acesso. A tecnologia de reconhecimento facial oferece uma forma de autenticação sem contato, o que se tornou especialmente relevante em um contexto pós-pandemia, além de ser intuitiva e de difícil falsificação quando implementada com algoritmos robustos (GROTHER; NGAN; HANAOKA, 2019).

Academicamente, este projeto contribui significativamente para a área de Visão Computacional e Inteligência Artificial, explorando a integração e otimização de algoritmos de ponta para uma aplicação prática. A pesquisa e o desenvolvimento de soluções baseadas em biometria facial, como a proposta, alinham-se às tendências atuais da computação, que buscam sistemas mais autônomos e inteligentes (GOODFELLOW; BENGIO; COURVILLE, 2016).

A relevância social do trabalho reside na oferta de uma solução que pode aprimorar a segurança e a conveniência em diversos ambientes, como empresas, instituições de ensino e eventos, ao mesmo tempo em que promove a discussão sobre a privacidade e o uso responsável de dados biométricos, um tema de crescente importância na sociedade contemporânea. Estudos como os de Buolamwini e Gebru (2018) e Grother, Ngan e Hanaoka

(2019) destacam a importância de abordar vieses e garantir a equidade em sistemas de reconhecimento facial, o que reforça a necessidade de um desenvolvimento responsável e ético. A aplicação de técnicas de aprendizado de máquina e visão computacional para resolver problemas do mundo real, como o controle de acesso, demonstra a aplicabilidade e o impacto dessas áreas, contribuindo para o avanço tecnológico e para a melhoria da qualidade de vida (JAIN; ROSS; PRABHAKAR, 2004).

Cientificamente, o projeto apresenta a validação de uma arquitetura de sistema que integra algoritmos de detecção e reconhecimento facial de última geração, avaliando seu desempenho em um contexto prático. A análise das métricas de acurácia, FAR e FRR em um protótipo funcional forneceu dados empíricos valiosos para a comunidade científica, contribuindo para o entendimento das capacidades e limitações desses sistemas em cenários reais. Além disso, a metodologia de construção de uma base de dados em conformidade com a LGPD pode servir de modelo para futuras pesquisas que envolvam dados sensíveis, promovendo a pesquisa responsável e a inovação ética (BRASIL, 2018). A inovação apresentada neste trabalho, ao integrar tecnologias de ponta para um controle de acesso eficiente e seguro, alinha-se às demandas de um mundo cada vez mais digitalizado, onde a agilidade e a confiabilidade são cruciais para a otimização de processos e a proteção de indivíduos e patrimônios (JAIN; ROSS; PRABHAKAR, 2004). Dessa forma, o trabalho não apenas desenvolve uma solução tecnológica inovadora, mas também estabelece um modelo de implementação ética e legalmente responsável, alinhado às exigências atuais de segurança digital e proteção de dados.

1.2 Objetivos

A seguir são apresentados os objetivos gerais e específicos relacionados a este trabalho.

1.2.1 Objetivo Geral

Desenvolver e validar um sistema automático de reconhecimento facial para eliminar a necessidade de verificação manual de documentos para o acesso a um ambiente com necessidade de identificação.

1.2.2 Objetivos Específicos

1. Analisar e selecionar a tecnologia de reconhecimento facial, considerando critérios como precisão, desempenho em tempo real e compatibilidade com os recursos disponíveis.

2. Construir uma base de dados própria e segura contendo informações autorizadas dos

usuários, composta por imagens faciais e vídeos, coletados com consentimento.

3. Desenvolver um sistema capaz de realizar a identificação facial das bases de dados já presentes na literatura e adaptá-la para o conjunto de dados próprio, formado pelas imagens e vídeos já coletados.

4. Elaborar um protótipo funcional da solução proposta, validando a capacidade do sistema de detectar e reconhecer rostos em tempo real.

2 Revisão de literatura

Este capítulo apresenta os fundamentos teóricos e as principais tecnologias que embasam o desenvolvimento do sistema de controle de acesso proposto. São abordados os conceitos centrais relacionados à biometria facial, o funcionamento dos sistemas de reconhecimento facial e as etapas que compõem seu pipeline.

2.1 Sistemas de Reconhecimento Facial

O reconhecimento facial é uma tecnologia biométrica que permite identificar ou verificar a identidade de uma pessoa com base em sua face capturada por imagem ou vídeo (JAIN; ROSS; PRABHAKAR, 2004). De acordo com Jain, Ross e Prabhakar (2004), a biometria representa uma alternativa mais segura e robusta em comparação aos métodos tradicionais, como senhas e cartões.

As etapas de detecção e extração de características são fundamentais para o funcionamento adequado desses sistemas.

2.1.1 Detecção e Pré-processamento da Face

A detecção de face consiste em localizar uma ou mais faces dentro de uma imagem. Após a detecção, a imagem é pré-processada para padronização, com etapas como alinhamento facial, ajuste de escala e correção de iluminação (GONZALEZ; WOODS, 2018). Tais transformações buscam garantir que a face esteja em condições apropriadas para análise e comparação.

Redes Neurais Convolucionais (CNNs) revolucionaram essa etapa, proporcionando maior precisão e robustez. Algoritmos modernos, como o MTCNN (ZHANG et al., 2016), conseguem detectar múltiplas faces com alto grau de precisão mesmo em condições adversas, como variações de pose e iluminação.

2.1.2 Extração de Características e Comparação

Após o pré-processamento, o sistema realiza a extração de características. Essa etapa consiste em transformar a imagem facial em um vetor numérico, denominado *embedding*, que captura as informações essenciais e exclusivas do rosto.

Esse vetor é então comparado com *embeddings* armazenados em uma base de dados. A comparação geralmente é feita por meio de uma métrica de distância (como a euclidiana). Quando a distância entre dois *embeddings* está abaixo de um determinado

limiar, considera-se que houve correspondência (BISHOP, 2006).

2.2 Pipeline de Reconhecimento Facial

Sistemas de reconhecimento facial operam de maneira estruturada por meio de um conjunto de etapas interdependentes, conhecidas como pipeline (KORTLI et al., 2020). O pipeline consiste, geralmente, nas seguintes fases (Figura 9):

1. Aquisição da imagem: Captura da imagem facial por meio de uma câmera ou vídeo.
2. Detecção de face: Localização da(s) face(s) presente(s) na imagem.
3. Pré-processamento: Normalização da imagem facial detectada (alinhamento, escala, equalização de histograma, etc.).
4. Extração de características: Conversão da imagem da face em uma representação vetorial (*embedding*) com características discriminantes.
5. Comparação ou classificação: Verificação (1:1) ou identificação (1:N) da face com base na similaridade entre *embeddings*.

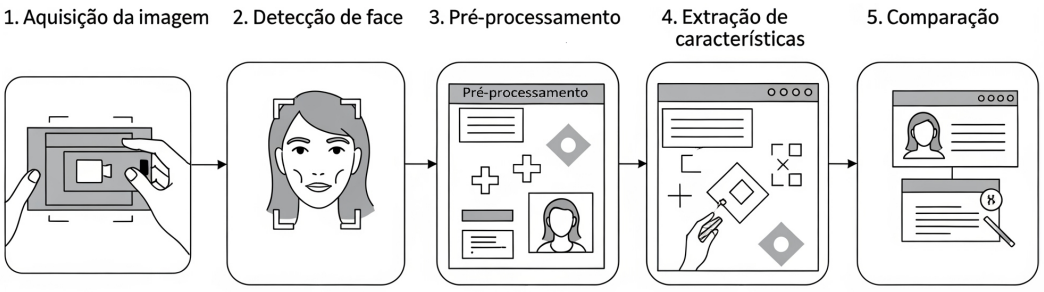


Figura 1 – Pipeline de um sistema de reconhecimento facial. Imagem final produzida pelo autor, montagem elaborada a partir de partes de imagens geradas por IA (Gemini).

Cada uma dessas etapas pode ser implementada com diferentes técnicas e algoritmos, que serão explorados nas seções seguintes.

2.2.1 Aquisição da Imagem

A aquisição da imagem é a etapa inicial e fundamental em qualquer sistema de reconhecimento facial, pois a qualidade e as características da imagem de entrada impactam diretamente o desempenho das fases subsequentes. Essa etapa envolve a captura de dados

visuais de uma face por meio de diversos dispositivos e em diferentes contextos, cada um com suas particularidades e desafios (HASSABALLAH; ABDELMGEID; AL-TURKI, 2015).

2.2.1.1 Câmeras Dedicadas e Webcams

Tradicionalmente, sistemas de reconhecimento facial utilizam câmeras dedicadas, como as empregadas em sistemas de segurança e vigilância, ou *webcams* conectadas a computadores (Swiftlane, 2020). Essas câmeras variam em resolução, taxa de quadros (*frames per second* - FPS) e capacidade de adaptação a diferentes condições de iluminação. Câmeras de segurança, por exemplo, são frequentemente otimizadas para capturar imagens em ambientes externos ou com iluminação variável, podendo incluir recursos como visão noturna (infravermelho) ou amplo alcance dinâmico (HDR) para lidar com contrastes extremos. *Webcams*, por sua vez, são comuns em aplicações de *desktop* e videoconferência, oferecendo uma solução de baixo custo, mas geralmente com menor qualidade de imagem e sensibilidade à iluminação ambiente.

2.2.1.2 Câmeras de Celulares

Com a proliferação de *smartphones*, as câmeras de celulares tornaram-se uma fonte ubíqua de imagens para reconhecimento facial. A conveniência e a alta resolução das câmeras frontais e traseiras dos celulares as tornam ideais para aplicações como desbloqueio de tela, autenticação em aplicativos bancários e controle de acesso. No entanto, as imagens capturadas por celulares podem apresentar variações significativas devido a fatores como a distância do usuário, o ângulo de captura (*selfie*), a iluminação ambiente (muitas vezes não controlada) e a qualidade do sensor. Apesar desses desafios, algoritmos modernos de reconhecimento facial são projetados para lidar com a variabilidade inerente a essas imagens, muitas vezes de baixa resolução ou com movimento, como mencionado por Lacerda (2024).

2.2.1.3 Imagens de Vídeos

A aquisição de imagens a partir de fluxos de vídeo é crucial para sistemas de reconhecimento facial em tempo real, como os utilizados em monitoramento de segurança, controle de multidões e autenticação contínua. Nesses cenários, o sistema não apenas detecta e reconhece faces em quadros individuais, mas também rastreia o movimento e a identidade ao longo do tempo. Os desafios incluem a degradação da qualidade da imagem devido ao movimento (*motion blur*), variações rápidas de iluminação, oclusões parciais e a necessidade de processamento em alta velocidade para manter a fluidez. A capacidade de lidar com vídeos com movimento e imagens de baixa resolução é essencial (OKABE; CARRO, 2015).

2.2.1.4 Outras Fontes e Sensores

Além das fontes convencionais, a literatura na área de reconhecimento facial explora uma variedade de outras tecnologias de aquisição de imagem para aprimorar a robustez e a segurança dos sistemas:

- Sensores Infravermelhos (IR): Câmeras infravermelhas capturam imagens com base no calor emitido pelos objetos, sendo menos suscetíveis a variações de iluminação visível. Isso as torna úteis para reconhecimento facial em ambientes com pouca luz ou escuridão total, e também para detecção de vivacidade (*liveness detection*), diferenciando uma face real de uma fotografia ou máscara (SINOSEEN, 2024).
- Sensores 3D: Tecnologias como câmeras de profundidade (RGB-D), luz estruturada e tempo de voo (*Time-of-Flight* - ToF) capturam informações tridimensionais do rosto. Isso permite a criação de modelos 3D da face, que são mais robustos a variações de pose, iluminação e maquiagem, além de serem mais difíceis de serem enganados por ataques de *spoofing* com fotos ou vídeos 2D. A biometria facial 3D aumenta a precisão ao considerar profundidade e contornos (SISPONTO, 2024).
- Imagens Multiespectrais e Hiperespectrais: Essas abordagens utilizam diferentes bandas do espectro eletromagnético (além do visível e infravermelho) para capturar informações adicionais sobre a pele e os tecidos faciais. Isso pode revelar características únicas que não são visíveis a olho nu, aumentando a capacidade discriminativa do sistema e a resistência a falsificações (GOWDA; MULLA; CLAY, 2006).
- Imagens Térmicas: Similar aos sensores infravermelhos, as câmeras térmicas capturam a assinatura de calor do rosto. Embora menos detalhadas que as imagens visíveis, podem ser úteis em condições extremas de iluminação e para detecção de vivacidade (CHEN; WU; ZHOU, 2006).

A escolha da tecnologia de aquisição depende dos requisitos específicos da aplicação, como ambiente de operação, nível de segurança desejado, custo e recursos computacionais disponíveis. A integração de múltiplas fontes de imagem (fusão de sensores) é uma tendência crescente para construir sistemas de reconhecimento facial mais robustos e confiáveis.

2.2.2 Detecção da Face

A detecção da face é uma etapa crucial e frequentemente a primeira no pipeline de um sistema de reconhecimento facial após a aquisição da imagem. Seu objetivo principal é localizar e delimitar a presença de uma ou mais faces humanas em uma imagem ou fluxo de vídeo, retornando as coordenadas de uma caixa delimitadora (*bounding box*) ao redor de cada rosto detectado. A precisão e a eficiência desta etapa são fundamentais,

pois erros na detecção podem comprometer todas as fases subsequentes do processo de reconhecimento. O campo da detecção facial tem evoluído significativamente com as inovações em aprendizado profundo, que impulsionaram a precisão e a robustez dos detectores modernos (WANG et al., 2018b).

Ao longo dos anos, diversas abordagens foram desenvolvidas para a detecção de faces, evoluindo de métodos baseados em regras e características manuais para modelos complexos de aprendizado profundo.

2.2.2.1 Métodos Clássicos

Antes do avanço das redes neurais profundas, a detecção e o reconhecimento facial eram dominados por métodos clássicos baseados em extração manual de características e classificadores tradicionais. Esses algoritmos geralmente combinavam descritores projetados para capturar informações geométricas ou texturais das imagens com modelos de classificação supervisionados. Entre as vantagens desses métodos estão a menor demanda computacional e a facilidade de implementação, o que os tornou amplamente utilizados em aplicações em tempo real durante muitos anos. No entanto, sua performance tende a degradar em condições não controladas, como variações de iluminação, pose, escala e oclusões parciais (ZHAO et al., 2003), o que motivou a transição para abordagens baseadas em aprendizado profundo.

2.2.2.1.1 Haar Cascade

Um dos primeiros e mais influentes algoritmos de detecção de objetos em tempo real. Utiliza classificadores em cascata baseados em características de Haar (retângulos claros e escuros) para identificar padrões faciais. É rápido e eficiente em CPUs, mas sensível a variações de iluminação, pose e oclusões, resultando em menor precisão em comparação com métodos modernos. Foi um marco, mas raramente é a escolha principal para sistemas de alta acurácia hoje (VIOLA; JONES, 2001).

2.2.2.1.2 Histogram of Oriented Gradients (HOG) + Support Vector Machine (SVM)

O HOG descreve a forma e a aparência de objetos locais pela distribuição de gradientes de intensidade ou direções de borda. Combinado com um classificador SVM, mostrou-se mais robusto que o Haar Cascade para variações de pose e iluminação, mas ainda com limitações em cenários complexos e menor precisão em faces inclinadas ou distantes (DALAL; TRIGGS, 2005).

2.2.2.2 Métodos Baseados em Redes Neurais Convolucionais (CNNs)

Os métodos baseados em CNNs revolucionaram a detecção de objetos, incluindo faces, devido à sua capacidade de aprender características complexas diretamente dos dados, superando as limitações dos métodos clássicos em termos de precisão e robustez.

2.2.2.2.1 R-CNN (*Regions with CNN features*) e suas variações (Fast R-CNN, Faster R-CNN)

- R-CNN (GIRSHICK et al., 2014): Foi um dos primeiros a aplicar CNNs para detecção de objetos. Ele gera um grande número de propostas de região (*region proposals*) e aplica uma CNN a cada uma. Embora preciso, era muito lento.
- Fast R-CNN (GIRSHICK, 2015): Melhorou a velocidade ao processar a imagem inteira com uma única CNN e, em seguida, projetar as propostas de região no mapa de características da CNN. Ainda assim, a geração das propostas de região era um gargalo.
- Faster R-CNN (REN et al., 2015): Introduziu a Region Proposal Network (RPN), que integra a geração de propostas de região na própria rede neural, tornando o processo de detecção de objetos quase em tempo real. É altamente preciso, mas ainda pode ser computacionalmente intensivo para algumas aplicações em tempo real.

2.2.2.2.2 YOLO

Proposto por Redmon et al. (2016), revolucionou a detecção de objetos ao tratar o problema como uma única tarefa de regressão. Em vez de dividir o processo em etapas separadas (geração de propostas, classificação, etc.), o YOLO processa a imagem inteira de uma só vez, prevendo simultaneamente as caixas delimitadoras e as probabilidades de classe para cada objeto. Essa característica confere ao YOLO uma elevada eficiência operacional, tornando-o tecnicamente adequado para aplicações que exigem processamento em tempo real, como sistemas de controle de acesso.

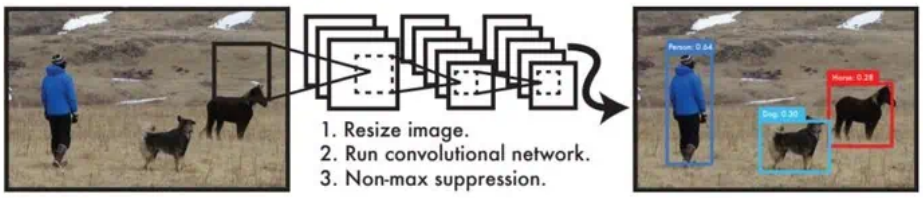


Figura 2 – Funcionamento do YOLO, onde uma imagem contendo uma pessoa e um cachorro é processada por uma rede convolucional em três etapas: Redimensionamento, detecção e supressão não máxima, resultando na identificação dos objetos com caixas delimitadoras coloridas. Imagem de Redmon et al. (2016)

Como ilustrado na Figura 2, o sistema realiza o processamento de forma simples e direta: (1) redimensiona a imagem de entrada para 448 × 448, (2) executa uma única rede convolucional na imagem e (3) limita as detecções resultantes com base na confiança do modelo.

Ele divide a imagem de entrada em uma grade (*grid*) de células e, para cada célula, prevê simultaneamente:

1. Caixas Delimitadoras (*Bounding Boxes*): As coordenadas *(x, y, w, h)* que definem a localização e o tamanho do objeto.
2. Pontuação de Confiança (*Confidence Score*): A probabilidade de que a caixa delimitadora de fato contenha um objeto e a precisão com que ela se ajusta a ele.
3. Probabilidades de Classe: A probabilidade de o objeto detectado pertencer a cada uma das classes para as quais o modelo foi treinado (neste caso, a classe “face”).

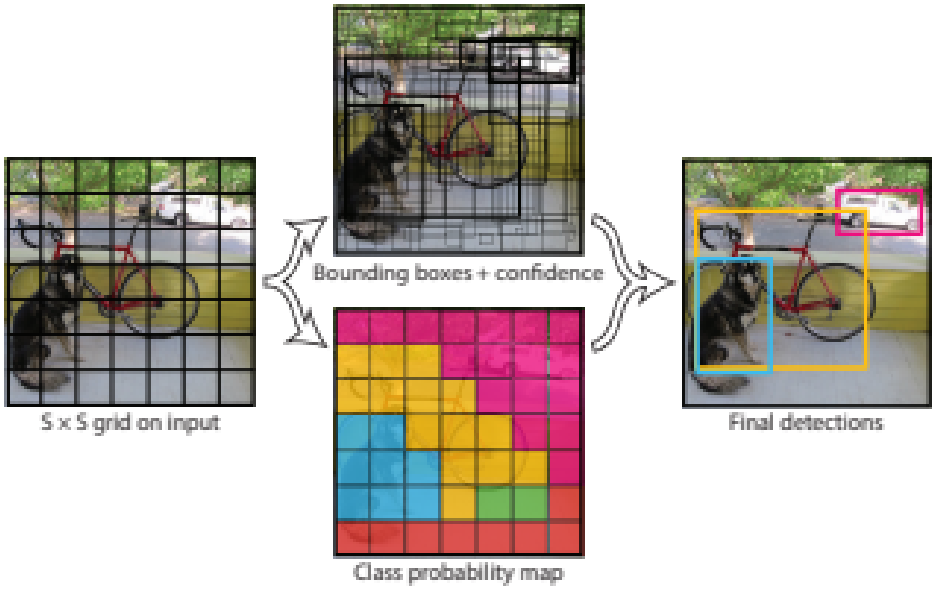


Figura 3 – Ilustra o funcionamento do modelo YOLO. A imagem de entrada é dividida em uma grade 5×5, onde cada célula prevê caixas delimitadoras e valores de confiança. Em seguida, é gerado um mapa de probabilidade por classe. Por fim, são obtidas as detecções finais, destacando os objetos identificados, como a bicicleta e o cachorro, com caixas delimitadoras coloridas. Imagem de Redmon et al. (2016)

A Figura 3 apresenta o modelo utilizado pelo sistema YOLO para detecção de objetos. O sistema divide a imagem em uma grade S × S e, para cada célula, prevê B caixas delimitadoras, suas respectivas confiabilidades e C probabilidades de classe.

Essa arquitetura de passagem única (*single-pass*) permite que o YOLO processe imagens em tempo real, o que viabiliza seu uso em sistemas de monitoramento de vídeo e

controle de acesso. O modelo analisa a imagem inteira de uma só vez, o que lhe confere um contexto global sobre os objetos e suas localizações, reduzindo significativamente os erros de detecção de fundo (falsos positivos). Versões subsequentes do YOLO, como YOLOv12 (TIAN; YE; DOERMANN, 2025), aprimoraram ainda mais a precisão e a velocidade, consolidando a família YOLO como uma das principais escolhas para detecção em tempo real.

2.2.2.2.3 MediaPipe:

Desenvolvido pelo Google, o MediaPipe é uma estrutura de código aberto para a construção de pipelines de aprendizado de máquina para processamento de dados multimodais, incluindo vídeo, áudio e dados de sensores. Para detecção facial, o MediaPipe oferece soluções eficientes e leves, otimizadas para dispositivos móveis e aplicações com restrições de hardware (LUGARESI et al., 2019).

Diferente de arquiteturas de detecção de propósito geral como o YOLO, o MediaPipe emprega modelos convolucionais otimizados para dispositivos de borda (*edge computing*), destacando-se pela eficiência no rastreamento facial e detecção de *landmarks* (pontos faciais chave) com reduzido custo computacional. Ele emprega uma abordagem de pipeline que pode incluir um detector de face leve seguido por um modelo de *landmark* facial, permitindo uma detecção e rastreamento contínuos e eficientes. Sua leveza e otimização o tornam uma excelente opção para aplicações embarcadas e em tempo real.

2.2.2.2.4 MTCNN (Multi-task Cascaded Convolutional Networks)

O MTCNN é um método popular e eficaz para detecção e alinhamento facial. Ele consiste em uma cascata de três redes neurais convolucionais que trabalham em conjunto:

1. P-Net (*Proposal Network*): Gera propostas de caixa delimitadora e seus escores de confiança.
2. R-Net (*Refine Network*): Filtra as propostas de caixa delimitadora e refina suas coordenadas.
3. O-Net (*Output Network*): Produz as caixas delimitadoras finais, os escores de confiança e os cinco *landmarks* faciais (olhos, nariz, cantos da boca).

O MTCNN é conhecido por sua alta precisão na detecção de múltiplas faces, mesmo em condições adversas como variações de pose, iluminação e oclusões parciais. Sua arquitetura em cascata permite um refinamento progressivo das detecções, tornando-o robusto para diversas aplicações (ZHANG et al., 2016).

2.2.3 Pré-processamento

Após a detecção da face em uma imagem ou vídeo, a etapa de pré-processamento torna-se essencial para padronizar e otimizar a imagem facial antes que as características sejam extraídas. O objetivo principal do pré-processamento é reduzir a variabilidade indesejada nas imagens causadas por fatores externos (como iluminação, pose, expressão facial e oclusões) e intrínsecos (como ruído do sensor) (PHILLIPS et al., 2005), garantindo que o sistema de reconhecimento facial opere com dados mais consistentes e de maior qualidade (ZHAO et al., 2003). Essa padronização é crucial para melhorar a acurácia e a robustez dos algoritmos subsequentes (TAIGMAN et al., 2014).

2.2.3.1 Etapas do Pré-Processamento

O pré-processamento consiste em uma série de transformações aplicadas à região da face detectada. Essas transformações visam normalizar a imagem, removendo ou atenuando os efeitos que podem dificultar a comparação entre faces. A ideia é que, após o pré-processamento, duas imagens da mesma pessoa, capturadas em condições diferentes, sejam o mais semelhantes possível, enquanto imagens de pessoas diferentes permaneçam distintas. Os métodos de pré-processamento são aplicados para remover artefatos e desvios, especialmente em imagens de ambientes não controlados (AMARAL et al., 2009).

As principais operações realizadas no pré-processamento incluem:

1. Alinhamento Facial: Corrige a rotação e a inclinação da face para uma pose frontal padrão. Isso é feito identificando pontos fiduciais (*landmarks*) no rosto, como os cantos dos olhos e da boca, e aplicando transformações geométricas (rotação, translação e escala) para alinhar esses pontos a uma configuração de referência. O alinhamento garante que as características faciais estejam na mesma posição relativa em todas as imagens, independentemente da pose original da pessoa (GONZALEZ; WOODS, 2018).

2. Ajuste de Escala (Redimensionamento): Padroniza o tamanho da face detectada para uma dimensão fixa (por exemplo, 160x160 *pixels* para o FaceNet (SCHROFF; KALENICHENKO; PHILBIN, 2015)). Isso é importante porque algoritmos de extração de características podem ser sensíveis a variações de escala. O redimensionamento pode ser feito usando interpolação bilinear ou bicúbica (GONZALEZ; WOODS, 2008).

3. Correção de Iluminação (Normalização de Iluminação): A iluminação é um dos fatores mais desafiadores no reconhecimento facial. Variações na intensidade e direção da luz podem alterar drasticamente a aparência de uma face. Métodos de correção de iluminação visam equalizar a distribuição de luz na imagem, tornando-a menos suscetível a sombras e brilhos excessivos. Isso pode incluir técnicas como equalização

de histograma, normalização de brilho e contraste, ou o uso de filtros homomórficos (MENDONçA, 2008); (ANDREZZA, 2015).

4. Remoção de Ruído: A remoção de ruído é uma etapa crucial em sistemas de visão computacional (GONZALEZ; WOODS, 2008). Filtros são aplicados para reduzir o ruído presente na imagem, que pode ser causado por condições de baixa luz, compressão ou imperfeições do sensor. A remoção de ruído ajuda a preservar as características importantes da face e a melhorar a qualidade geral da imagem, o que é um desafio conhecido no reconhecimento facial (ZHAO et al., 2003; CAO et al., 2018).

2.2.3.2 Métodos de pré-processamento

Diversos métodos e técnicas são empregados na etapa de pré-processamento, cada um com sua finalidade específica:

- Equalização de Histograma: Uma técnica comum para melhorar o contraste de uma imagem, distribuindo uniformemente os valores de intensidade dos *pixels*. Isso ajuda a realçar detalhes em regiões muito claras ou muito escuras, tornando a imagem mais adequada para análise (GOMES; ARAUJO, 2021); (JESUS et al., 2015).
- Filtros Espaciais (Suavização e Realce): Tais filtros são utilizados para modificar o espectro espacial da imagem, sendo ferramentas essenciais no pré-processamento para suavização e remoção de ruído, conforme detalhado na literatura clássica de processamento de imagens de Gonzalez e Woods (2008).
  - Filtro Gaussiano: Um filtro de suavização que reduz o ruído e os detalhes finos da imagem, aplicando uma média ponderada dos *pixels* vizinhos. É útil para remover ruído aleatório e preparar a imagem para outras operações.
  - Filtro da Mediana: Outro filtro de suavização, eficaz na remoção de ruído impulsivo (sal e pimenta) sem borrar as bordas da imagem de forma significativa.
  - Filtros Lineares: Incluem filtros de média e outros filtros passa-baixa que suavizam a imagem, reduzindo o ruído e detalhes de alta frequência.
- Conversão para Escala de Cinza: A conversão para escala de cinza, uma técnica fundamental em processamento de imagens (GONZALEZ; WOODS, 2008), é uma etapa comum no pré-processamento. Embora muitos sistemas modernos de reconhecimento facial possam trabalhar com imagens coloridas (RGB), a conversão para escala de cinza é uma etapa comum para reduzir a dimensionalidade dos dados e o custo computacional, sem perda significativa de informações para a maioria das tarefas de reconhecimento facial (VIOLA; JONES, 2001; ZHAO et al., 2003). A cor geralmente não é uma característica discriminativa primária para a identidade facial.

- Normalização Geométrica: Além do alinhamento, a normalização geométrica pode envolver a correção de distorções de perspectiva e a padronização da pose da cabeça (ZHAO et al., 2003; PHILLIPS et al., 2005). Isso é particularmente importante para garantir que a face esteja em uma orientação consistente, independentemente do ângulo de captura. Essa padronização é uma etapa crucial para aumentar a robustez de sistemas de reconhecimento facial (TAIGMAN et al., 2014).

- Normalização de *Pixels*: A normalização de *pixels*, uma técnica fundamental em processamento de imagens (GONZALEZ; WOODS, 2008), ajusta os valores de intensidade dos *pixels* para um intervalo padrão (por exemplo, entre 0 e 1 ou -1 e 1). Esta etapa é crucial para o treinamento de redes neurais, pois ajuda a estabilizar o processo de otimização e a garantir que os algoritmos não sejam indevidamente influenciados por grandes variações nos valores de pixel (LECUN et al., 1998). A normalização é uma etapa padrão em arquiteturas de reconhecimento facial baseadas em *deep learning*, como a utilizada no FaceNet (SCHROFF; KALENICHENKO; PHILBIN, 2015).

O pré-processamento é uma fase crítica que prepara os dados para a extração de características, garantindo que as variações na aparência da face sejam atribuídas à identidade da pessoa e não a fatores externos (ZHAO et al., 2003). A escolha e a sequência dos métodos de pré-processamento dependem da qualidade das imagens de entrada e dos requisitos específicos do algoritmo de reconhecimento facial a ser utilizado, sendo uma etapa fundamental tanto em abordagens clássicas quanto em sistemas baseados em aprendizado profundo (ZAFEIRIOU et al., 2015).

2.2.4 Extração de Características

Após a detecção e o pré-processamento da face, a próxima etapa crucial em um sistema de reconhecimento facial é a extração de características (ZHAO et al., 2003). Esta fase é responsável por transformar a imagem facial, que é uma representação visual, em um formato numérico compacto e discriminativo, conhecido como vetor de características ou *embedding* (SCHROFF; KALENICHENKO; PHILBIN, 2015). O objetivo é capturar as informações essenciais e únicas do rosto de uma pessoa, que permitem distingui-la de outras, enquanto minimiza a variabilidade intra-classe (variações na mesma pessoa devido a pose, iluminação, expressão, etc.) e maximiza a variabilidade inter-classe (diferenças entre pessoas distintas) (DUDA; HART; STORK, 2001).

2.2.4.1 Processo de Extração de Características Faciais

A extração de características pode ser entendida como o processo de identificar e quantificar os atributos mais relevantes de uma face. Em vez de usar a imagem inteira, que

contém uma grande quantidade de dados redundantes e sensíveis a ruídos, esta etapa foca em extrair um conjunto menor de valores que representam a identidade facial de forma eficiente. Esse vetor numérico serve como uma "impressão digital" matemática do rosto, que será posteriormente utilizada para comparação.

Historicamente, a extração de características era feita manualmente ou com algoritmos baseados em regras que identificavam pontos fiduciais (como a distância entre os olhos, a largura do nariz, a forma da boca) ou padrões de textura (ZHAO et al., 2003). Com o avanço do aprendizado de máquina e, mais recentemente, do aprendizado profundo, os métodos evoluíram para abordagens que aprendem as características mais discriminativas diretamente dos dados.

O processo geralmente envolve:

1. Identificação de Atributos Relevantes: O algoritmo analisa a imagem pré-processada para encontrar padrões, texturas, formas e relações espaciais que são distintivas para cada indivíduo.

2. Mapeamento para um Espaço de Baixa Dimensionalidade: As características extraídas são mapeadas para um vetor de números (o *embedding*), que possui uma dimensionalidade muito menor do que a imagem original, mas retém a informação crucial para a identificação.

3. Otimização para Discriminação: Os métodos modernos, especialmente os baseados em redes neurais, são treinados para gerar *embeddings* onde faces da mesma pessoa são agrupadas (distância pequena) e faces de pessoas diferentes são separadas (distância grande) (SCHROFF; KALENICHENKO; PHILBIN, 2015).

2.2.4.2 Métodos de Extração de Características

Diversas técnicas foram desenvolvidas para a extração de características, variando em complexidade e desempenho.

2.2.4.3 Eigenfaces (Análise de Componentes Principais - PCA)

Um dos primeiros métodos de reconhecimento facial, proposto por Turk e Pentland (1991). Ele trata as imagens faciais como pontos em um espaço de alta dimensionalidade e usa PCA para encontrar um conjunto de vetores ortogonais (eigenfaces) que representam as principais variações entre as faces. Cada face é então projetada nesse subespaço, resultando em um vetor de características. Embora historicamente importante, é sensível a variações de iluminação e pose e menos robusto que os métodos modernos.

2.2.4.3.1 Fisherfaces (Análise Discriminante Linear - LDA)

Desenvolvido para superar algumas limitações do Eigenfaces, o Fisherfaces (BELHUMEUR; HESPANHA; KRIEGMAN, 1997) utiliza a Análise Discriminante Linear (LDA) para encontrar uma projeção que maximiza a separabilidade entre as classes (pessoas diferentes) e minimiza a variabilidade dentro da mesma classe. Isso o torna mais robusto a variações de iluminação e expressão do que o PCA.

2.2.4.3.2 DeepFace

Um dos primeiros sistemas de reconhecimento facial baseados em CNNs a atingir desempenho próximo ao humano. O DeepFace, em sua evolução desde a proposta original do Facebook AI Research (FAIR) (TAIGMAN et al., 2014) até implementações modernas como a biblioteca Deepface (SERENGIL; OZPINAR, 2020), consolidou-se como uma abordagem eficaz, combinando técnicas de pré-processamento avançado e aprendizado métrico.

Como ilustrado na Figura 4, o sistema DeepFace consiste em quatro módulos principais: alinhamento 2D, alinhamento 3D, frontalização e rede neural. Uma imagem de um rosto passa por esses módulos em sequência, resultando em um vetor de características de 4.096 dimensões que representa o rosto.

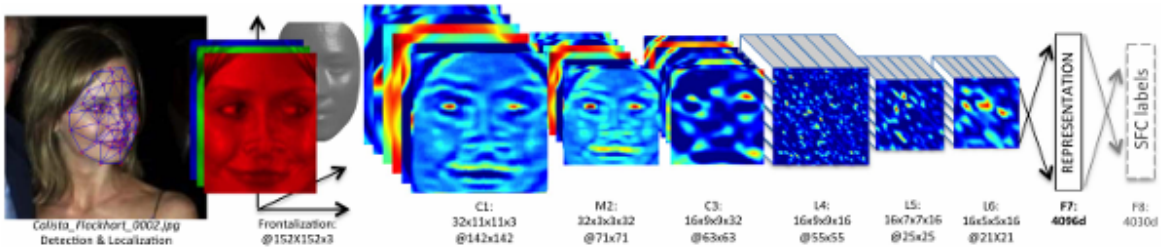


Figura 4 – O diagrama representa o fluxo de processamento do modelo DeepFace, que começa com o alinhamento 3D do rosto para normalizar a imagem. Em seguida, a imagem passa por camadas de convolução e *pooling* que extraem características faciais, culminando em uma representação vetorial (*embedding*). Imagem de Taigman et al. (2014)

O alinhamento facial começa com um alinhamento 2D, e então continua com o alinhamento 3D e a frontalização. Isso significa que o processo do DeepFace é em duas etapas: primeiro, ele corrige os ângulos de uma imagem para que o rosto na foto esteja olhando para frente, utilizando um modelo 3D de um rosto.

A frontalização tem o objetivo de reduzir erros causados por projeções de perspectiva não modeladas, buscando distorcer as imagens 2D com as menores distorções possíveis. Além disso, a câmera P é capaz de substituir partes da imagem e misturá-las com suas contrapartes simétricas (TAIGMAN et al., 2014). A câmera P refere-se a um modelo

matemático que representa a projeção da imagem 3D do rosto para a imagem 2D capturada, permitindo corrigir e misturar partes da face para minimizar distorções causadas pela perspectiva.

A rede neural é composta por uma sequência de camadas organizadas da seguinte forma: camada convolucional, *max pooling*, outra camada convolucional, três camadas localmente conectadas e, por fim, uma camada totalmente conectada. A entrada é uma imagem RGB do rosto, redimensionada para uma resolução específica, e a saída é um vetor real de dimensão 4.096, que representa as características extraídas da imagem facial. No artigo original, uma camada totalmente conectada adicional é adicionada no final para classificar a imagem do rosto em uma das 4.030 pessoas possíveis que a rede havia visto durante o treinamento (TAIGMAN et al., 2014).

O DeepFace foi treinado em um dos maiores conjuntos de dados faciais da época, contendo quatro milhões de imagens rotuladas por identidade, pertencentes a mais de 4.000 indivíduos. O objetivo do treinamento é aprender a mapear diretamente as imagens para um espaço vetorial euclidiano compacto, onde a distância entre os vetores (ou *embeddings*) de duas imagens da mesma pessoa é pequena, enquanto a distância entre vetores de pessoas diferentes é grande. Essa metodologia de aprendizado de representações discriminativas em um espaço de *embedding* é fundamental para o DeepFace (TAIGMAN et al., 2014).

Implementações modernas, como a biblioteca Deepface (SERENGIL; OZPINAR, 2020), expandiram significativamente as capacidades do *framework* original. Essas versões atualizadas incorporam *backends* alternativos como ArcFace (DENG et al., 2019) e FaceNet (SCHROFF; KALENICHENKO; PHILBIN, 2015), que produzem *embeddings* mais compactos (128D a 512D) sem comprometer a acurácia.

Essa abordagem de ponta a ponta, combinada com as contínuas melhorias em arquiteturas de redes neurais e técnicas de treinamento, mantém o DeepFace como uma das soluções mais eficientes para reconhecimento facial em aplicações que demandam tanto precisão quanto eficiência computacional.

2.2.4.3.3 VGG-Face

O modelo VGG-Face é um marco no reconhecimento facial baseado em Redes Neurais Convolucionais (CNNs) profundas, desenvolvido pelo Visual Geometry Group (VGG) da Universidade de Oxford (PARKHI; VEDALDI; ZISSERMAN, 2015). O trabalho, apresentado por Parkhi, Vedaldi e Zisserman (2015) em 2015, demonstrou a eficácia de arquiteturas "muito profundas" (inspiradas na VGG-16) treinadas em um conjunto de dados massivo para a tarefa de verificação e identificação facial.

A principal contribuição do VGG-Face reside em dois pilares: a criação de um dataset em larga escala e a aplicação de uma arquitetura de CNN profunda. O VGG

Face Dataset, compilado pelos autores, consistiu em aproximadamente 2,6 milhões de imagens de 2.622 identidades, superando em ordem de magnitude os datasets publicamente disponíveis na época. A utilização de um volume tão grande de dados foi crucial para treinar uma rede com alta capacidade de representação.

A arquitetura da rede VGG-Face é uma adaptação da arquitetura VGG, caracterizada por sua profundidade (16 camadas com pesos) e pelo uso exclusivo de filtros convolucionais pequenos de 3×3. Essa profundidade permite que a rede aprenda hierarquias complexas de características faciais, desde bordas e texturas de baixo nível até representações semânticas de alto nível.

Diferentemente de abordagens que utilizam a função de perda tripla (*Triplet Loss*) desde o início, o VGG-Face foi inicialmente treinado como um classificador para distinguir as 2.622 identidades presentes no seu dataset. Após o treinamento, a camada de classificação (*softmax*) é removida, e o vetor de características extraído da camada anterior (penúltima camada totalmente conectada) é utilizado como o embedding facial. Este vetor de 1024 dimensões atua como uma representação compacta e discriminativa da face, onde a similaridade entre duas faces é medida pela distância Euclidiana ou similaridade de cosseno entre seus embeddings (PARKHI; VEDALDI; ZISSERMAN, 2015).

2.2.4.3.4 FaceNet

O FaceNet, um sistema desenvolvido pelo Google que alcançou desempenho estado da arte em *benchmarks* de reconhecimento facial (SCHROFF; KALENICHENKO; PHILBIN, 2015). Schroff, Kalenichenko e Philbin (2015) demonstraram que o FaceNet é capaz de gerar *embeddings* de alta qualidade que permitem o reconhecimento facial com elevada precisão.

A principal inovação do FaceNet é que ele aprende a mapear diretamente as imagens de faces para um espaço vetorial euclidiano compacto de 128 dimensões (conhecido como *embedding space*). Nesse espaço, a distância entre os vetores (ou *embeddings*) de duas imagens da mesma pessoa é pequena, enquanto a distância entre os vetores de imagens de pessoas diferentes é grande. Essa propriedade permite que tarefas de reconhecimento facial, como verificação e identificação, sejam resolvidas de forma simples usando técnicas de agrupamento (*clustering*) ou comparação de distâncias (GOODFELLOW; BENGIO; COURVILLE, 2016). Goodfellow, Bengio e Courville (2016) ressaltam a importância do aprendizado de representações em redes neurais profundas para tarefas complexas como o reconhecimento de padrões.

Para alcançar esse mapeamento, o FaceNet utiliza uma arquitetura de Rede Neural Convolucional (CNN) profunda. A estrutura geral desse modelo é composta por uma camada de entrada em lote, seguida de uma CNN profunda e uma normalização $L2$,

resultando na incorporação de faces. Durante o treinamento, aplica-se a função de perda tripla (*Triplet Loss*) como pode ser vista na figura 5

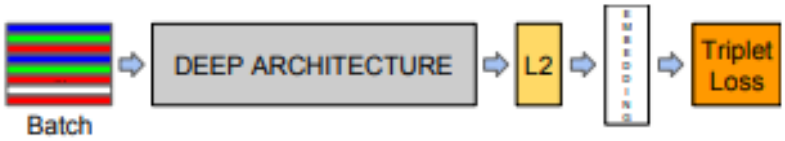


Figura 5 – O diagrama representa um modelo que utiliza a função de custo *Triplet Loss*. Um lote de imagens é processado por uma arquitetura de rede neural para gerar vetores de características. A normalização L2 é aplicada, e então a *Triplet Loss* é calculada para treinar o modelo, minimizando a distância entre *embeddings* de rostos semelhantes e maximizando a distância entre os diferentes. Imagem de Schroff, Kalenichenko e Philbin (2015)

Em vez de classificar cada imagem individualmente, o método de treinamento considera três imagens por vez:

1. Âncora (*Anchor*): Uma imagem de referência de um indivíduo.
2. Positiva (*Positive*): Outra imagem do mesmo indivíduo da âncora.
3. Negativa (*Negative*): Uma imagem de um indivíduo diferente.

O objetivo da função de perda tripla é ajustar os pesos da rede neural para que a distância entre a âncora e a positiva (d(A, P)) seja sempre menor do que a distância entre a âncora e a negativa (d(A, N)), com uma margem de segurança (α). A condição a ser satisfeita é:

$$d(A, P) + \alpha < d(A, N)$$

A Figura 6 representa o funcionamento do *Triplet Loss*, técnica usada no treinamento do modelo FaceNet que minimiza a distância entre uma âncora e um positivo (mesma identidade) e maximiza a distância entre a âncora e um negativo (identidade diferente).

Figura 6 – A imagem demonstra o objetivo da *Triplet Loss*: durante o aprendizado, o modelo ajusta os *embeddings* para aproximar o vetor "*Positive*" (mesma classe) do vetor "*Anchor*" e afastar o vetor "*Negative*" (classe diferente) do "*Anchor*". Schroff, Kalenichenko e Philbin (2015)

Ao treinar a rede com milhões desses "*triplets*", o FaceNet aprende a gerar *embeddings* altamente discriminativos e generalizáveis, que são robustos a variações de pose, iluminação e expressão facial. Uma vez que o modelo está treinado, o processo de reconhecimento consiste em aplicar a imagem da face detectada na rede para gerar seu vetor de 128 dimensões, que é então comparado com os vetores armazenados na base de dados.

2.2.4.3.5 ArcFace

Uma abordagem mais recente que aprimora o treinamento de CNNs para reconhecimento facial, focando em maximizar a separação angular entre as classes no espaço de *embedding*. O ArcFace introduz uma função de perda baseada em margem angular aditiva (*Additive Angular Margin Loss*), que impõe uma separação mais rigorosa entre as caracte-

rísticas de diferentes identidades. Isso resulta em *embeddings* ainda mais discriminativos e um desempenho superior em bases de dados grandes e desafiadoras (DENG et al., 2019).

A principal inovação do ArcFace reside na sua função de perda, que tem uma clara interpretação geométrica devido à sua correspondência exata com a distância geodésica em uma hiperesfera. Ao contrário da função de perda *softmax* tradicional, que não otimiza explicitamente a similaridade intra-classe e a diversidade inter-classe, o ArcFace adiciona uma penalidade de margem angular fixa diretamente ao ângulo entre a característica de entrada e o vetor de peso da classe correta. Isso força o modelo a aprender características mais robustas, pois a condição para uma classificação correta se torna mais rigorosa: a distância angular entre as faces da mesma pessoa deve ser ainda menor do que o mínimo exigido pelo *softmax*, o que se traduz em fronteiras de decisão mais nítidas entre as identidades. O ArcFace alcançou resultados de ponta em diversos *benchmarks*, sendo, ao mesmo tempo, de fácil implementação e com um custo computacional insignificante durante o treinamento (DENG et al., 2019).

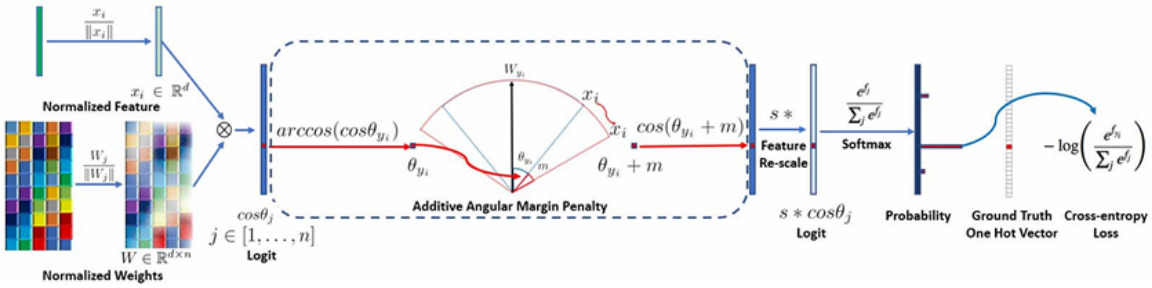


Figura 7 – Pipeline de treinamento de uma DCNN para reconhecimento facial supervisionado pela perda ArcFace. A imagem ilustra o fluxo de dados desde a normalização das características até o cálculo da perda de entropia cruzada, com o passo central do penalizador de margem angular aditiva (*Additive Angular Margin Penalty*). Imagem de Deng et al. (2019)

O diagrama acima (Figura 7) ilustra o processo detalhado de como a ArcFace Loss atua no treinamento de uma rede neural. Inicialmente, as características da DCNN (*Deep Convolutional Neural Network*) e os pesos da última camada totalmente conectada são normalizados por L_2. Isso faz com que a previsão dependa apenas do ângulo entre o vetor de características e o vetor de peso, projetando-os em uma hiperesfera de raio fixo. A seguir, a função arc-coseno é utilizada para calcular o ângulo (θ_{y_i}) entre a característica de entrada e o peso da classe correspondente (*ground truth*).

É neste ponto que a penalidade de margem angular aditiva (‘m‘) é aplicada diretamente ao ângulo, ou seja, $\theta_{y_i} + m$. Isso aumenta a separação angular entre as classes, forçando a rede a aprender características mais discriminativas. Depois de aplicar essa penalidade, o cosseno do novo ângulo é calculado e todos os *logits* são reescalados por um fator ‘s‘. Finalmente, esses valores processados são submetidos à função *softmax* para

calcular a probabilidade e, em seguida, a perda de entropia cruzada, completando o ciclo de treinamento. Este processo garante que o ArcFace otimiza diretamente a distância geodésica na hiperesfera, o que resulta em um desempenho superior e em uma interpretação geométrica clara (DENG et al., 2019).

2.2.4.3.6 Abordagens Alternativas com CNNs

Além dos citados, existem inúmeras outras arquiteturas e funções de perda para treinar CNNs para extração de características, como VGGFace (CAO et al., 2018), SphereFace (LIU et al., 2018), CosFace (WANG et al., 2018a), entre outros. Todos buscam otimizar a representação facial para que seja robusta a variações e altamente discriminativa.

2.3 Comparação e Métricas de Avaliação de Desempenho

Após a extração das características faciais e sua representação como *embeddings* (vetores numéricos) (SCHROFF; KALENICHENKO; PHILBIN, 2015), a etapa final do pipeline de reconhecimento facial é a comparação. Esta fase determina se duas faces pertencem à mesma pessoa ou, em um cenário de identificação, qual identidade em um banco de dados corresponde à face de entrada (ZHAO et al., 2003). A eficácia dessa comparação é quantificada por meio de métricas de avaliação de desempenho, que são cruciais para validar a acurácia e a confiabilidade do sistema (PHILLIPS et al., 2005).

2.3.1 Como é Realizada a Comparação

A comparação no reconhecimento facial é fundamentalmente baseada na similaridade ou distância entre os *embeddings* faciais. Dada uma face de entrada, seu *embedding* é gerado e, em seguida, comparado com um ou mais *embeddings* armazenados em uma base de dados de referência (SCHROFF; KALENICHENKO; PHILBIN, 2015). Existem dois cenários principais de comparação (PHILLIPS et al., 2005):

1. Verificação (1:1): Neste cenário, o sistema compara a face de entrada com uma única face de referência (declarada pelo usuário, por exemplo, em um sistema de autenticação). O objetivo é verificar se a pessoa é quem ela afirma ser. A comparação resulta em um escore de similaridade ou distância. Se esse escore estiver acima de um determinado limiar (para similaridade) ou abaixo de um limiar (para distância), a identidade é confirmada.

2. Identificação (1:N): Aqui, a face de entrada é comparada com todas as faces em um banco de dados de referência. O objetivo é identificar a pessoa desconhecida. O sistema calcula o escore de similaridade/distância entre o *embedding* de entrada e

cada *embedding* no banco de dados, e a identidade com o escore mais alto (ou menor distância) é considerada a correspondência. Um limiar também pode ser aplicado para determinar se a correspondência é suficientemente forte para ser considerada uma identificação válida, evitando falsos positivos.

As métricas de distância mais comuns utilizadas para comparar *embeddings* incluem (DUDA; HART; STORK, 2001; SCHROFF; KALENICHENKO; PHILBIN, 2015):

- Distância Euclidiana: Calcula a distância em linha reta entre dois pontos em um espaço multidimensional. Quanto menor a distância euclidiana entre dois *embeddings*, maior a similaridade entre as faces.
- Similaridade de Cosseno: Mede o cosseno do ângulo entre dois vetores. Um valor de 1 indica que os vetores são idênticos (apontam na mesma direção), 0 indica ortogonalidade (sem similaridade), e -1 indica que são opostos. É amplamente utilizada porque é menos sensível à magnitude dos vetores e mais focada na direção, o que pode ser vantajoso para *embeddings* de redes neurais.

2.3.2 Métricas de Avaliação de Desempenho

A validação da eficácia de um sistema de reconhecimento facial exige o uso de métricas quantitativas que avaliem sua precisão e confiabilidade (JAIN; ROSS; PRABHAKAR, 2011; ZHAO et al., 2003). A performance do protótipo desenvolvido foi analisada com base em um conjunto de métricas padrão da indústria, calculadas a partir dos resultados dos testes de identificação (PHILLIPS et al., 2005). As principais métricas a serem utilizadas são:

- Acurácia (*Accuracy*): Representa a proporção de todas as identificações corretas (tanto positivos verdadeiros quanto negativos verdadeiros) em relação ao número total de tentativas. É uma medida geral da performance do sistema.

  Fórmula: *Acurácia = (Verdadeiros Positivos + Verdadeiros Negativos) / Total de Amostras*

- Taxa de Falsa Aceitação (*False Acceptance Rate* - FAR): Também conhecida como erro do tipo I, mede a frequência com que o sistema identifica incorretamente um indivíduo não autorizado como um usuário válido. Em um sistema de controle de acesso, uma FAR alta representa uma falha grave de segurança.

  Fórmula: *FAR = Falsos Positivos / (Falsos Positivos + Verdadeiros Negativos)*

- Taxa de Falsa Rejeição (*False Rejection Rate* - FRR): Também chamada de erro do

tipo II, mede a frequência com que o sistema falha em reconhecer um usuário autorizado, negando-lhe o acesso indevidamente. Uma FRR alta prejudica a usabilidade e a conveniência do sistema.

Fórmula: *FRR = Falsas Rejeições / (Falsas Rejeições + Verdadeiros Positivos)*

- Taxa de Reconhecimento Correto (*True Positive Rate* - TPR ou *Recall*): Mede a proporção de usuários autorizados que são corretamente identificados pelo sistema. É o inverso da FRR (TPR = 1 - FRR).

O comportamento do sistema é frequentemente ajustado por um limiar de decisão (*threshold*). Um limiar mais baixo torna o sistema mais permissivo, diminuindo a FRR, mas aumentando a FAR. Inversamente, um limiar mais alto torna o sistema mais rigoroso, reduzindo a FAR, mas elevando a FRR (JAIN; ROSS; PRABHAKAR, 2004). O processo de calibração visa atingir o equilíbrio entre essas métricas, conhecido como Taxa de Erro Igual (*Equal Error Rate* - EER), ponto em que FAR e FRR se igualam. A análise dessas métricas foi fundamental para validar o protótipo e ajustar seus parâmetros para para uma operação consistente em um cenário de aplicação real.

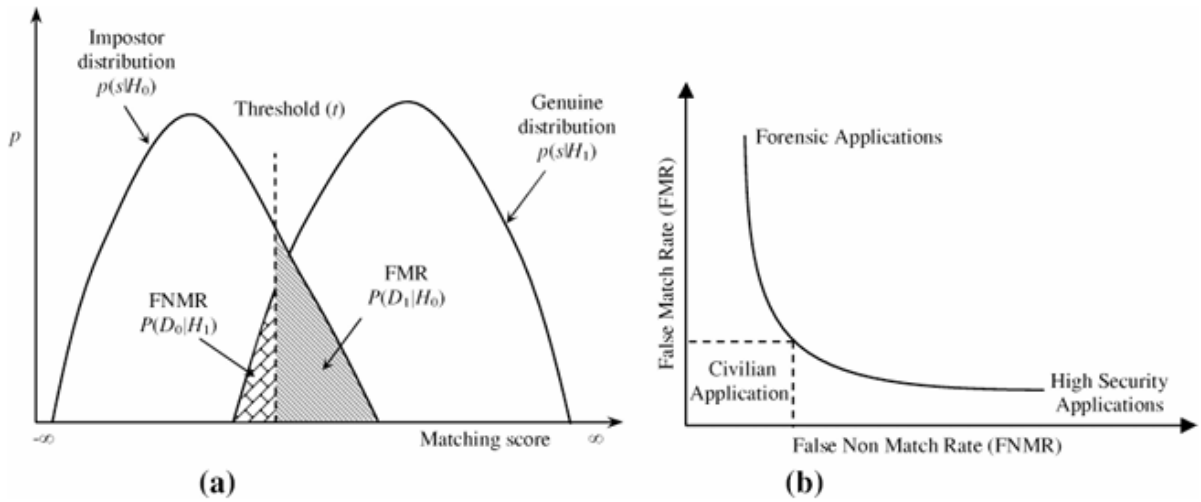


Figura 8 – No gráfico (a), são mostradas as distribuições de impostores e genuínos em função da pontuação de correspondência, destacando o limiar (*threshold*) que separa os dois grupos e indicando as taxas de FMR (*False Match Rate*) e FNMR (*False Non-Match Rate*). No gráfico (b), observa-se a relação inversa entre FMR e FNMR, indicando diferentes aplicações conforme o nível de segurança exigido, indo de aplicações civis até aplicações de alta segurança. Imagem de Jain, Ross e Prabhakar (2004)

A Figura 8 ilustra o comportamento dessas métricas em relação ao limiar adotado. Na parte (a), visualiza-se a sobreposição entre as distribuições de escores de similaridade para pares genuínos (mesma identidade) e impostores (identidades diferentes), destacando como FMR (*False Match Rate*, equivalente à FAR) e FNMR (*False Non-Match Rate*,

equivalente à FRR) são afetadas por um dado limiar t. A FMR corresponde à proporção de pares impostores com escores acima ou iguais a t, enquanto a FNMR representa a proporção de pares genuínos com escores abaixo de t.

Na parte (b), observa-se que diferentes valores de limiar resultam em diferentes combinações entre FMR e FNMR, formando a curva característica do receptor (Curva ROC - *Receiver Operating Characteristic*). Essa curva permite visualizar os possíveis pontos de operação de um sistema biométrico, que variam conforme os requisitos de segurança e conveniência. A compreensão inadequada dessas taxas de erro é, frequentemente, uma das principais causas de equívocos na avaliação de desempenho por parte de usuários e fornecedores.

3 Metodologia

A presente pesquisa foi classificada como aplicada, de natureza experimental e com abordagem quantitativa. O objetivo principal consistiu no desenvolvimento de um sistema de controle de acesso baseado em reconhecimento facial e na avaliação de seu desempenho por meio de métricas objetivas. Nesse processo, foram testados e comparados diversos métodos e algoritmos de visão computacional, visando identificar a configuração mais eficiente para o problema proposto.

A metodologia adotada foi estruturada em quatro fases sequenciais e interdependentes. A etapa inicial baseou-se na coleta e estruturação de dados, em que foram obtidas imagens e vídeos dos participantes para compor uma base representativa para a galeria e validação do sistema. Em seguida, procedeu-se à implementação do sistema, integrando as técnicas selecionadas ao protótipo funcional. Posteriormente, foram realizados os testes e validação, culminando na análise de resultados e métricas, fase na qual a consistência e a robustez das abordagens foram mensuradas. Cada uma dessas etapas é detalhada nas seções subsequentes, apresentando os procedimentos, as ferramentas e as justificativas técnicas que orientaram o desenvolvimento da pesquisa.

3.1 Coleta e Estruturação de Dados

A fase de coleta de dados foi fundamental para a criação de um conjunto de dados próprio, essencial tanto para o cadastro dos usuários quanto para a validação do sistema. Todo o procedimento foi conduzido com o consentimento informado e por escrito dos 10 (dez) voluntários participantes, seguindo rigorosamente os princípios de privacidade e segurança estabelecidos pela Lei Geral de Proteção de Dados (LGPD), conforme disposto em Brasil (2018).

As coletas foram realizadas nas dependências da Universidade Federal do Espírito Santo (UFES), seguindo um protocolo de captura padronizado. O ambiente utilizado possuía iluminação consistente e os voluntários foram posicionados contra um fundo neutro (branco) para minimizar ruídos visuais. Para a aquisição das mídias, utilizou-se a mesma webcam externa (Chipset Jieli VID 1224) para todos os sujeitos. Diferente de abordagens estritamente rígidas, optou-se por não fixar uma distância exata entre o rosto e a câmera; os participantes foram instruídos a se posicionarem de forma natural, simulando uma interação típica de controle de acesso (com distância estimada entre 40 cm e 80 cm). Essa variabilidade intencional permitiu avaliar a robustez do sistema frente a pequenas alterações de escala e postura, aproximando o experimento de um cenário real de aplicação.

Para a composição do banco de dados, foram capturadas 5 (cinco) imagens estáticas de cada participante. Essas imagens, destinadas à Galeria, foram utilizadas para gerar os vetores de características base dos usuários autorizados durante o processo de cadastro. Além das imagens, também foram realizadas gravações curtas, com duração de 10 segundos, de cada indivíduo. Esses vídeos foram empregados na etapa de validação para simular o acesso em tempo real, testando a capacidade de reconhecimento do sistema em fluxo contínuo.

Por fim, no que tange aos quesitos de segurança e armazenamento dos dados, todo o material coletado foi tratado de forma pseudonimizada. As informações pessoais foram desvinculadas dos arquivos de mídia bruta, sendo o armazenamento realizado em local seguro com acesso restrito exclusivamente aos pesquisadores responsáveis pelo projeto.

3.2 Implementação do Sistema

O desenvolvimento do sistema seguiu um *pipeline* modular de cinco etapas, estruturado sequencialmente conforme ilustrado na Figura 9. Esta arquitetura de ponta a ponta foi projetada para operar em tempo real, iniciando-se na captura da imagem e culminando na decisão binária de acesso. Para a execução e avaliação do protótipo, utilizou-se um notebook Lenovo Ideapad S145 equipado com processador AMD Ryzen 5 3500U e 12 GB de memória RAM. A escolha por este hardware de médio desempenho foi intencional, visando demonstrar a viabilidade técnica da solução proposta em dispositivos com recursos computacionais limitados. O ecossistema de software foi fundamentado na linguagem Python (v3.10.11), integrando bibliotecas de visão computacional de alto desempenho sobre o sistema operacional Windows 11 Pro.

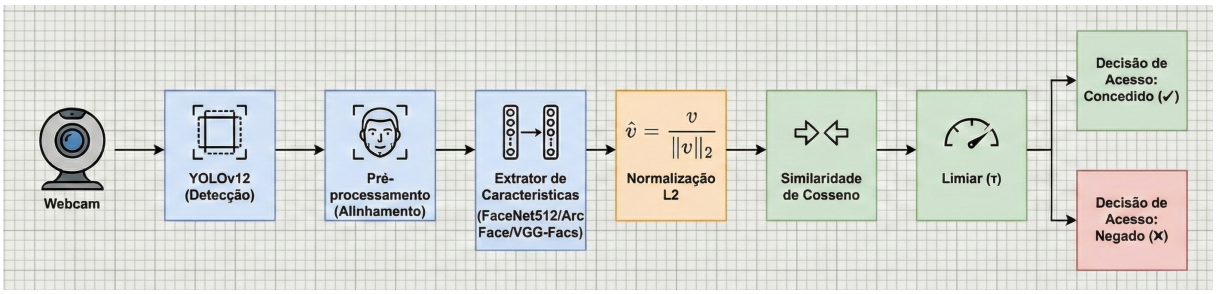


Figura 9 – Pipeline completo do sistema de reconhecimento facial proposto.

Fonte: Elaborado pelo autor.

O fluxo de processamento inicia-se pela aquisição de dados via webcam externa (Chipset Jieli), justificada pela necessidade de reduzir o ruído digital inerente aos sensores integrados de notebooks. Imediatamente após a captura, o quadro é submetido ao módulo de detecção representado na figura pelo bloco *yolov12n-face*. A escolha estratégica pela variante *nano* da família YOLOv12 permitiu o isolamento preciso da região de interesse

facial com baixa latência de inferência, o que é crucial para a fluidez do sistema. Diferente de métodos clássicos, esta abordagem baseada em redes neurais convolucionais demonstrou maior robustez a variações de iluminação, entregando apenas as coordenadas da face para a etapa subsequente.

Uma vez isolada, a imagem da face segue para o módulo de pré-processamento, onde técnicas de visão computacional realizam a padronização necessária. As operações incluem o redimensionamento da face para a dimensão de entrada exigida (160x160 pixels), a normalização dos valores de intensidade dos pixels e o alinhamento geométrico baseada na posição dos olhos. Na sequência, a imagem tratada é transformada em um vetor numérico utilizando a biblioteca *DeepFace* sobre o *framework* TensorFlow. A escolha por esta biblioteca deve-se à sua capacidade de abstrair a complexidade das redes profundas, facilitando a troca ágil entre diferentes arquiteturas de extração.

Visando uma avaliação robusta e representativa do estado da arte, este trabalho optou por não se limitar a um único extrator de características, mas sim realizar uma análise comparativa entre três arquiteturas que simbolizam diferentes paradigmas na evolução do reconhecimento facial. A escolha do VGG-Face justificou-se por seu papel histórico como *baseline* das Redes Neurais Convolucionais (CNNs) profundas clássicas, permitindo aferir o quanto as novas técnicas evoluíram em relação às abordagens tradicionais. O FaceNet512 foi selecionado para representar a eficiência da aprendizagem baseada em métricas de distância (*Triplet Loss*), focada no agrupamento espacial das faces. Por fim, o ArcFace foi incluído por representar o estado da arte em funções de perda baseadas em margem angular (*Angular Margin Loss*), prometendo maior discriminação entre classes. Vale ressaltar que o modelo *DeepFace* original foi desconsiderado devido a incompatibilidades técnicas com a versão atualizada do TensorFlow. Em todos os modelos utilizados nos experimentos, aplica-se a Normalização L2 ($\hat{v}$) sobre o vetor resultante, projetando-o em uma hiperesfera de raio unitário para garantir a estabilidade matemática necessária.

O estágio final do processamento envolve o cálculo da métrica de Similaridade de Cosseno entre o vetor da face capturada e os vetores da galeria. O resultado obtido é submetido a uma verificação de limiar (τ), valor ajustável definido na calibração. Como demonstrado na ramificação final da Figura 9, o sistema toma uma decisão determinística: caso a similaridade supere o limiar, o acesso é concedido (bloco verde); caso contrário, o acesso é negado (bloco vermelho).

3.3 Protocolo de Testes e Validação Automatizada

A validação do sistema proposto não foi realizada manualmente, mas sim através do desenvolvimento de um algoritmo de automação em Python, projetado especificamente para processar o conjunto de vídeos coletados e gerar métricas em lote. Este software de

validação opera varrendo o diretório de vídeos e instanciando o fluxo de reconhecimento facial para cada arquivo encontrado. Para garantir a fidelidade dos dados e simular um ambiente de produção, o algoritmo carrega previamente a base de dados de faces na memória e, para cada vídeo processado, gera um arquivo de registro detalhado em formato CSV. Esse arquivo armazena, quadro a quadro, o tempo de inferência do detector, o tempo de processamento do reconhecimento, a identidade predita pelo sistema e a pontuação exata de similaridade de cosseno, permitindo uma auditoria posterior precisa dos resultados.

Uma inovação metodológica importante deste trabalho foi a implementação dinâmica dos cenários de teste via lógica de programação, eliminando a necessidade de manipulação física dos diretórios de imagens. O algoritmo foi programado para alternar entre dois contextos distintos. No cenário de Usuário Genuíno, o vetor de características do participante presente no vídeo é mantido na galeria ativa, testando a capacidade de identificação correta (Verdadeiro Positivo). Já no cenário de Impostor Desconhecido, o software executa uma exclusão lógica temporária, removendo o vetor do sujeito da lista de busca durante a execução daquele vídeo específico. Isso força o sistema a comparar a face capturada apenas com os demais indivíduos do banco de dados, simulando um teste de mundo aberto onde o objetivo é negar o acesso (Verdadeiro Negativo).

Os parâmetros operacionais foram definidos através de uma etapa preliminar de calibração. Inicialmente, o sistema foi configurado com um limiar de decisão base de 0,40 para todos os modelos. A partir dessa execução inicial, analisou-se a distribuição das pontuações de similaridade geradas nos cenários de impostores para identificar a pontuação máxima atingida por um indivíduo não autorizado (o pior caso de falso positivo). O limiar final definitivo foi então ajustado individualmente para cada arquitetura (FaceNet512, ArcFace e VGG-Face), consistindo na soma desse valor máximo observado com uma margem de segurança adicional. Essa abordagem assegura que o sistema opere em um regime de segurança máxima, priorizando a eliminação de falsas aceitações. Complementarmente, a estratégia de amostragem temporal foi fixada em um a cada cinco quadros, visando estressar a capacidade de rastreamento do sistema.

3.4 Métricas de Avaliação

A análise quantitativa dos dados gerados nos arquivos de registro foi realizada por meio de métricas objetivas padronizadas na literatura de biometria. A métrica primária adotada foi a Acurácia (ACC), que oferece uma visão global do desempenho do classificador. Ela é calculada pela proporção total de predições corretas do sistema, somando-se os casos em que o usuário legítimo foi aceito (Verdadeiros Positivos - TP) e os casos em que o impostor foi corretamente barrado (Verdadeiros Negativos - TN), em relação ao volume total de amostras processadas, incluindo os erros de Falso Positivo (FP) e Falso Negativo

(FN).

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.1}$$

Além da acurácia global, é fundamental avaliar as falhas de segurança através da Taxa de Falsa Aceitação (FAR). Esta métrica representa a probabilidade de o sistema falhar em seu objetivo principal de proteção, permitindo indevidamente o acesso a um indivíduo não autorizado. No contexto do algoritmo desenvolvido, a FAR é contabilizada sempre que, no cenário de Impostor Desconhecido, o sistema encontra uma similaridade superior ao limiar calibrado com qualquer outra pessoa da base, gerando uma quebra de segurança.

$$FAR = \frac{FP}{FP + TN} \tag{3.2}$$

Por fim, calculou-se a Taxa de Falsa Rejeição (FRR), que reflete a eficiência operacional e a usabilidade da solução. Esta taxa indica a frequência com que o sistema nega acesso a um indivíduo legitimamente cadastrado e presente na galeria. Uma FRR elevada implica em transtornos para o usuário final, que precisa tentar o acesso múltiplas vezes. O equilíbrio entre a segurança (FAR) e a usabilidade (FRR) é determinado diretamente pela rigidez do limiar definido na etapa de calibração, demonstrando o compromisso inerente aos sistemas biométricos de alta segurança.

$$FRR = \frac{FN}{FN + TP} \tag{3.3}$$

4 Resultados e Discussão

Este capítulo apresenta os resultados obtidos nos experimentos realizados com o sistema de reconhecimento facial desenvolvido. A validação do sistema ocorreu em dois cenários distintos: (1) Autenticidade, visando avaliar a capacidade de reconhecimento de usuários cadastrados, e (2) Impostor, focado na capacidade de rejeição de indivíduos não autorizados.

Como etapa preliminar a esses cenários de reconhecimento, avaliou-se o desempenho do detector YOLOv12, componente que desempenhou um papel crucial no pré-processamento das imagens. A análise dos dados operacionais, fundamentada na média aritmética dos tempos de inferência registrados nos logs de auditoria de todos os vídeos processados, indicou que o tempo médio de detecção foi de 392,22 ms. Esta etapa foi determinante para o sucesso do sistema, pois garantiu que apenas faces de alta qualidade e corretamente recortadas fossem enviadas para os classificadores, contribuindo diretamente para a estabilidade e a alta acurácia reportada nas seções a seguir.

4.1 Análise Estatística de Similaridade

A robustez de um sistema biométrico baseia-se fundamentalmente na separabilidade entre as distribuições de pontuação de usuários genuínos e impostores. A Tabela 1 detalha a análise estatística dos *scores* de Similaridade de Cosseno obtidos.

Tabela 1 – Estatística descritiva dos scores e Gap de separabilidade entre cenários.

Modelo	Cenário	Média	Máximo	Mínimo	Gap
FaceNet512	Genuíno (C1)	0,8471	0,9357	0,7034	+0,1632
	Impostor (C2)	0,3349	0,5402	0,1577	
ArcFace	Genuíno (C1)	0,7502	0,9004	0,5079	-0,0686
	Impostor (C2)	0,2652	0,5765	0,1513	
VGG-Face	Genuíno (C1)	0,6802	0,9177	0,4180	-0,0291
	Impostor (C2)	0,2634	0,4471	0,1423	

Fonte: Elaborado pelo autor.

Observa-se que o modelo FaceNet512 apresentou a maior média de similaridade para usuários genuínos (0,8471), indicando alta confiança na identificação positiva. Em contraste, o VGG-Face demonstrou médias inferiores (0,6802), reflexo de uma arquitetura mais antiga e menos otimizada para espaços latentes compactos.

Uma métrica fundamental para avaliar a confiabilidade do sistema é o Gap de Separação (Margem de Segurança). Ele é definido como a distância entre o "pior caso" de cada classe: o menor score de similaridade registrado entre genuínos (o par de fotos da mesma pessoa que o modelo quase não reconheceu) e o maior score registrado entre impostores (o par de fotos de pessoas diferentes que o modelo quase confundiu).

Matematicamente, o Gap é calculado pela diferença entre o mínimo dos genuínos e o máximo dos impostores. O FaceNet512 destacou-se como o único modelo a apresentar um Gap Positivo (+0,1632), o que significa que existe um intervalo de valores onde não há sobreposição entre as classes, permitindo a definição de um limiar (threshold) que garante 0% de erro de classificação para esta base. Já o ArcFace (-0,0686) e o VGG-Face (-0,0291) apresentaram Gaps negativos, indicando que as distribuições se interceptam e que qualquer escolha de limiar resultará em algum nível de Falso Positivo ou Falsa Rejeição.

4.2 Definição de Limiares e Calibração de Segurança

Antes de mensurar a acurácia global do sistema, foi necessário estabelecer os critérios de corte (*thresholds*) definitivos que separam um acesso autorizado de uma tentativa de intrusão. Conforme estipulado na metodologia, a prioridade deste projeto foi estabelecer uma política de segurança rígida, visando anular a ocorrência de Falsos Positivos (acesso indevido).

Para isso, realizou-se uma análise preliminar da distribuição dos scores de similaridade obtidos exclusivamente no Cenário 2 (Impostor). O objetivo foi identificar qual foi a pontuação máxima atingida por um indivíduo não autorizado em cada modelo, ou seja, o "pior caso" de vulnerabilidade.

A Tabela 2 apresenta os resultados dessa calibração. A segunda coluna exibe o maior score de similaridade registrado entre uma face desconhecida e o banco de dados. Observa-se, por exemplo, que no modelo ArcFace, um impostor chegou a atingir uma similaridade de 0,5765. Se o limiar padrão de 0,50 fosse utilizado, esse intruso teria acessado o sistema.

Tabela 2 – Ajuste de limiares para regime de segurança máxima.

Modelo	**Máx. Score Impostor**	**Limiar Definido**	**Margem Segurança**
FaceNet512	0,5402	**0,60**	+0,06
ArcFace	0,5765	**0,60**	+0,03
VGG-Face	0,4471	**0,45**	+0,01

Fonte: Elaborado pelo autor.

Com base nesses valores críticos, os limiares operacionais (terceira coluna) foram

fixados ligeiramente acima do máximo registrado, garantindo uma margem de segurança (quarta coluna). Para o FaceNet512 e o ArcFace, fixou-se o limiar em 0,60, o que garante uma distância segura em relação aos scores de impostores (0,54 e 0,57, respectivamente). Já para o VGG-Face, que apresentou scores gerais mais baixos, o limiar de 0,45 foi suficiente para superar o máximo de 0,4471. Estes são os valores de referência utilizados para todas as métricas de acurácia apresentadas nas seções subsequentes.

4.3 Resultados Quantitativos Comparativos

Com os limiares calibrados para o regime de segurança máxima, avaliou-se o desempenho final dos modelos considerando as métricas de Acurácia, Taxa de Falsa Aceitação (FAR), Taxa de Falsa Rejeição (FRR) e latência média de processamento. A Tabela 3 sumariza o desempenho global obtido durante a validação automatizada.

Tabela 3 – Comparação final de desempenho entre os modelos.

Modelo	Acurácia Global	FAR (Impostor)	FRR (Genuíno)	Latência (ms)
FaceNet512	100,00%	0,00%	0,00%	739,3 ms
ArcFace	99,39%	0,00%	1,21%	481,7 ms
VGG-Face	98,27%	0,00%	3,47%	931,0 ms

Fonte: Elaborado pelo autor.

A análise dos dados revela que o modelo FaceNet512 se consolidou como a arquitetura mais robusta para a aplicação proposta. Foi o único extrator a atingir 100% de acurácia global, demonstrando capacidade de generalização suficiente para não cometer erros de falsa aceitação (falha de segurança) nem de falsa rejeição (falha de usabilidade) nas condições testadas. Em termos de desempenho temporal, sua latência média de 739,3 ms situa-se em um patamar intermediário. Embora não seja o modelo mais veloz do comparativo, este tempo de resposta representa um compromisso aceitável considerando a segurança absoluta proporcionada, sendo viável para aplicações de controle de acesso onde a precisão é o fator crítico.

Em contrapartida, o modelo ArcFace destacou-se pela eficiência computacional superior. Com uma latência média de apenas 481,7 ms, ele mostrou-se aproximadamente 48% mais rápido que o VGG-Face e 35% mais veloz que o FaceNet512. Apesar dessa vantagem em velocidade, o modelo apresentou uma taxa de falsa rejeição marginal de 1,21%, o que corresponde a 7 erros em 577 tentativas de acesso genuíno. Este valor resulta da divisão do total de *frames* dos vídeos de teste por cinco, uma vez que se adotou a estratégia de processar apenas 1 a cada 5 quadros para otimizar a fluidez em tempo real. Esses dados sugerem que o ArcFace apresenta viabilidade técnica para sistemas embarcados com recursos de processamento restritos ou cenários de alto fluxo de pessoas,

onde a velocidade de inferência é prioritária, desde que se tolere uma pequena taxa de rejeição de usuários legítimos.

Por fim, o modelo VGG-Face, utilizado neste estudo como linha de base histórica. O modelo apresentou o desempenho menos favorável em todos os quesitos, registrando a maior latência do comparativo (931,0 ms) e a maior taxa de falsa rejeição (3,47%). Na prática, essa taxa de erro implica que o usuário legítimo precisaria realizar mais tentativas para obter o acesso, prejudicando a experiência de uso. Portanto, conclui-se que sua utilização em novos projetos de reconhecimento facial deve ser restrita a fins de comparação acadêmica, sendo superado pelas arquiteturas mais modernas baseadas em margem angular e triplet loss.

4.4 Limitações Encontradas

Durante a execução dos experimentos, identificou-se uma limitação no Hardware de Captura, inicialmente, a webcam integrada do notebook (baixa resolução e ruído em baixa luz) degradou o desempenho dos modelos mais sensíveis. A substituição por uma webcam externa de maior fidelidade foi determinante para alcançar os resultados de 100% com o FaceNet512.

5 Conclusão

O objetivo desta pesquisa foi projetar e implementar uma solução de controle de acesso baseada em visão computacional que aliasse alta precisão na identificação biométrica à eficiência operacional necessária para aplicações reais. Diante desse desafio, o trabalho atingiu seu propósito ao desenvolver e validar um sistema capaz de operar com segurança e robustez. A integração do detector YOLOv12n-face com a arquitetura FaceNet512 provou ser a configuração ideal para o ambiente proposto, atendendo plenamente aos requisitos definidos no início da investigação.

A eficácia da solução foi comprovada isolando-se a métrica de separabilidade como prova definitiva: o sistema obteve uma acurácia global de 100% e, fundamentalmente, uma margem de separação positiva de +0,1632. Este resultado indica que, para as amostras analisadas, é possível distinguir usuários autorizados de tentativas de acesso não autorizado sem sobreposição de escores. Tal métrica corrobora a viabilidade do sistema para operar no ponto operacional estabelecido ($FAR \approx 0$), ressalvadas as limitações impostas pelo tamanho da amostra e pelas condições de ambiente controlado.

Para trabalhos futuros, sugere-se a implementação de mecanismos de detecção de vivacidade (*liveness detection*) para prevenir ataques de apresentação e a exploração de técnicas de quantização para portabilidade do sistema em dispositivos de borda com recursos limitados.

Referências

AMARAL, V. et al. Normalização espacial de imagens frontais de face em ambientes controlados e não-controlados. *FaSCi-Tech*, v. 1, n. 1, 2009. Citado na página 23.

ANDREZZA, I. L. P. *Análise de técnicas de normalização aplicadas ao reconhecimento facial*. Tese (Doutorado) — Universidade Federal da Paraíba, 2015. Citado na página 24.

BELHUMEUR, P. N.; HESPANHA, J. P.; KRIEGMAN, D. J. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 19, n. 7, p. 711–720, 1997. Citado na página 27.

BISHOP, C. M. *Pattern Recognition and Machine Learning*. [S.l.]: Springer, 2006. Citado na página 16.

BRASIL. *Lei nº 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais (LGPD)*. 2018. Diário Oficial da União, Brasília, DF, 15 ago. 2018. Citado 3 vezes nas páginas 12, 13 e 37.

BUOLAMWINI, J.; GEBRU, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In: *Conference on Fairness, Accountability and Transparency*. [S.l.: s.n.], 2018. p. 77–91. Citado na página 12.

CAO, Q. et al. Vggface2: A dataset for recognising faces across pose and age. *arXiv preprint arXiv:1710.08092*, 2018. Citado 2 vezes nas páginas 24 e 33.

CHEN, X.; WU, X.; ZHOU, H. Thermal infrared imaging for face recognition: A survey. *Pattern Recognition*, Elsevier, v. 39, n. 10, p. 1882–1894, 2006. Citado na página 18.

DALAL, N.; TRIGGS, B. Histograms of oriented gradients for human detection. In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2005. p. 886–893. Citado na página 19.

DENG, J. et al. Arcface: Additive angular margin loss for deep face recognition. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2019. p. 4685–4694. Citado 3 vezes nas páginas 28, 32 e 33.

DUDA, R. O.; HART, P. E.; STORK, D. G. *Pattern classification*. [S.l.]: John Wiley Sons, 2001. Citado 2 vezes nas páginas 25 e 34.

GIRSHICK, R. Fast r-cnn. In: *Proceedings of the IEEE International Conference on Computer Vision*. [S.l.: s.n.], 2015. p. 1440–1448. Citado na página 20.

GIRSHICK, R. et al. Rich feature hierarchies for accurate object detection and semantic segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2014. p. 580–587. Citado na página 20.

GOMES, J. S.; ARAUJO, F. H. D. Análise de técnicas de pré-processamento de imagem para reconhecimento facial baseada em vgg faces e ball tree. *Encontro Unificado de Computação do Piauí*, 2021. Citado na página 24.

GONZALEZ, R. C.; WOODS, R. E. *Digital image processing*. [S.l.]: Pearson Prentice Hall, 2008. Citado 3 vezes nas páginas 23, 24 e 25.

GONZALEZ, R. C.; WOODS, R. E. *Digital Image Processing*. 4. ed. [S.l.]: Pearson, 2018. Citado 2 vezes nas páginas 15 e 23.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. [S.l.]: MIT Press, 2016. Citado 2 vezes nas páginas 12 e 29.

GOWDA, P. H.; MULLA, D.; CLAY, D. E. Hyperspectral and multispectral remote sensing. In: *Applied Remote Sensing*. [S.l.]: American Society of Agronomy, Crop Science Society of America, Soil Science Society of America, 2006. p. 1–17. Citado na página 18.

Grand View Research. *Face Recognition Market Size, Share & Trends Analysis Report By Component (Hardware, Software), By Technology (2D, 3D, AI), By Application, By End-use, By Region, And Segment Forecasts, 2024 - 2030*. 2024. <https://www.grandviewresearch.com/industry-analysis/facial-recognition-market>. Acesso em: 12 jul. 2025. Citado na página 11.

GROTHER, P.; NGAN, M.; HANAOKA, K. *Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects*. [S.l.], 2019. Citado 2 vezes nas páginas 12 e 13.

HASSABALLAH, M.; ABDELMGEID, M. A.; AL-TURKI, S. Face recognition: challenges, achievements and future directions. *IET Computer Vision*, v. 9, n. 4, p. 14–30, 2015. Citado na página 17.

JAIN, A. K.; ROSS, A.; PRABHAKAR, S. An introduction to biometric recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, v. 14, n. 1, p. 4–20, 2004. Citado 5 vezes nas páginas 11, 12, 13, 15 e 35.

JAIN, A. K.; ROSS, A.; PRABHAKAR, S. *Handbook of fingerprint recognition*. [S.l.]: Springer Science & Business Media, 2011. Citado na página 34.

JESUS, L. et al. Análise de métodos de processamento de imagens para reconhecimento facial utilizando fisherfaces em imagens sob condições desfavoráveis. *ERBASE*, 2015. Citado na página 24.

KORTLI, Y. et al. Face recognition systems: A survey. *Sensors*, MDPI, v. 20, n. 342, p. 1–36, 2020. Citado na página 16.

LACERDA, H. W. Sistema automático móvel para gestão de presença com reconhecimento facial. *Repositório UFCG*, 2024. Disponível em: <http://dspace.sti.ufcg.edu.br:8080/xmlui/handle/riufcg/38146>. Citado na página 17.

LECUN, Y. et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, IEEE, v. 86, n. 11, p. 2278–2324, 1998. Citado na página 25.

LIU, W. et al. Sphereface: Deep hypersphere embedding for face recognition. *arXiv preprint arXiv:1704.08063*, 2018. Citado na página 33.

LUGARESI, C. et al. Mediapipe: A framework for perceiving and processing reality. In: *Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2019. Citado 2 vezes nas páginas 11 e 22.

MENDONçA, M. M. *Comparação entre métodos de normalização de iluminação utilizados para melhorar a taxa do reconhecimento facial*. Tese (Doutorado) — Universidade de São Paulo, 2008. Citado na página 24.

OKABE, R. K.; CARRO, S. A. Reconhecimento facial em imagens capturadas por câmeras digitais de rede. *Colloquium Exactarum*, v. 7, n. 2, p. 1–10, 2015. ISSN 2178-8332. Citado na página 17.

PARKHI, O. M.; VEDALDI, A.; ZISSERMAN, A. Deep face recognition. In: BMVA. *British Machine Vision Conference (BMVC)*. [S.l.], 2015. Citado 2 vezes nas páginas 28 e 29.

PHILLIPS, P. J. et al. Face recognition vendor test 2002: Evaluation results. *Face Recognition Vendor Test (FRVT)*, 2005. Citado 4 vezes nas páginas 23, 25, 33 e 34.

REDMON, J. et al. You only look once: Unified, real-time object detection. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2016. p. 779–788. Citado 3 vezes nas páginas 11, 20 e 21.

REN, S. et al. Faster r-cnn: Towards real-time object detection with region proposal networks. In: *Advances in Neural Information Processing Systems (NeurIPS)*. [S.l.: s.n.], 2015. v. 28, p. 91–99. Citado na página 20.

SCHROFF, F.; KALENICHENKO, D.; PHILBIN, J. Facenet: A unified embedding for face recognition and clustering. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2015. p. 815–823. Citado 10 vezes nas páginas 11, 23, 25, 26, 28, 29, 30, 31, 33 e 34.

SERENGIL, S. I.; OZPINAR, A. Lightface: A hybrid deep face recognition framework. In: IEEE. *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*. 2020. p. 23–27. Disponível em: <https://ieeexplore.ieee.org/document/9259802>. Citado 2 vezes nas páginas 27 e 28.

SINOSEEN. *Quantos tipos de sensores de imagem existem?* 2024. Disponível em: <https://www.sinoseen.com/pt/how-many-types-of-image-sensors-are-there>. Citado na página 18.

SISPONTO. *Biometria Facial vs. Biometria Digital: Qual o Melhor?* 2024. Disponível em: <https://www.sisponto.com.br/biometria-facial-vs-biometria-digital-qual-o-melhor>. Citado na página 18.

Swiftlane. *Limitations of Face Recognition in Access Control Systems*. 2020. <https://swiftlane.com/blog/limitations-of-facial-recognition/>. Accessed: [14/08/2025]. Citado na página 17.

TAIGMAN, Y. et al. Deepface: Closing the gap to human-level performance in face verification. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2014. p. 1701–1708. Citado 5 vezes nas páginas 11, 23, 25, 27 e 28.

TIAN, Y.; YE, Q.; DOERMANN, D. Yolov12: Attention-centric real-time object detectors. *arXiv preprint arXiv:2502.12524*, 2025. Citado na página 22.

TURING, A. M. Computing machinery and intelligence. *Mind*, v. 59, n. 236, p. 433–460, 1950. Citado na página 4.

TURK, M.; PENTLAND, A. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, v. 3, n. 1, p. 71–86, 1991. Citado na página 26.

VIOLA, P.; JONES, M. Rapid object detection using a boosted cascade of simple features. In: *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2001. v. 1, p. 511. Citado 2 vezes nas páginas 19 e 24.

WANG, H. et al. Cosface: Large margin cosine loss for deep face recognition. *arXiv preprint arXiv:1801.09414*, 2018. Citado na página 33.

WANG, X. et al. Deep learning for face detection: A survey. *arXiv preprint arXiv:1807.03901*, 2018. Citado na página 19.

ZAFEIRIOU, S. et al. A survey on face recognition pipelines: From traditional methods to deep learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 37, n. 10, p. 1925–1944, 2015. Citado na página 25.

ZHANG, K. et al. Joint face detection and alignment using multi-task cascaded convolutional networks. *IEEE Signal Processing Letters*, v. 23, n. 10, p. 1499–1503, 2016. Citado 2 vezes nas páginas 15 e 22.

ZHAO, W. et al. Face recognition: A literature survey. *ACM Computing Surveys (CSUR)*, ACM, v. 35, n. 4, p. 399–458, 2003. Citado 7 vezes nas páginas 19, 23, 24, 25, 26, 33 e 34.