

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO CEARÁ
PRÓ-REITORIA DE PESQUISA, PÓS-GRADUAÇÃO E INOVAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Iago Belarmino Lucena

**Análise de Abordagem Baseada em Rede Convolucional Aplicada a
Soluções de Reconhecimento Facial Frontal**

Fortaleza – CE

Dados Internacionais de Catalogação na Publicação
Instituto Federal do Ceará - IFCE
Sistema de Bibliotecas - SIBI
Ficha catalográfica elaborada pelo SIBI/IFCE, com os dados fornecidos pelo(a) autor(a)

- L935a Lucena, Iago.
Análise de Abordagem Baseada em Rede Convolucional Aplicada a Soluções de Reconhecimento Facial Frontal / Iago Lucena. - 2022.
70 f. : il. color.
- Dissertação (Mestrado) - Instituto Federal do Ceará, Mestrado em Ciência da Computação, Campus Fortaleza, 2022.
Orientação: Prof. Dr. Pedro Pedrosa Rebouças Filho.
1. detecção facial. 2. reconhecimento facial. 3. oclusão facial. 4. facenet. I. Título.

CDD 004

Iago Belarmino Lucena

Análise de Abordagem Baseada em Rede Convolucional Aplicada a Soluções de Reconhecimento Facial Frontal

Dissertação de Mestrado submetida ao Programa de Pós-Graduação em Ciência da Computação do Instituto Federal de Educação, Ciência e Tecnologia do Ceará, como requisito parcial para obtenção do grau de Mestre em Ciência da Computação.

Linha de Pesquisa: Processamento de Imagens e Sinais

Orientador: Prof. Dr. Pedro Pedrosa Rebouças Filho

Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE)

Pró-Reitoria de Pesquisa, Pós-Graduação e Inovação (PRPI)

Programa de Pós-Graduação em Ciência da Computação (PPGCC)

Fortaleza – CE

20 de dezembro de 2022



Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE)
Pró-Reitoria de Pesquisa, Pós-Graduação e Inovação (PRPI)
Programa de Pós-Graduação em Ciência da Computação (PPGCC)

Iago Belarmino Lucena

Esta defesa de dissertação foi julgada adequada para a obtenção do título de Mestre em Ciência da Computação, sendo aprovada pela Coordenação do Programa de Pós-Graduação em Ciência da Computação do Instituto Federal de Educação, Ciência e Tecnologia do Ceará e pela banca examinadora:

**Orientador: Prof. Dr. Pedro Pedrosa Rebouças
Filho**

Instituto Federal de Educação, Ciência e Tecnologia do
Ceará (IFCE)

**Prof. Dr. Carlos Mauricio Jaborandy de Mattos
Dourado Junior**

Instituto Federal de Educação, Ciência e Tecnologia do
Ceará (IFCE)

Prof. Dr. Roger Moura Sarmiento

Instituto Federal de Educação, Ciência e Tecnologia do
Ceará (IFCE)

Fortaleza – CE
20 de dezembro de 2022

Agradecimentos

Os agradecimentos desse trabalho vão inicialmente à minha esposa Caroline e aos meus pais João e Fátima, que estão sempre me apoiando a enfrentar novos desafios.

Agradecimento também ao Prof. Dr. Pedro Pedrosa, por todos os conselhos durante esse tempo de mestrado.

Por último, aos colegas do LAPISCO pela vivência e constante ajuda no dia-a-dia, em especial ao Fabrício, Adriell, Matheus, Lucas, Jerovanne, João Matheus, Leandro, Aldísio e Navar que me ajudaram bastante no desenvolvimento e estudo dessa dissertação.

Resumo

A tecnologia de reconhecimento facial vem constantemente sendo empregada nos mais diversos setores, desde análise de expressões faciais para medir satisfação do cliente em lojas, autorização de pagamentos, autenticação de login de celulares e até instrumento policial de identificação de criminosos. De um modo resumido, a técnica consiste em, a partir de uma imagem obtida por uma câmera identificar rostos contidos naquele foto e confrontá-los com uma base de dados de faces cadastradas previamente no sistema. A partir de avanços em termos de algoritmos e *hardwares* obtidos nos últimos anos, foi possível disponibilizar através de celulares modernos soluções para verificação e reconhecimento facial. O presente trabalho visa aprofundar a análise de uma solução de reconhecimento facial fazendo o uso da Rede Neural Convolucional *FaceNet*, serão avaliadas diferentes bases públicas (AR Dataset, Labeled Face in the Wild, CelebFaces e ExtendedYale) e uma base privada. O estudo procurou validar a utilização da FaceNet nessa base privada, com grande quantidade de faces, além de avaliar efeito da oclusão parcial em rostos nas bases públicas a partir do uso de máscaras faciais, tendo em vista o cenário de pandemia de Covid-19 que massificou a utilização desse item. Por meio de um fluxo inicial de detecção facial utilizando o método HOG, foi possível fazer uma análise exploratória dos dados de faces da base privada, medindo sua qualidade. Ainda utilizando HOG, também foi possível gerar máscaras faciais de forma artificial, utilizando a técnica *MaskTheFace*. Por fim, foi aplicada a rede *FaceNet* com pesos pré-treinados em ambas as bases para o problema, obtendo ótimos resultados utilizando diferentes de distâncias vetoriais (Euclidiana, Manhattan e distância dos cossenos) por meio do algoritmo de baseado em grafos que superam 90% de acurácia nos casos sem oclusão. Para casos com oclusão parcial da face, a mesma técnica empregada alcançou resultados de acurácia de aproximadamente 60% nos melhores casos.

Palavras-chave: Detecção e Reconhecimento facial, Oclusão Facial, FaceNet.

Abstract

Facial recognition technology is constantly used in the most diverse sectors, from the face expression analysis of customers in stores, authorization of payments, cell phone login authentication, and even a police tool for identifying criminals. In a nutshell, the technique consists of, from an image controlled by a camera, identifying faces in that photo and confronting them with a database of faces previously registered in the system. From advances in algorithms and *hardware* obtained in recent years, it was possible to provide solutions for verification and facial recognition through modern cell phones. The present work aims to deepen the analysis of a facial recognition solution using the Convolutional Neural Network *FaceNet*, different public databases (AR Dataset, Labeled Face in the Wild, CelebFaces, and ExtendedYale), and a private database. The study sought to validate the use of FaceNet in a database with a large number of faces, in addition to evaluating the effect of partial occlusion on faces from the use of face masks, because of the scenario of the Covid-19 pandemic that increased the use of that item. Through the proposed facial detection flow by using Histogram of Oriented Gradients technique, it was possible to make an explanatory data analysis of private database, measuring its quality. By still using HOG, artificial face masks were created, with a technique named *MaskTheFace*. Finally, FaceNet was applied to both bases, achieving good results with different vector distance (Euclidean, Manhattan and Cosine Distance) which overcome 90% of accuracy in the best results by applying a graph-based matching algorithm, for not occluded faces. By applying the same method for occluded faces, the maximum accuracy obtained is about 60%.

Palavras-chave: face detection and recognition ,facial occlusion, FaceNet.

Lista de Figuras

Figura 1	– Estrutura utilizada pela FaceNet. Após a passagem do conjunto de imagens pela arquitetura da rede, os atributos passam por uma camada de normalização L2 para então formar o vetor de 128 características. A função de custo (<i>Triplet Loss Function</i>) então é utilizada para atualizar os pesos da rede, baseada na premissa de que imagens de uma mesma pessoa têm de ter valores de distância menores quando comparados dois indivíduos diferentes. Fonte: Elaborado pelo autor.	19
Figura 2	– A imagem apresenta os passos intermediários do algoritmo de extração de características HOG. Fonte: Elaborado pelo autor	22
Figura 3	– Face original cinco primeiras <i>Eigenfaces</i> geradas. Fonte: (HESELTINE et al., 2003).	24
Figura 4	– Processo de extração de características utilizando LBP. Fonte: (SALTON; PRADO, 2018).	25
Figura 5	– Processo de criação do Histograma resultante do LBP sobre a imagem processada. Fonte: (SALTON; PRADO, 2018).	25
Figura 6	– Arquitetura utilizada DeepFace. Cada cor ilustra os mapas de características produzidos em cada camada. Fonte: (TAIGMAN et al., 2014).	26
Figura 7	– Arquitetura utilizada na DeepID2 e DeepID2+. Fonte: (SUN et al., 2014).	27
Figura 8	– Oclusão de faces em um reconhecimento facial. Fonte: (WANG; YUAN; YU, 2017).	28
Figura 9	– Diferença de luminosidade sobre faces da base de dados Yale. Fonte: (FORCZMANSKI; GEORGY; SHCHEGOLEVA, 2012).	29
Figura 10	– Mesma face com resoluções de pixels diferentes para cada figura. Fonte: (BOOM et al., 2007).	29
Figura 11	– Envelhecimento de um famoso jogador de futebol produzidas por um aplicativo chamado FaceApp.	30
Figura 12	– Face com os pontos fiduciais marcados. Fonte: (AQUINO et al., 2017)	31
Figura 13	– Exemplo de problema de busca em uma grande base de dados. Fonte: (WANG; OTTO; JAIN, 2016)	33
Figura 14	– Fluxo do processo de análise do reconhecimento facial. Fonte: Elaborado pelo autor.	36
Figura 15	– Fluxo de vetorização da imagem de face	36
Figura 16	– Representação gráfica em um espaço 2-D da medida de distância dos cossenos. Fonte: Elaborado pelo autor	37
Figura 17	– Ilustração da base de dados privada, representando os diferentes tipos de qualidade de fotos presentes. Fonte : Elaborado pelo autor.	38

Figura 18 – AR dataset	40
Figura 19 – LFW dataset	41
Figura 20 – Extended Yale Dataset	41
Figura 21 – CelebFaces-A Dataset	42
Figura 22 – Fluxo de mascaramento artificial de faces através da técnica <i>MaskThe-Face</i> . Fonte: Elaborado pelo Autor	43
Figura 23 – Histograma dos Scores de todas as detecções no conjunto de dados. Esta figura ilustra o total de Scores obtidos com filtragem de Scores negativos.	46
Figura 24 – Relação entre resolução da imagem e score obtido. O valor do score utilizado é o valor médio por dimensão.	48
Figura 25 – Distribuição do conjunto de imagem quanto aos Scores de detecção por filtragem com $\text{Score} \geq T$	48
Figura 26 – Histograma dos Scores de todas as detecções no conjunto de dados. Esta figura ilustra o total de Scores obtidos sem filtragem de Score mínimo.	49
Figura 27 – Distribuição do conjunto de imagem quanto aos Scores de detecção por filtragem com $\text{Score} \geq T$	50
Figura 28 – Histograma dos Scores de todas as detecções no conjunto de dados. Esta figura ilustra o total de Scores obtidos com filtragem de Scores negativos.	50
Figura 29 – Esquema de geração e organização da partição V1.1 com os <i>targets</i>	53
Figura 30 – Esquema de geração e organização da partições V2.1 e V3.1 com os <i>targets</i>	53
Figura 31 – Resultados para faces oclusas do dataset LFW	59
Figura 32 – Resultados para faces oclusas do dataset Yale Faces	60
Figura 33 – Resultados para faces oclusas do dataset AR Faces	61
Figura 34 – Resultados para faces oclusas do dataset CelebFaces A	62

Lista de Tabelas

Tabela 1	–	Variação de VAL de acordo com a resolução das imagens	20
Tabela 2	–	Visão geral da base de dados de imagens faciais.	39
Tabela 3	–	Lista dos 20 postos de coleta que apresentaram os maiores scores de detecção, em média. A listagem está ordenada em forma decrescente pelo Score Médio de detecção da face.	46
Tabela 4	–	Lista dos 20 postos de coleta que apresentaram os menores scores de detecção, em média. A listagem está ordenada em forma crescente pelo Score Médio de detecção da face.	47
Tabela 5	–	Estatísticas quanto ao total de detecções quando aplicada a filtragem por limiar T do Score de detecção.	51
Tabela 6	–	Resultados de acurácia obtidos pelo método NGT com os <i>datasets</i> de consulta V1.1, V2.1, V3.1 e V-mask usando como alvo o <i>dataset</i> com 7M de amostras.	55
Tabela 7	–	Resultados de acurácia obtidos pelo método NGT com os <i>datasets</i> de consulta V1.1, V2.1 e V-mask usando como alvo o <i>dataset</i> com 4.9M de amostras.	55
Tabela 8	–	Resultados para a LFW	56
Tabela 9	–	Resultados de acurácia obtidos pelo método NGT para o dataset LFW.	56
Tabela 10	–	Resultados de acurácia obtidos pelo método NGT para o dataset Yale Faces.	57
Tabela 11	–	Resultados de acurácia obtidos pelo método NGT para o dataset AR Faces.	57
Tabela 12	–	Resultados de acurácia obtidos pelo método NGT para o dataset CelebFaces A.	58

Lista de abreviaturas e siglas

CNN	Rede neural convolucional, traduzida do inglês Convolutional neural network
HOG	Histogramas de gradientes Orientados, do inglês Histogram oriented gradients
GPU	Graphics Processing Unit
LFW	Labeled Faces in the Wild
SVM	Support Vector Machine
NGT	Grafo e árvore de vizinhança, do inglês Neighborhood Graph and Tree
kNN	Método dos "k" vizinhos mais próximos, do inglês k-nearest neighbor
PB	Tons de "Cinza ou Preto e Branco" para imagens

Sumário

1	INTRODUÇÃO	13
1.1	Motivação	15
1.2	Descrição do Problema	15
1.3	Objetivos	15
1.3.1	Objetivo Geral	15
1.3.2	Objetivos Específicos	16
1.4	Organização da Dissertação	16
2	FUNDAMENTAÇÃO TEÓRICA	17
2.1	Técnicas para Reconhecimento Facial	18
2.2	Algoritmos de Detecção de Face	21
2.2.1	Histograma de Gradientes Orientados	21
2.2.2	Detecção de Face usando CNN	22
2.2.3	Algoritmos de extração de Atributos da Face	23
2.2.3.1	Extração de Atributos Utilizando Métodos Clássicos	24
2.2.3.2	Extração de atributos utilizando métodos baseados em Redes Neurais	26
2.3	Condições Necessárias para um Reconhecimento Efetivo	27
2.3.1	Principais Desafios	27
2.3.2	Qualidade da Detecção da Face	31
2.3.3	Análise de Performance	32
2.3.3.1	Técnicas de Análise de Tempo e Performance dos Métodos de Busca em Grandes Bases	33
3	METODOLOGIA	35
3.1	Fluxo de Análise do Problema de Reconhecimento Facial com a FaceNet	35
3.2	Extração de Características e Medição de Similaridade Vetorial	36
3.2.1	Formato de Matching	38
3.3	Bases de Dados	38
3.3.1	Base Privada de Imagens de Rostos	38
3.3.2	Base Pública de Imagens de Rostos	40
3.4	Experimentos de Reconhecimento Facial com Oclusão por Máscaras Faciais	42
4	RESULTADOS E DISCUSSÕES	44
4.1	Base de dados privadas	44

4.1.1	Score de Detecção de Faces	44
4.1.2	Resultados Preliminares	45
4.1.3	Análises Estatísticas nas Bases de Dados	45
4.1.3.1	Análise de Detecção Facial da Base Privada 1	45
4.1.3.2	Análise de Detecção Facial da Base Privada 2	48
4.1.4	Partições da Base Privada para Aplicação da FaceNet	51
4.1.4.1	<i>Dataset</i> de Consulta	51
4.1.4.2	Imagens em Tons de Cinza ou Imagens Preto e Branco	51
4.1.4.3	Idade	52
4.1.4.4	Qualidade de Detecção	52
4.1.5	Partições	52
4.2	Análise dos Resultados das Bases Públicas e Privadas.	54
4.2.1	Partições da Base de Imagens Privada para Fins de <i>Matching</i>	54
4.2.2	Experimentos com o <i>dataset</i> 1	55
4.2.3	Experimentos com o <i>dataset</i> 2	55
4.2.4	Bases Públicas com Máscaras	55
5	CONCLUSÕES	63
5.1	Trabalhos Futuros	63
	REFERÊNCIAS	65

1 Introdução

Reconhecimento facial é uma técnica de identificação biométrica que utiliza informações morfológicas e antropométricas extraídas da face humana com o objetivo de estabelecer o grau de similaridade entre imagens faciais diferentes (JUNIOR, 2020). Trata-se de uma temática de estudo motivada principalmente por aplicações de segurança e de verificação pessoal (JAIN; LI, 2011). O fato de ser um método de reconhecimento de pessoas menos invasivo comparado a outras identificações de biometria como por impressão digital e íris, combinado a disponibilidade e o desenvolvimento de computadores com alta capacidade de processamento e custos acessíveis, assim como sistemas embarcados, contribuíram para o aumento do interesse em diversas aplicações de processamento digital de imagens e vídeos.

Historicamente, os problemas de detecção e reconhecimento facial são resolvidos por três diferentes tipos de abordagem: local, holística e abordagem híbrida (KORTLI et al., 2020). A primeira leva em consideração somente certas características ou regiões da face e não ela toda. A segunda utiliza toda a região facial como dado de entrada e realiza uma transformação em um sub-espço vetorial menor para análise. Já a última utiliza características faciais locais e gerais para aumentar a acurácia do reconhecimento.

Existem diversos algoritmos e processos que são realizados até que seja possível o reconhecimento. O processo principal pode ser dividido em duas etapas básicas: detecção e reconhecimento da face. A detecção consiste na identificação de uma face dada uma imagem, a partir da utilização de algoritmo específico, sendo um problema bem mais complexo que a detecção de um objeto estático visto que a face de um indivíduo é um objeto dinâmico que apresenta diversos formatos e tonalidades de cores (MULLER; MAGAIA; HERBST, 2004).

Após a detecção, a extração de características faciais é necessária para que a classificação e o reconhecimento da pessoa sejam realizados. Nesse sentido, a identificação de pontos específicos da região facial é fundamental, sendo utilizados não só para a identificação desses pontos como também para o rastreamento da face e estimação do posicionamento da face (ZHU; RAMANAN, 2012). Entretanto, problemas no reconhecimento das componentes essenciais para a descrição da face, assim como na predição do formato correto da face da pessoa analisada dificultam o cálculo desses pontos (KAZEMI; SULLIVAN, 2014).

O reconhecimento facial ainda sofre com problemas que envolvem a iluminação do ambiente e a variação das posições dos rostos, assim como a oclusão da face por objetos de uso comum como óculos, bonés e cachecóis (MI et al., 2019; SCHROFF; KALENICHENKO; PHILBIN, 2015). Isso fica mais crítico no contexto da vigilância, análise forense e de

identificação com foto é uma tarefa cuja precisão é crítica. Muitas vezes, as limitações na qualidade geral das imagens reduzem a capacidade de tomar decisões sobre a identidade de uma pessoa (SUMA et al., 2019).

Embora diversos avanços tenham sido feitos para solucionar o problema do reconhecimento facial, muitos desafios ainda permanecem em aberto e necessitam de atenção, para que seja possível criar um sistema que consiga atender às necessidades de segurança e de verificação pessoal. Recentemente, por exemplo, com a pandemia de covid-19 iniciada em 2020, o uso de máscaras faciais como forma de proteção primária à transmissão da doença adicionou um desafio a mais ao problema de reconhecimento facial, devido ao feito de oclusão causado pela máscara, uma vez que até mesmo a percepção e reconhecimento de pessoas ao olho humano podem ser afetadas (FITOUSHI et al., 2021).

Nesse contexto, o *matching* em faces é uma técnica para reconhecimento facial que faz o confronto de duas imagens e analisa semelhanças entre determinadas áreas da face. Quanto maior as áreas correspondente *matching*, menor são as distâncias dos espaços dimensionais das imagens comparadas. Essa técnica são implementados para executar essa comparação em diversas faces e diferentes ambientes com fluxo de pessoas ou buscar um indivíduo no banco de dados facial a partir da sua foto (VOKHMINTCEV et al., 2016).

Levando em consideração a crescente relevância dos trabalhos relacionados ao reconhecimento facial, seja para aplicações de segura, autenticação, vigilância e entretenimento, os trabalhos relacionados para este estudo abordam essa área, bem como o avanço obtido o longo dos anos. Em geral, as técnicas buscam replicar o comportamento do olho e cérebro humano de interpretação e processamento do que é visto, e assimilação com a memória adquirida. A análise em questão trata do método desenvolvido pelo funcionários do Google em (SCHROFF; KALENICHENKO; PHILBIN, 2015), resultando na rede conhecida como *Facenet*. Trata-se de uma rede neural convolucional (CNN) que usa nas etapas de treinamento uma função de perda baseada em *triplets*, onde a partir de 3 imagens, sendo duas do mesmo indivíduo. O objetivo dela é fazer com que indivíduos similares ou iguais tenham representações de distância num dado espaço vetorial cada vez menor. A partir dessa abordagem, no mesmo trabalho foi possível obter uma acurácia de 99,38% na base LFW (HUANG et al., 2008a). Um dos principais ganhos do modelo diz respeito a redução considerável da representação matemática da imagem, que é um vetor 1x128 no espaço Euclidiano. Isso permite a medição da similaridade entre duas fotos através do cálculo simples da distância Euclidiana, por exemplo, o que possibilita inclusive ganhos de processamento para soluções de reconhecimento facial de escopo limitado e baixo custo, como em smartphones.

1.1 Motivação

O reconhecimento facial vem sendo uma importante ferramenta popularizada nos últimos anos. Diversas aplicações comerciais vêm sendo apresentadas ao públicas com funções que variam do entretenimento, segurança pública, controle de acesso entre outros.

Dentre as várias formas de se tentar resolver o problema do reconhecimento facial, foi desenvolvida a rede neural FaceNet (SCHROFF; KALENICHENKO; PHILBIN, 2015), cujo principal ganho é diminuir a dimensionalidade de imagens para fins de comparação, o que traz como consequência a possibilidade de uso de um sistema de reconhecimento com processamento em servidores ou até mesmo embarcado.

Esse trabalho visa avaliar o potencial de uso da FaceNet, tendo em vista bases públicas e uma base privada, de forma a comprovar sua eficácia, tendo em vista seu baixo custo computacional e potencial de utilização em aplicações em nuvem e uso comercial. Adicionalmente é também estudada uma situação típica de problema de reconhecimento facial envolvendo oclusão parcial de faces, através de máscaras.

1.2 Descrição do Problema

O problema em questão nesse trabalho diz respeito a analisar o reconhecimento facial de forma a validar o uso de rede neural pré-treinada. O grande benefício disso é justamente não precisar preparar a etapa de treinamento com sua própria base, o que pode consumir recursos computacionais, além de não depender amostras variadas de um mesmo indivíduo para que se consiga fazer a identificação. Serão utilizadas bases públicas e privadas na condução dos experimentos. Adicionalmente, procura-se também verificar se o efeito de oclusão causado por máscaras faciais afeta o desempenho dessa mesma rede neural, mantendo-se o fluxo de detecção e reconhecimento exatamente os mesmos de quando não há nenhum obstáculo na face.

1.3 Objetivos

1.3.1 Objetivo Geral

O objetivo geral desse trabalho consiste em avaliar uma abordagem para detecção e reconhecimento facial a partir de fotos frontais com a utilização da rede neural convolucional FaceNet. O estudo proposto busca a observação de cenários para detecção e reconhecimento facial humano, baseado em diferentes problemas e incluindo soluções para oclusão em imagens de faces.

1.3.2 Objetivos Específicos

- Apresentar soluções para detecção e reconhecimento facial frontal de faces humanas.
- Avaliar a performance da Rede Neural Convolucional FaceNet em diferentes situações de captura frontal de faces.
- Propor técnicas de oclusão para os problemas de detecção e reconhecimento facial, avaliando o impacto no desempenho da FaceNet.
- Avaliar possíveis métricas são melhores aplicáveis para resolução do problema em questão
- Identificar como diferentes padrões de imagens para captura de fotos podem ser definidos a partir de resultados obtidos para a referida rede.

1.4 Organização da Dissertação

Além do capítulo introdutório (Capítulo 1) esta dissertação conta com outros 5 capítulos e está dividida conforme se segue:

- Capítulo 2: o capítulo faz um apanhado sobre o problema de reconhecimento e detecção facial com as etapas relacionadas ao processol
- Capítulo 3: apresenta a metodologia utilizada, com as respectivas métricas aplicadas as bases de faces trabalhadasl
- Capítulo 4: apresenta os resultados obtidos propostos pela abordagem.
- Capítulo 5: apresenta a conclusão e sugestões de trabalhos futuros.

2 Fundamentação Teórica

Esta seção apresenta um apanhado sobre o método objeto de análise e as principais tecnologias que envolvem o processo de reconhecimento de faces, desde a detecção da face até a etapa de reconhecimento, abrangendo também os principais desafios presentes neste tema.

Mukherjee et al (MUKHERJEE et al., 2019) propuseram uma técnica que permite reduzir em 93% o gasto de energia de dispositivos móveis com aplicações de reconhecimento facial. Basicamente, esta técnica se divide em duas etapas: Na primeira fase, o dispositivo móvel é capaz de detectar a face e enviar apenas a região do rosto, no caso, a região de interesse, para um servidor na nuvem, e posteriormente essa face é utilizada no reconhecimento facial, este fluxograma diminui o processamento ocasionando uma redução de gasto de energia.

Schroff et al (SCHROFF; KALENICHENKO; PHILBIN, 2015), elaboraram uma arquitetura composta por camadas de entrada em lotes em uma CNN profunda capaz de detectar faces depois de treinadas. Em seus resultados os autores utilizaram um conjunto de 200 milhões de imagens de rostos de 8 milhões de diferentes pessoas. Em sua melhor arquitetura, os autores atingiram 99% de acurácia, o que comprova a eficácia do método para seu devido escopo. Vale ressaltar que o agrupamento facial coletado consistia de variações abundantes de condições de captura, como pose, oclusão, iluminação e faixa etária.

Amos et al (AMOS et al., 2016) desenvolveram uma biblioteca denominada OpenFace para o reconhecimento de faces. Essa biblioteca contava com uma rede altamente precisa e bom baixo tempo de treinamento e detecção. A arquitetura dessa é inspirada no modelo do RNA, denominado detector facial Dlib. Este mais usado devido à sua maior precisão em relação à OpenCV, na qual os experimentos indicaram que o tempo de treinamento para OpenFace também é menor. O tempo de previsão das outras técnicas aumenta quando o número de pessoas aumenta, mas o tempo de previsão do OpenFace permanece constante. Por fim, os autores descobriram que o OpenFace demonstrou precisão competitiva com resultados do estado da arte atual, mesmo usando um conjunto de dados duas ordens de magnitude menor

Em seu estudo, Peixoto et al (PEIXOTO et al., 2020) propuseram um novo método que realiza a compressão adaptativa dos atributos obtidos a partir de sistemas de reconhecimento facial, de forma a possibilitar a integração de sistemas IoT e técnicas de aprendizado de máquina e profundo para aplicações de reconhecimento facial, reduzindo custos computacionais, de rede, e energia. A compressão foi aplicada nos atributos extraídos

através do extrator FaceNet (SCHROFF; KALENICHENKO; PHILBIN, 2015), que gera em sua saída um vetor de atributos de dimensão 128, composto de características discriminatórias de cada face. Ao todo, para avaliar o método proposto, os autores utilizaram quatro conjuntos de dados, obtendo resultados significativos na compactação, com uma redução de 90% dos arquivos em relação ao seu tamanho original.

Além dos algoritmos citados até agora, o DeepID, que produz maior precisão com uma diferença relativamente pequena de 97,45% com um desvio padrão de 0,26% (SUN et al., 2014). O FaceNet, que teve a maior precisão de 99,63% (SCHROFF; KALENICHENKO; PHILBIN, 2015) que foi treinada com um conjunto de dados privado composto por milhões de imagens tiradas de mídias sociais (AMOS et al., 2016). Enquanto o Openface é um modelo desenvolvido a partir do DeepFace e GoogleNet que foi redesenhado com várias alterações que são treinadas usando o conjunto de dados CASIA e FaceScrub (AMOS et al., 2016).

Uma pesquisa sobre um sistema para detecção de presença em sala de aula estudou o uso da CNN FaceNet e Support Vector Machine para a detecção de faces. Após obtidas as imagens, os resultados para detecção de face eram extraídos com o auxílio da Facenet e então classificados através da SVM. Este estudo compara o desempenho de 3 arquiteturas de modelos de deep learning para a detecção de faces, cujos modelos são Facenet, VGG16 e Rede Neural Convolutacional (CNN). Os melhores resultados de precisão foram obtidos com o modelo Facenet, com uma taxa de precisão de 99,6%, apenas olhando para o desempenho do modelo de arquitetura de aprendizagem profunda, faltando avaliar o reconhecimento em tempo real e a determinação de limiares para reconhecimento facial (NYEIN; OO, 2019).

Por fim, uma pesquisa sobre um sistema de atendimento em tempo real, fez uso do reconhecimento facial com auxílio de técnicas de *deep learning*. O conjunto de dados utilizado foi era constituído de imagens de 28 alunos, com 10 fotos cada, constituindo 280 imagens. O reconhecimento facial fez uso do cálculo da distância euclidiana com um limite inferior a 0,6 para considerar o acerto, a precisão obtida foi de 95%(KUANG; BAUL, 2020).

2.1 Técnicas para Reconhecimento Facial

Desde o século 20, pesquisas tem sido realizadas para a detecção de padrões e formas em imagens computacionais. Um dos principais objetivos desse estudo é a detecção e reconhecimento facial, por exemplo: com apenas uma foto ou vídeo, conseguir a habilidade de reconhecer quem são as pessoas presentes.

Existem várias técnicas para reconhecimento facial, mas as principais utilizadas nos dias de hoje são a detecção com técnicas de aprendizagem de máquina, visão computacional

ou redes neurais. Porém nos primórdios, uma das primeiras técnicas foi a utilização do Histograma de Gradientes Orientados (do inglês: *Histogram of Oriented Gradients (HOG)*) (FREEMAN, 1994) para reconhecer faces em uma imagem.

O Histograma de Gradientes Orientados tenta construir uma caracterização mais detalhada da estrutura espacial com uma pequena janela de imagem. É uma etapa útil de pré-processamento para algoritmos que detectam objetos com estrutura quase-regular. Porém, com o aumento do uso de unidades de processamento gráfica (do inglês: *Graphics Unit Processing (GPU)*), o uso de redes neurais profundas em inteligência artificial, seguido pela maior disposição de dados disponíveis e maior poder computacional com o passar dos anos, a tarefa de reconhecimento facial começou a utilizar métodos de segmentação utilizando redes neurais convolucionais, como a FaceNet (SCHROFF; KALENICHENKO; PHILBIN, 2015).

Criada por pesquisadores do Google, o sistema é capaz de identificar faces mesmo com a presença de sombras na imagem ou o posicionamento do corpo tendendo para a direita. Com esse sistema, os pesquisadores conseguiram alcançar a métrica recorde de 95.73% de acurácia no dataset público chamado *AR Face Database* (MARTINEZ; BENAVENTE, 1998).

Uma característica importante da FaceNet é que ela utiliza uma rede convolucional treinada para otimizar os vetores de características, ao invés de utilizar camadas intermediárias para fazer a mesma tarefa como em implementações anteriores. Essa rede gera um vetor de atributos com apenas 128 características, tendo um baixo custo computacional comparado com outras redes de aprendizado profundo que apresentam milhares de atributos. A Figura 1 apresenta a arquitetura utilizada pela FaceNet.

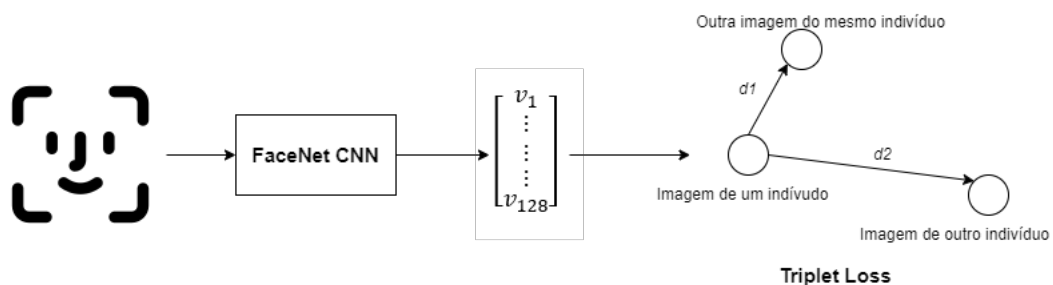


Figura 1 – Estrutura utilizada pela FaceNet. Após a passagem do conjunto de imagens pela arquitetura da rede, os atributos passam por uma camada de normalização L2 para então formar o vetor de 128 características. A função de custo (*Triplet Loss Function*) então é utilizada para atualizar os pesos da rede, baseada na premissa de que imagens de uma mesma pessoa têm de ter valores de distância menores quando comparados dois indivíduos diferentes. Fonte: Elaborado pelo autor.

A FaceNet é treinada de tal forma que a distância euclideana represente diretamente

a similaridade entre as faces, isto é, imagens da mesma face possuem vetores de atributos com distância pequena e imagens de faces diferentes possuem distâncias maiores. Essa otimização é parte da função de custo (*Triplet Loss Function*) utilizada para atualizar os pesos da rede. Os autores avaliaram o modelo desenvolvido através de verificação facial que é uma comparação entre duas faces para definir se as faces pertencem as mesmas pessoas ou a pessoas distintas. A métrica é definida pela taxa de validação $VAL(d)$:

$$VAL(d) = \frac{|TA(d)|}{|P_{same}|},$$

em que P_{same} são os pares de faces que pertencem a mesma pessoa e o conjunto de todos os verdadeiros positivos $TA(d)$ é definido por

$$TA(d) = (i, j) \in P_{same} \text{ com } D(x_i, x_j) \leq d,$$

em que (i, j) é o par de faces a ser comparado, $D(x_i, x_j)$ é a distância L2 entre os vetores de 128 características extraídos pela FaceNet e d é um limiar definido pelos autores.

Um dos resultados importantes apresentados pelos autores é o resultado obtido com diferentes tamanhos de imagens de faces. Esse resultado é apresentado na Tabela 1. Os resultados obtidos com imagens com até 120x120 pixels apresentam resultados próximos e VAL acima de 80% e imagens com 40x40 pixels apresentaram VAL de 37,8%, bem abaixo dos demais valores, não sendo uma boa opção para realizar o reconhecimento facial.

Tabela 1 – Variação de VAL de acordo com a resolução das imagens

#pixels	VAL (%)
1600	37,8
6400	79,5
14400	84,5
25600	85,7
65536	86,4

O método ainda atingiu 99.63% de acurácia na base LFW (HUANG et al., 2008b), ultrapassando os resultados obtidos pela DeepID2+.

Por fim, levando em consideração os trabalhos abordados ao longo desta secção, bem como o apelo por técnicas para reconhecimento facial de bons resultados, baixo custo computacional e energético, além do uso de dispositivos com fácil integração com IoT, visando ainda a contribuição para a comunidade acadêmica com uma técnica de baixo gasto de processamento e resultados satisfatórios, o presente estudo contempla a criação de uma técnica de captura, envio e análise de fotos frontais capturadas através de celular para fins de reconhecimento facial, baseada na FaceNet.

2.2 Algoritmos de Detecção de Face

Aqui serão apresentados os algoritmos hoje disponíveis na literatura para a primeira etapa do processo: a detecção da face. Neste sentido, o objetivo é encontrar a face a partir de uma imagem estática ou oriunda de um *frame* de vídeo, de forma a minimizar o quanto possível as detecções que representem falsos positivos. A seguir, será apresentado um algoritmo baseado em técnicas clássicas, conhecido como HOG. Posteriormente, será apresentada uma tecnologia mais recente de detecção, baseada em um modelo do tipo *deep learning*.

2.2.1 Histograma de Gradientes Orientados

O método de extração de características de imagens HOG (do inglês: *Histogram of Oriented Gradients*) foi proposto por Dalal e Triggs (2005) cujo objetivo era a detecção de pedestres. Na prática, esse método atua extraindo informações referentes à orientação das arestas contidas em uma imagem, sendo estas arestas calculadas por meio de métodos de detecção de bordas como o Sobel.

O método recebe uma imagem como entrada e retorna um vetor de características n dimensional. O método é baseado na distribuição (histogramas) das direções dos gradientes (gradientes orientados). Os gradientes (derivados x e y) de uma imagem são úteis porque a magnitude dos gradientes é grande em torno das bordas e cantos (regiões com mudanças bruscas de intensidade) e sabe-se que as bordas e os cantos contêm muito mais informações sobre a forma do objeto do que as regiões planas.

Assumindo que as imagens estão escala de cinza, o método está dividido em quatro fases: (1) cálculo da orientação e magnitude das arestas na imagem; (2) divisão da imagem em blocos e células; (3) cálculo do histograma de orientação dos gradientes por células, posteriormente agrupados em blocos; (4) por fim, a concatenação destes histogramas, formando assim o vetor descritor HOG (CRUZ; GUIMARÃES; SHIGUEMORI, 2012).

A Figura 2 apresenta as etapas seguidas pelo algoritmo de extração de atributos HOG. Para calcular o descritor HOG, inicialmente é necessário calcular os gradientes horizontal e vertical para obter o histograma de gradientes. Para isso, pode-se aplicar o filtro de Sobel. Com isso, é possível realizar o cálculo da orientação e magnitude das arestas da imagem em tons de cinza. Pode-se observar que a imagem de gradiente (segunda imagem da esquerda para a direita na Figura 2) removeu muitas informações não essenciais (por exemplo, regiões com intensidade constante), mas contornos destacados.

Posteriormente, duas estruturas são definidas denominadas células e blocos. A cada pixel, o gradiente tem uma magnitude e uma direção. Para fornecer uma representação mais compacta e robusta ao ruído, o histograma de gradientes é calculado sobre blocos (janelas na cor vermelha) de dimensões $v \times v$ que agrupa várias células (janelas na cor

verde).

O próximo passo é criar um histograma de gradientes sobre cada bloco $v \times v$. Para que a extração seja independente das variações de iluminação, deve-se normalizar tornando o descritor menos invariante à iluminação e a sombras. Cada bloco agora pode ser representado por um histograma da orientação do gradiente de cada célula.

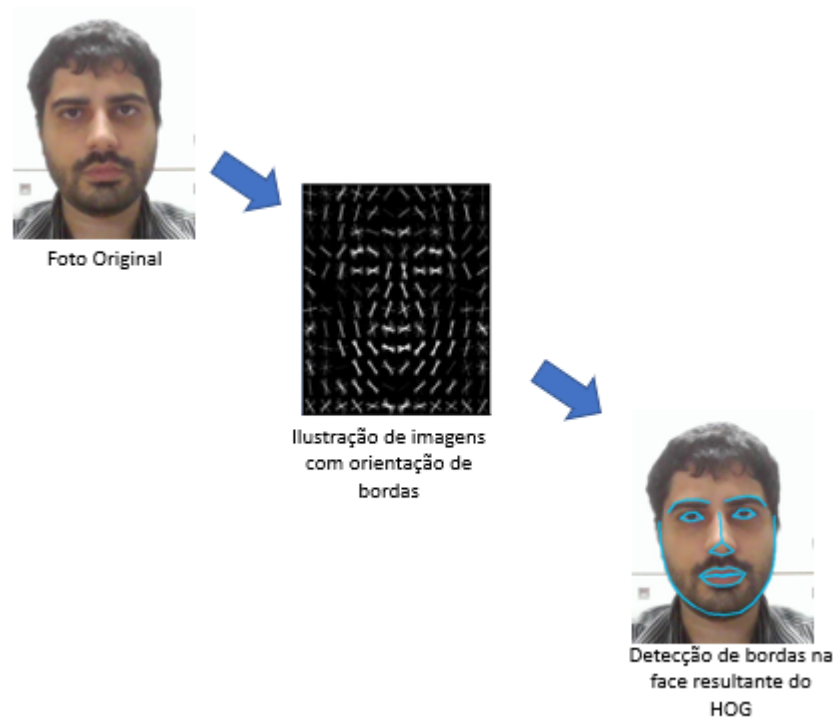


Figura 2 – A imagem apresenta os passos intermediários do algoritmo de extração de características HOG. Fonte: Elaborado pelo autor

2.2.2 Detecção de Face usando CNN

Uma CNN (do inglês: *Convolutional Neural Network*) (LECUN; BENGIO et al., 1995) é um tipo especial de rede neural. Ela incorpora conhecimento prévio sobre o sinal de entrada e suas distorções em sua arquitetura. As redes neurais convolucionais combinam características locais e pesos compartilhados, bem como utilizam subamostras espaciais para garantir um certo nível de movimento, escala e invariância à deformação. Usando os campos receptivos locais, os neurônios podem extrair características visuais simples, como cantos e pontos finais. Essas características elementares são então vinculadas pelas camadas seguintes para detectar características mais complicadas.

Uma rede convolucional típica contém um conjunto de camadas, das quais cada uma das camadas consiste em um ou mais filtros. Cada unidade no filtro é conectada à uma vizinhança local na camada anterior. A unidade pode ser vista como um detector de características local cuja característica de ativação é determinada no estágio de aprendizado.

As saídas desse conjunto de unidades constituem um mapa de características. As unidades em um mapa de características são restritas a executar a mesma operação em diferentes partes da imagem de entrada ou nos mapas de características anteriores, extraíndo diferentes atributos da mesma imagem. Um mapa de características pode ser obtido de maneira sequencial através da varredura da imagem de entrada por uma única unidade com pesos formando um campo receptivo local e armazenando as saídas desta unidade em locais correspondentes no mapa de características. Esta operação é equivalente a uma convolução com um pequeno kernel. O mapa de características pode ser tratado como um plano de unidades que compartilham pesos. As camadas de subamostragem que normalmente seguem as camadas com mapas de características locais e convolucionais introduzem um certo nível de invariância a distorções e translações.

As CNNs provaram ser muito poderosas no aprendizado discriminativo de características visuais. Por exemplo, (ZHANG; ZHANG, 2014) construíram uma CNN que pode aprender simultaneamente a decisão de face/não-face, o problema de estimativa de pose de face e o problema de localização de marcos faciais. (YANG et al., ; RANJAN; PATEL; CHELLAPPA, 2015) propuseram a detecção facial utilizando CNN na tarefa de extração de características. (FARFADE; SABERIAN; LI, 2015) realizou um *fine tuning* em uma CNN treinada na tarefa de classificação do ImageNet voltado para a detecção facial. Outro trabalho utilizando redes convolucionais totalmente conectadas foi proposto por (YANG et al., 2015) para gerar mapa de calor de partes faciais e, em seguida, usa o mapa de calor para gerar propostas de rostos.

Neste trabalho, alguns métodos de detecção facial foram estudados, dentre eles, a detecção facial baseada nos algoritmos HOG e CNN, disponibilizados pelo *framework dlib* (KING, 2009). De acordo com a descrição disposta pela biblioteca ¹ a arquitetura possui três camadas de downsampling. Essas camadas reduzem o tamanho da imagem em 8x e produz um mapa de características com 32 dimensões. Em seguida, uma sequência de quatro camadas convolucionais são adicionadas antes da camada final da rede. A última camada possui apenas um canal e a função de custo para o cálculo do erro é *Maximum-Margin Object Detector* (MMOD) (KING, 2015).

2.2.3 Algoritmos de extração de Atributos da Face

Passada a etapa de detecção da face, agora é necessário com base na região que representa a face, identificar atributos de relevância tal que permita diferenciar duas faces de pessoas distintas. Tais atributos, representarão as características únicas do indivíduo. Neste sentido, há um grau de preocupação ainda maior para evitar técnicas que produzam conjuntos de atributos ambíguos.

¹ Disponível em: <<http://dlib.net/>>

2.2.3.1 Extração de Atributos Utilizando Métodos Clássicos

Eigenfaces refere-se a uma abordagem baseada na aparência para reconhecimento de rosto, que busca capturar a variação em uma coleção de imagens de rosto e usa essas informações para codificar e comparar imagens de rostos individuais de maneira holística. Especificamente, as *Eigenfaces* são os principais componentes de uma distribuição de faces, ou equivalentemente, os autovetores da matriz de covariância do conjunto de imagens de faces, em que uma imagem com N pixels é considerada um ponto (ou vetor) no espaço N -dimensional. A ideia de usar componentes principais para representar rostos humanos foi desenvolvida por Sirovich e Kirby (SIROVICH; KIRBY, 1987), sendo usada por Turk e Pentland (TURK; PENTLAND, 1991) e Heseltine et al. (HESELTINE et al., 2003) para detecção e reconhecimento de rostos, como mostra a Figura 3.



Figura 3 – Face original cinco primeiras *Eigenfaces* geradas. Fonte: (HESELTINE et al., 2003).

A abordagem utilizando *Eigenfaces* é considerada por muitos como a primeira tecnologia de reconhecimento facial em funcionamento e serviu de base para um dos principais produtos comerciais de tecnologia de reconhecimento facial. Desde seu desenvolvimento e publicação inicial, houveram muitas extensões ao método original e muitos novos desenvolvimentos em sistemas de reconhecimento de rosto automático. Porém *Eigenfaces* ainda é frequentemente considerado como um método de comparação base para demonstrar o desempenho mínimo esperado qualquer sistema de reconhecimento facial.

O Histograma é uma representação da distribuição de pixels em uma imagem, criada por Pearson et al. (PEARSON, 1895). É amplamente usada para representação numérica de dados numéricos e estima a distribuição de probabilidade de uma variável contínua. Para imagens, se possuir apenas tons de cinza, o histograma irá representar o número de pixels que possuem cada valor da faixa de tons de cinza, porém se a imagem possuir mais de um canal, cada canal deverá possuir um histograma para representar a distribuição de pixels da imagem (JULINA; SHARMILA, 2017).

Suponha que tenhamos uma imagem facial em escala de cinza. Podemos obter parte dessa imagem como uma janela de 3×3 pixels. Também pode ser representado como uma matriz 3×3 contendo a intensidade de cada pixel (0 a 255). Então, precisamos levar o valor central da matriz a ser usado como limite. Este valor será usado para definir os novos valores dos 8 vizinhos. Para cada vizinho do valor central (limite), definimos um novo valor binário. Definimos 1 para valores iguais ou superiores ao limite e 0 para valores

inferiores ao limite. Agora, a matriz conterá apenas valores binários (ignorando o valor central). Precisamos concatenar cada valor binário de cada posição da linha da matriz por linha em um novo valor binário (por exemplo, 10001101). Em seguida, convertemos esse valor binário em um valor decimal e o definimos como o valor central da matriz, que na verdade é um pixel da imagem original. No final deste procedimento (procedimento LBP), temos uma nova imagem que representa melhor as características da imagem original, como mostra a Figura 4.

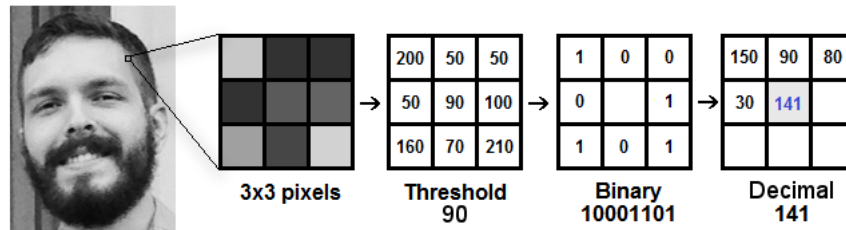


Figura 4 – Processo de extração de características utilizando LBP. Fonte: (SALTON; PRADO, 2018).

Com base na imagem acima, podemos extrair o histograma de cada região da seguinte maneira: Como temos uma imagem em escala de cinza, cada histograma (de cada grade) conterá apenas 256 posições (0 – 255) representando as ocorrências de cada intensidade de pixel. Então, precisamos concatenar cada histograma para criar um novo e maior histograma. Supondo que tenhamos grades 8x8, teremos $8 \times 8 \times 256 = 16.384$ posições no histograma final. O histograma final representa as características da imagem original, demonstrado na Figura 5.

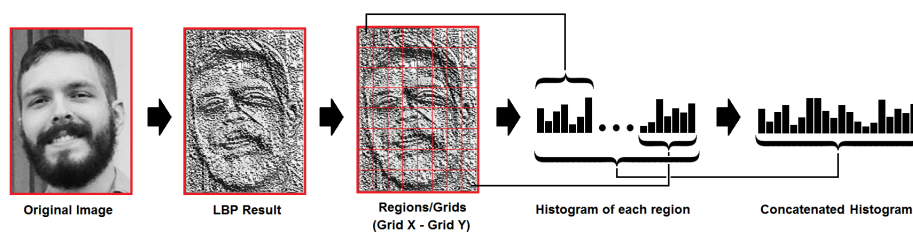


Figura 5 – Processo de criação do Histograma resultante do LBP sobre a imagem processada. Fonte: (SALTON; PRADO, 2018).

Outra forma de representar os dados de entrada é encontrar um subespaço que represente a maior parte da variação de dados. Isso pode ser obtido com o uso da Análise de Componentes Principais (PCA) (do inglês: *Principal Component Analysis*) (PEARSON, 1901). Quando aplicado a imagens de rosto, a Análise de Componentes Principais geram um conjunto de autofaces. Essas autofaces são os autovetores associados aos maiores autovalores da matriz de covariância dos dados de treinamento. Os vetores próprios encontrados correspondem à solução dos Mínimos Quadrados (do inglês: *Least Squares*)

(ETEMAD; CHELLAPPA, 1997). Essa maneira de representar os dados garante que a variação dos dados seja mantida, eliminando correlações existentes desnecessárias entre os recursos originais (dimensões) nos vetores de amostra.

Quando o objetivo é classificação e não representação, a solução dos Mínimos Quadrados pode não produzir os resultados mais desejáveis. Nesses casos, deseja-se encontrar um subespaço que mapeie os vetores de amostra da mesma classe em um único ponto da representação do recurso e os de classes diferentes, o mais afastados possível um do outro. As técnicas derivadas para atingir esse objetivo são conhecidas como Análise Discriminante (do inglês: *Discriminant Analysis*) (SWETS; WENG, 1996). A Análise Discriminante mais conhecida é a Análise Discriminante Linear (LDA) (do inglês: *Linear Discriminant Analysis*), que pode ser derivada de uma ideia sugerida por R.A. Fisher em 1936 (FISHER, 1936). Quando o LDA é usado para encontrar a representação do subespaço de um conjunto de imagens de face, os vetores de base resultantes que definem esse espaço são conhecidos como *Fisherfaces* (BELHUMEUR; HESPANHA; KRIEGMAN, 1997).

2.2.3.2 Extração de atributos utilizando métodos baseados em Redes Neurais

Taigman et al. (TAIGMAN et al., 2014) propôs uma rede para alinhamento 3D de faces para ajudar no processo de reconhecimento facial criando representações de faces através da utilização de uma rede neural profunda composta por 9 camadas. A rede é composta por mais de 120 milhões de parâmetros e é treinada para a tarefa de reconhecimento facial de mais 4000 pessoas. A Figura 6 apresenta a arquitetura utilizada pelos autores.

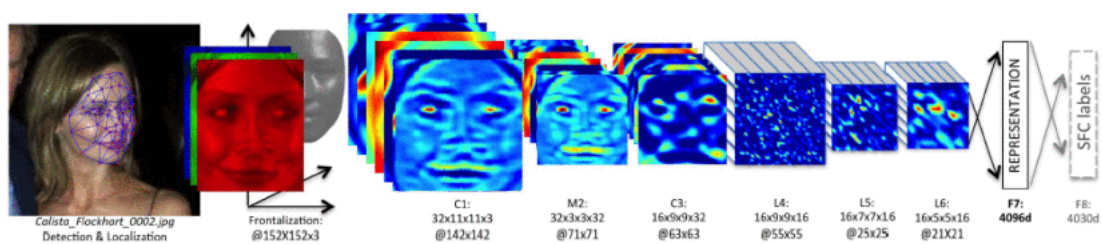


Figura 6 – Arquitetura utilizada DeepFace. Cada cor ilustra os mapas de características produzidos em cada camada. Fonte: (TAIGMAN et al., 2014).

Os autores também utilizaram uma rede Siamesa. Uma vez aprendida, a rede é replicada duas vezes e os atributos são utilizados diretamente para verificar se as imagens pertencem as mesmas pessoas ou a pessoas distintas. Os autores atingiram 97.35% de acurácia na base *Labeled Faces in the Wild* (LFW) que é uma das bases mais utilizadas no ambiente acadêmico na tarefa de verificação facial.

Sun et al. propuseram uma rede convolucional chamada DeepID2+ (SUN; WANG; TANG, 2015) que herda as características da DeepID2 (SUN et al., 2014), que possui 4

camadas convolucionais, com 20, 40, 60 e 80 vetores de características seguidos por uma camada com dimensão de 160 totalmente conectadas com a terceira e quarta camada convolucional. Essa última camada com dimensão de 160 é supervisionada tanto pela tarefa de identificação quanto pela tarefa de verificação facial. A Figura 7 apresenta a arquitetura utilizada pela DeepID2 e DeepID2+.

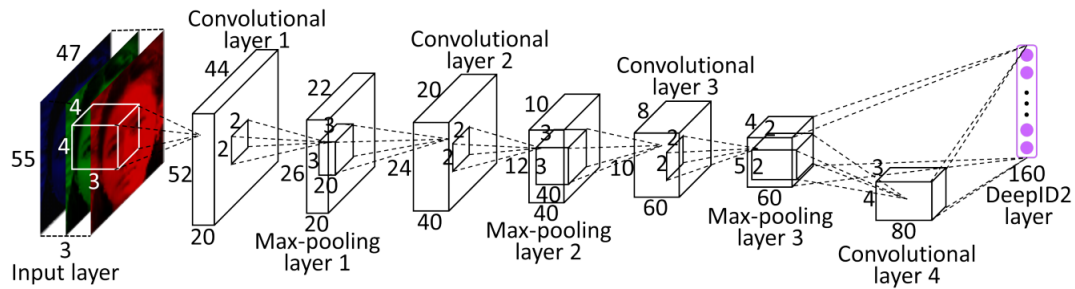


Figura 7 – Arquitetura utilizada na DeepID2 e DeepID2+. Fonte: (SUN et al., 2014).

Segundo os autores, esse modelo não requer alinhamento 2D/3D das faces detectadas. Os autores atingiram 99.47% de acurácia na base LFW.

2.3 Condições Necessárias para um Reconhecimento Efetivo

Esta seção apresenta os principais desafios dentro da área de pesquisa de detecção e reconhecimento de faces. Exploraremos a seguir breves definições dos principais problemas e as formas previstas na literatura como possibilidades de solução a estas questões.

2.3.1 Principais Desafios

Existem alguns problemas para tornar o reconhecimento facial efetivo, tal como a oclusão, brilho intenso, qualidade insatisfatória da câmera utilizada para capturar cada imagem, em diante.

Segundo (QUIRITA, 2014), um sistema de reconhecimento facial é considerado ideal se, para qualquer imagem de entrada, a identidade fornecida pelo comparador for a correta. Porém, situações não ideais como a inconstância na iluminação do ambiente, a oclusão, a resolução da imagem, a idade do indivíduo no período em que as amostras foram coletadas, o tipo de captura, bem como o uso de acessórios externos são barreiras que devem ser contornadas para que o reconhecimento de pessoas seja eficaz.

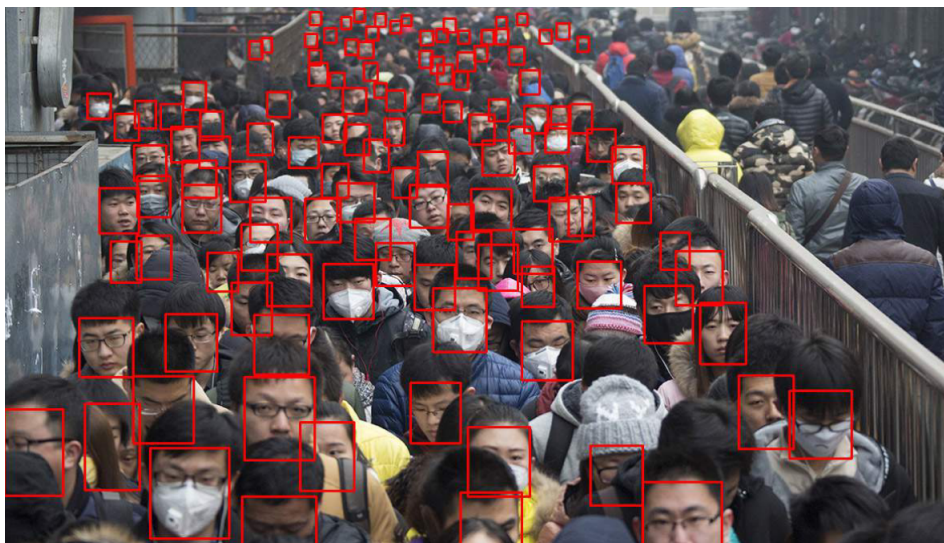


Figura 8 – Oclusão de faces em um reconhecimento facial. Fonte: (WANG; YUAN; YU, 2017).

Oclusão significa que há algo que você deseja ver, mas não pode devido a alguma propriedade da configuração do sensor ou a algum evento. Exatamente como ele se manifesta ou como você lida com o problema variará devido ao problema em questão. No desenvolvimento de um sistema que rastreia objetos (pessoas, carros), a oclusão ocorre se um objeto que você estiver rastreando estiver oculto (ocluído) por outro objeto, como mostra a Figura 8. Como duas pessoas passando um pelo outro, ou um carro que passa debaixo de uma ponte (WANG; YUAN; YU, 2017). O problema nesse caso é o que você faz quando um objeto desaparece e reaparece novamente.

Utilizando uma câmera de alcance, a oclusão é uma área em que você não possui nenhuma informação. Algumas câmeras de alcance a laser funcionam transmitindo um feixe de laser para a superfície que você está examinando e, em seguida, configurando uma câmera que identifica o ponto de impacto desse laser na imagem resultante (GHIASI; FOWLKES, 2015). Isso fornece as coordenadas 3D desse ponto. No entanto, como a câmera e o laser não estão necessariamente alinhados, pode haver pontos na superfície examinada que a câmera pode ver, mas o laser não pode atingir (occlusão). O mesmo pode ocorrer na imagem estéreo se houver partes da cena que são vistas apenas por uma das duas câmeras. Obviamente, nenhum dado de intervalo pode ser coletado a partir desses pontos.



Figura 9 – Diferença de luminosidade sobre faces da base de dados Yale. Fonte: (FORCZMANSKI; GEORGY; SHCHEGOLEVA, 2012).

Luminosidade pode ser definida como a quantidade de energia emitida por uma fonte de luz em relação à fonte à qual estamos comparando. Utilizando um vídeo de uma câmera para reconhecimento facial, a luminosidade pode afetar tanto positivamente quanto negativamente para o reconhecimento efetivo de um rosto (FINLAYSON, 2018), como mostra a Figura 9. Se uma fonte de luz muito forte estiver sendo emitido sobre uma face, os contornos do rosto podem ser mal capturados ou não capturados, tornando a tarefa de reconhecimento facial mais difícil. Para um reconhecimento facial efetivo, é importante que as condições de luminosidade e contraste estejam equilibradas, de tal modo que as formas do rosto sejam capturadas corretamente (FORCZMANSKI; GEORGY; SHCHEGOLEVA, 2012).

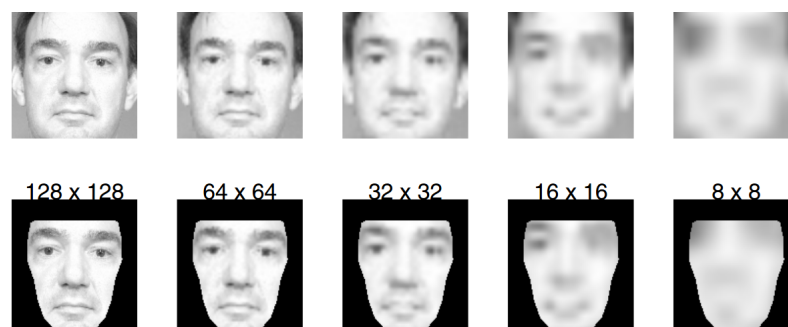


Figura 10 – Mesma face com resoluções de pixels diferentes para cada figura. Fonte: (BOOM et al., 2007).

Resolução da imagem é definida no Processamento Digital de Imagens, como a quantidade de pixels por linhas e colunas em uma matriz que representa a imagem. Imagens de câmeras de segurança normalmente possuem uma resolução HD (High Definition) de 1280x720 pixels (1920 pixels em cada linha e 720 pixels em cada coluna da matriz). Para o

reconhecimento facial efetivo, é desejável que a câmera possua uma resolução HD ou FHD (1920x1080 pixels), pois com uma resolução menor a quantidade de pixels que representa uma pessoa ou sua face pode possuir um grau insuficiente de detalhes, dependendo da sua distância com a câmera de captura (LI et al., 2018), como é representado na Figura 10.

Como o modelo de reconhecimento facial deve identificar o indivíduo instantaneamente e enquanto o mesmo estiver sendo procurado, a análise pela idade é de fundamental importância. Isso se dá pelo fato de não ser possível precisar o tempo até que a pessoa procurada seja encontrada, logo as características faciais dela podem sofrer alterações e isso é consequência imediata também do envelhecimento comum a todo indivíduo. Esse e outros fatores como expressões faciais, óculos, bonés, pelos faciais, por exemplo, são classificados como intrapessoais, pois não dependem do observador (NUNES, 2016).

Nesse sentido, técnicas de envelhecimento facial baseadas em deep learning podem ser aplicadas nas imagens para reproduzir um envelhecimento no indivíduo. O objetivo disso é aumentar as amostras da determinada pessoa com o objetivo de identificá-la mesmo se ela estiver mais velha. Um exemplo desse tipo de técnica pode ser visto na Figura 11 onde a foto de um jogador de futebol foi envelhecida usando-se um aplicativo chamado FaceApp que foi desenvolvido na Rússia. Esse aplicativo utiliza redes neurais convolucionais para analisar milhares de imagens aprendendo aspectos específicos de rostos envelhecidos como rugas e aplicando essas características nas imagens, dando a elas um aspecto envelhecido.



Figura 11 – Envelhecimento de um famoso jogador de futebol produzidas por um aplicativo chamado FaceApp.

Outro fator determinante no reconhecimento facial de pessoas é a posição em que a face se encontra na imagem. Essas posições podem variar de muitas formas, lateralmente à esquerda ou à direita, frontalmente para cima ou para baixo, dentre outras posições. Dessa forma, as imagens que compõem o banco de dados devem ser filtradas de forma a se obter o melhor enquadramento dos rostos. Isso é importante porque cada indivíduo tem características particulares e que os diferencia de outros.

Essas características podem ser analisadas com a localização dos chamados "pontos

fiduciais", que são definidos por (RIBEIRO et al., 2012) como "pontos de controle sobre um objeto que definem regiões características com propriedades interessantes à detecção. No caso da face, características faciais". Como pode ser visto na Figura 12, na face humana, esses pontos dão a localização dos olhos, nariz, boca, dentre outros.

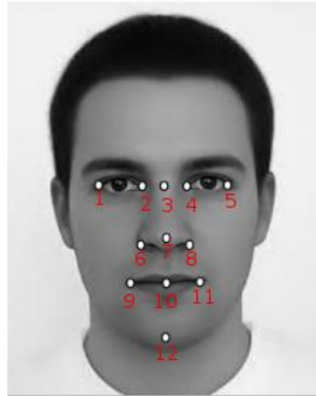


Figura 12 – Face com os pontos fiduciais marcados. Fonte: (AQUINO et al., 2017)

Além desses fatores, a detecção e o reconhecimento facial através de vídeos configura-se como dificultosa pois os níveis de contraste da imagem, bem como os de oclusão podem ser muito elevados, o que implica em falhas no sistema de reconhecimento, pois ele não terá sucesso em extrair as características das pessoas de forma eficaz. Esse problema não se configura por exemplo quando o sistema de aquisição de imagens para o banco de dados é controlado, ou seja, quando todas as variáveis problemáticas como a intensidade de luz, a oclusão, dentre outras, são controladas. Porém, esse é um cenário ideal e que não reproduz fidedignamente a realidade de um sistema de reconhecimento de pessoas.

Além disso, em sistemas de video monitoramento, a grande quantidade de dados para serem processados em um curto intervalo de tempo e com uma boa taxa de acerto dificulta ainda mais o reconhecimento (QUIRITA, 2014).

É comum as pessoas usarem acessórios como óculos, bonés ou deixarem seus cabelos e barbas crescerem, como pode ser visto na Figura 8. Esses fatores também são prejudiciais para um sistema de reconhecimento de pessoas que tem uma deficiência no número de imagens em seu banco de dados, pelo simples fato de o sistema não ter sido treinado com um mesmo indivíduo usando tais acessórios ou com suas características faciais modificadas.

2.3.2 Qualidade da Detecção da Face

Outro fator fundamental para garantir um reconhecimento facial efetivo é garantir que a detecção da face realizada tenha uma qualidade adequada. A biblioteca *dlib*, amplamente utilizada no ambiente acadêmico e de pesquisas envolvendo visão computacional, apresenta um método que retorna a qualidade das detecções de faces realizadas.

O algoritmo de detecção utilizado é construído a partir do método HOG, conforme apresentado na seção 2.2.1. O HOG foi treinado combinado com um classificador linear gerando uma pirâmide da imagem e um esquema de detecção janelada. Este detector é bem geral e capaz de detectar objetos semi rígidos, além de faces humanas, sendo composto por 5 filtros: face frontal, face lateral esquerda, face lateral direita, face frontal rotacionada para esquerda e face frontal rotacionada para direita. Assim, é possível, além de detectar as faces presentes em uma imagem, atribuir uma pontuação para a qualidade da detecção e determinar o tipo de face dentre os 5 filtros treinados. A pontuação e o tipo de face são importantes para garantir que o reconhecimento facial seja realizado apenas em faces que apresentem detecção de boa qualidade e tipo de face adequado (faces frontais são mais suscetíveis a apresentarem uma boa qualidade e portanto são mais adequadas para o reconhecimento).

2.3.3 Análise de Performance

O modelo FaceNet baseia-se no processo de *matching*, que basicamente consiste um cálculo de distâncias no espaço dimensional transformado da imagem da face e em uma busca pela menor distância aos vetores de interesse, também chamado de kNN (SCHROFF; KALENICHENKO; PHILBIN, 2015; PETERSON, 2009). Em sua proposição original os autores não propuseram uma forma de otimizar o *matching*. Sabe-se que este é um problema clássico de machine learning, comumente chamado de "*curse of dimensionality*", no qual o custo computacional de busca cresce a medida que a quantidade de dados aumenta.

Assumindo que no âmbito do reconhecimento facial, cada face possuirá seu próprio *encoding* gerado pela FaceNet, um cenário real de uma região metropolitana de uma cidade com 1 milhão de habitantes poderá inviabilizar a busca e tempo real. Em seu estudo, (KALAKUNTLA, 2017) elenca as problemáticas do kNN em grandes conjuntos de dados, como tempo de processamento e consumo de memória. O autor realiza um comparativo entre técnicas de otimização do kNN em grandes conjuntos de dados na CPU e GPU.

No trabalho de (KALAKUNTLA, 2017) a busca de um kNN em um conjunto de dados de 128 atributos com 200 mil amostras é 25 vezes mais rápido em um código escrito com CUDA para GPU do que um código paralelizado em CPU. O autor elenca um tempo de 84 segundos contra 2100 segundos, respectivamente.

As principais métricas encontradas na literatura para avaliação desse tipo de sistema relativas a tempo de processamento são: tempo para cálculo de distâncias; tempo para procura na base (*matching*). As métricas relativas a assertividade do sistema são: acurácia, matriz de confusão, taxa de falsos positivos e taxas de falso negativos.

Há também métricas relativa a qualidade mínima das imagens que devem serem usa-

das para reconhecimento facial (ABAZA; HARRISON; BOURLAI, 2012), como contraste, brilho, focus e iluminação.

2.3.3.1 Técnicas de Análise de Tempo e Performance dos Métodos de Busca em Grandes Bases

O problema de busca em grandes bases de dados é ilustrado na Figura 13. A ideia reside na necessidade de encontrar rostos similares, dada uma face de entrada. Os autores elencam que a principal problemática em reconhecimento facial é a incapacidade de escalar algoritmos para uma grande região de busca (centenas de milhões de faces). A algoritmo proposto por (WANG; OTTO; JAIN, 2016) baseia-se em aproximações dos candidatos mais prováveis, dada uma face de entrada e re-ordenação por similaridade da lista de possíveis indivíduos (YI et al., 2014; JEGOU; DOUZE; SCHMID, 2010).

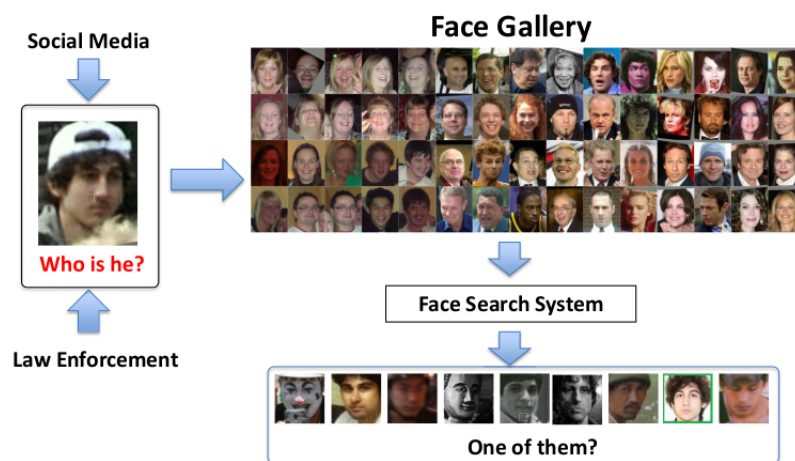


Figura 13 – Exemplo de problema de busca em uma grande base de dados. Fonte: (WANG; OTTO; JAIN, 2016)

Há métricas bastantes específicas para o problema de reconhecimento facial, que vão além das métricas comumente usadas para avaliação de sistemas de sistemas de reconhecimento de padrões (WANG; OTTO; JAIN, 2016).

No contexto de reconhecimento facial, **precision** é a fração do conjunto de busca que contém faces similares e **recall** é a fração entre todas as faces similares dada uma face de entrada. A definição de similaridade pode variar de acordo com as equações estabelecidas para o cálculo, bem como a escolha de um limiar mínimo para considerar que duas faces são da mesma pessoa.

Um sistema que preza por elevado *recall* é capaz de retornar um elevado número de possíveis candidatos. Por sua vez, um elevado número de candidatos propicia em baixa precisão. Portanto, (WANG; OTTO; JAIN, 2016) propõe utilizar uma métrica agregada,

chamada de *mean Average Precision* (mAP), já utilizado para avaliar algoritmos de busca (JEGOU; DOUZE; SCHMID, 2010).

mAP é definido da seguinte forma: dada n possíveis de interesse $Q = \{\vec{x}_q^1, \vec{x}_q^2, \dots, \vec{x}_q^n\}$ e um conjunto de N imagens de faces. Dessa forma (WANG; OTTO; JAIN, 2016) define mAP como,

$$mAP(Q) = \text{mean}(\text{avg}(P(\vec{x}_q^i)), i = 1, 2, \dots, n,$$

na qual $P(\vec{x}_q^i)$ é a combinação entre **precision** e **recall** de acordo com o elencado pelo autor.

Para a análise do tempo deve-se considerar o tempo médio de busca de uma face em uma determinada base de dados. No trabalho de (WANG; OTTO; JAIN, 2016) são reportados os seguintes resultados: $mAP = 0.40$ e 6.88 segundos para a busca em de uma face em uma base de dados de 80 milhões de faces.

3 Metodologia

Neste capítulo, serão evidenciados os passos realizados para a avaliação da FaceNet no problema de reconhecimento facial para 4 bases de fotos públicas e uma base privada com suas respectivas características. Também será mostrado como é feito o processo de oclusão das faces.

3.1 Fluxo de Análise do Problema de Reconhecimento Facial com a FaceNet

Serão feitas duas etapas de avaliação da FaceNet:

1. Aplicação da FaceNet na base privada, avaliando o problema de faces com e sem oclusão, validando a rede com pesos pré-treinados aplicada a uma base robusta em termos de quantidade. O *matching* é realizado utilizando a distância euclidiana. É necessário um trabalho de análise explanatória devido ao desconhecimento prévio da qualidade da base, principalmente focado na detecção facial.
2. Aplicação da FaceNet nas bases públicas, avaliando o problema de faces com oclusão por máscara, o *matching* sendo feito pelas distâncias Euclidiana, Manhattan e método dos cossenos com posterior comparação de resultados.

É proposta uma aplicação com finalidade para captura e processamento de faces frontais, procurando o rosto em uma base pré-cadastrada. Após verificação, é exibida uma lista de possíveis candidatos na base, com um score de classificação. O fluxo está na Figura 14

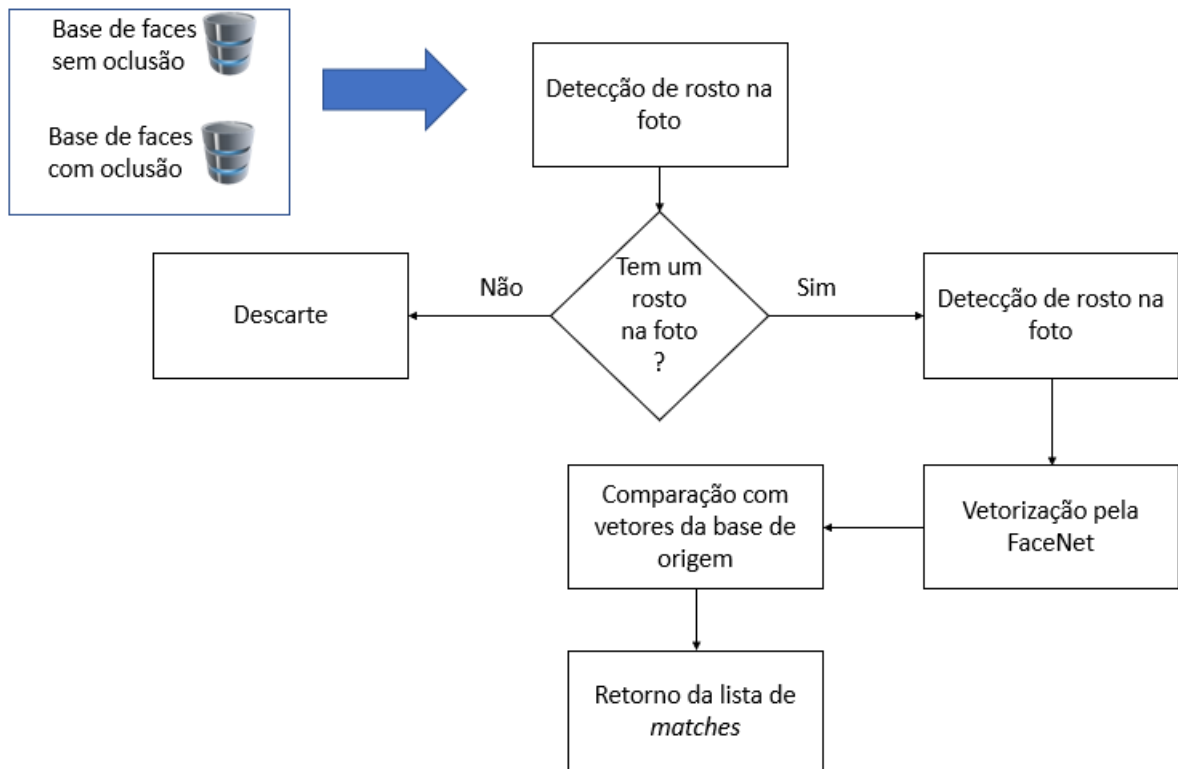


Figura 14 – Fluxo do processo de análise do reconhecimento facial. Fonte: Elaborado pelo autor.

3.2 Extração de Características e Medição de Similaridade Vetorial

A extração de características leva em consideração pontos da faces relevantes obtidos pela detecção facial são identificados através do método HoG, com o uso da biblioteca *dlib*¹. Assim, a etapa de vetorização pode ser exemplificada pela Figura 15.

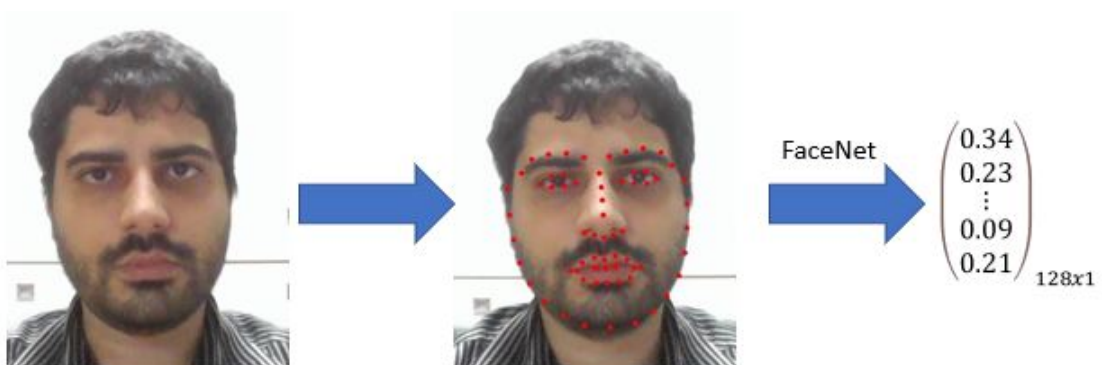


Figura 15 – Fluxo de vetorização da imagem de face

Considerando que o resultado pós-processamento da FaceNet é um vetor de dimensão 128 com números reais, propõe-se a aplicação de métricas cujo objetivo é entender

¹ Disponível em: <<http://dlib.net/>>

quais vetores tem um maior grau de similaridade entre si. São utilizadas no trabalho as distâncias Manhattan, Euclidiana e dos Cossenos.

A distância Manhattan se trata do somatório da diferença entre as respectivas coordenadas em um dado espaço vetorial, ou seja, dados vetores $x=(x_1,x_2,...,x_n)$ e $y=(y_1,y_2,...,y_n)$ a equação da distância Manhattan é representada por :

$$d_{xy} = \sum_{i=1}^n |x_i - y_i| \quad (3.1)$$

É mostrado em Malkauthekar (2013) que a distância Manhatthan pode obter resultados de até 97% de acurácia aplicada ao problema de reconhecimento facial quando comparada a outras métricas.

Já a distância Euclidiana pode ser definida como uma medida de similaridade, uma técnica normalmente atrelada à algoritmos de visão computacional pois se aplica melhor a dados não padronizados, como amostras sem tratamento de adaptação de escala, dessa forma, pode-se dizer que seu resultado final ser insensível à *outliers* (CARDARILLI et al., 2019). Por sua vez, essa medida de similaridade normalmente não é recomendada em casos na qual há diferença de escala ou de grandezas entre as dimensões. A equação da distância para dois pontos A e B em um espaço n dimensional pode ser vista na equação 3.2

$$d_{AB} = \sqrt{\sum_{i=1}^n (A_i - B_i)^2} \quad (3.2)$$

Por fim, a distância de cossenos é a medida do ângulo formado entre dois vetores em um espaço vetorial, representada pela figura 16. O cálculo é visto na equação 3.3

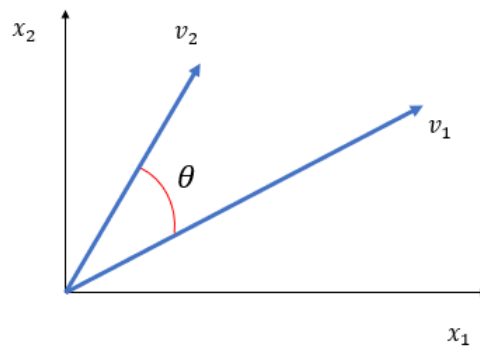


Figura 16 – Representação gráfica em um espaço 2-D da medida de distância dos cossenos.
Fonte: Elaborado pelo autor

$$\cos(\theta) = \frac{v_1 \cdot v_2}{|v_1| |v_2|} \quad (3.3)$$

3.2.1 Formato de Matching

A abordagem escolhida para fazer os cálculos de matching foi baseada no algoritmo NGT (do inglês *Neighborhood Graph and Tree*), no qual há uma busca aproximada dos vizinhos mais próximos. O principal benefício com relação a outros métodos, por exemplo, é a redução do tempo de classificação dos semelhantes, uma vez que não é feita uma busca exaustiva em toda base de dados, além de uma menor exigência de poder computacional, como visto em Iwasaki e Miyazaki (2018)

3.3 Bases de Dados

Serão trabalhadas uma base privada e bases públicas de imagens frontais de faces. O intuito da base privada é verificar a eficácia da FaceNet pré-treinada em uma base robusta em termos de quantidade. No caso das bases públicas, foram escolhidas bases já catalogadas em outros trabalhos relacionados a mesma temática desse trabalho. A análise do efeito de oclusão por máscara será feito tanto nas bases públicas quanto na base de imagens privada.

3.3.1 Base Privada de Imagens de Rostos

Nesta seção apresentamos uma análise preliminar da base de imagens privadas utilizadas. Por se tratar de base de dados real, a exibição dela nesse trabalho envolve questões de autorização dos titulares das fotos para o cumprimento de legislação vigente. O padrão de fotos está ilustrado e exemplificado conforme a Figura 17, uma vez que não pôde-se utilizar exemplos reais de imagens da base. A base total tem um conjunto de fotos de aproximadamente 7 milhões de imagens de faces.



Figura 17 – Ilustração da base de dados privada, representando os diferentes tipos de qualidade de fotos presentes. Fonte : Elaborado pelo autor.

Embora haja uma padrão requerido captura das fotos, a base foi construída pelo registro de faces em postos de coleta em locais diferentes, tendo o processo conduzido tanto por variados profissionais e equipamentos. Esse último ponto provocou uma divergência na qualidade das imagens ao longo do tempo, uma vez que registros mais antigos tendem

a ter uma qualidade menor que registros mais novos, devido ao emprego de equipamentos mais modernos para as fotos. Tendo isso em vista, foram trabalhadas diferentes amostras dessa base privada para que se tivesse um melhor entendimento do real estado das fotos, principalmente relativa ao que diz respeito ao problema da detecção de faces. Inicialmente, foi separada uma amostra composta por aproximadamente 15 GB de dados, totalizando 20.254 amostras. Na Tabela 2 apresentou-se uma visão das estatísticas gerais deste conjunto. Destacam-se o total de amostras por posto de coleta, com aproximadamente 100 casos para cada posto, o tamanho médio de espaço em disco consumido pelas imagens que é 858.9KB, além das dimensões espaciais médias da imagem com 729 de largura por 957 de altura. Identificou-se que em todos os casos foi utilizado os tipos de compressão: JPEG e Zip, para este último tipo, identificou-se que as imagens estavam erroneamente rotuladas com extensão *.jpg*, para visualização, o nome foi ajustado para a extensão *.png*.

Tabela 2 – Visão geral da base de dados de imagens faciais.

Posto de Coleta	
Total de amostras	20.254
Total de postos	204
Amostras por Posto	≈ 100
Propriedade dos Arquivos	
Tamanho Médio	858.944 KB
Tamanho Máximo	1 MB
Tamanho Mínimo	83 KB
Espaço de cor	sRGB
Compressão	JPEG (5536), Zip (14718)
Dimensões espaciais Médias	729 x 957
Bits por Pixel	8 bits x 3 canais

Em seguida, uma nova amostra, dessa vez com 427.603 imagens com faces frontais também foi analisada. Nesse sentido, utilizando o algoritmo apresentado na Subseção 4.1.1, avaliou-se nessa base a qualidade e a quantidade de faces encontradas por imagem, além de associar esses indicadores com a resolução a fim de se detalhar ainda mais o perfil das imagens.

Há uma concentração de aproximadamente 33% com imagens de 767x1028 e outra concentração de 25% de imagens com dimensões de 472x630. As outras fatias distribuídas nesse gráficos são de variações próximas a de 472x630 e, portanto, serão interpretadas como se fossem de 472x630 devido seu grau de aproximação menor ou igual a 10% desse mesmo valor.

Por fim, uma outra amostra da base privada, com um total de 119.485 amostras também foi analisada. Este mesmo total foi obtido quando analisou-se o número de arquivos em que foram detectadas faces, independentemente do score de detecção, o que indica que toda a base utilizada apresenta pelo menos uma ocorrência de detecção.

3.3.2 Base Pública de Imagens de Rostos

São utilizados 4 diferentes bases de dados. A base AR (MARTINEZ, 1998), na Figura 18, contém imagens de 126 pessoas, das quais 70 são homens e 56 são mulheres, que foram obtidas a partir de 2 sessões fotográficas com uma diferença de 14 dias entre elas, com cada modelo tirando pelo menos 13 fotos por sessão com diferentes expressões faciais, condições de luminosidade e oclusão. A base LFW (HUANG et al., 2008a), Figura 19, tem 13.233 imagens de 5749 pessoas retiradas de diversas fontes online, sendo em sua maior parte fotos de personalidades. Apenas 1680 pessoas tem duas ou mais imagens. Essa base apresenta faces em variadas formas, com diferentes oclusões, angulações, ruído e posições. A base Extended YaleB (GEORGHIADES; BELHUMEUR; KRIEGMAN, 2001), Figura 20 tem 16.128 imagens de 28 pessoas em 9 diferentes posições com 64 variações de luminosidade. Já a base CelebFaces (LIU et al., 2015), Figura 21 possui 202.599 imagens de 10.177 pessoas, nas quais situações de variação de pose e formato de rosto, presença ou não de pelos faciais e uso de apetrechos como óculos e bonés podem ser encontradas .



Figura 18 – AR dataset



Figura 19 – LFW dataset



Figura 20 – Extended Yale Dataset



Figura 21 – CelebFaces-A Dataset

3.4 Experimentos de Reconhecimento Facial com Oclusão por Máscaras Faciais

A oclusão em faces pode ocorrer devido a inúmeros fatores, seja pelo uso de acessórios como óculos, diferentes cortes de cabelo que venham a obstruir parte do rosto. Contudo, durante o ano de 2020 devido a pandemia de Covid-19, vários governos adotaram recomendações e imposições de novos hábitos, principalmente no que diz respeito a medidas sanitárias. Uma das mais populares é justamente o uso de máscaras faciais como forma de se evitar o contágio da doença. Com isso, os desafios para os sistemas de reconhecimento facial até então desenvolvidos aumentaram. Nessa sessão, é avaliado o efeito da oclusão causada pelas máscaras, para as mesmas bases públicas e privadas. São realizados dois tipos de procedimentos para avaliação. O primeiro deles é através da geração artificial de máscaras, através da técnica *MaskTheFace* descrita em Anwar e Raychowdhury (2020) e exemplificada na Figura 22.

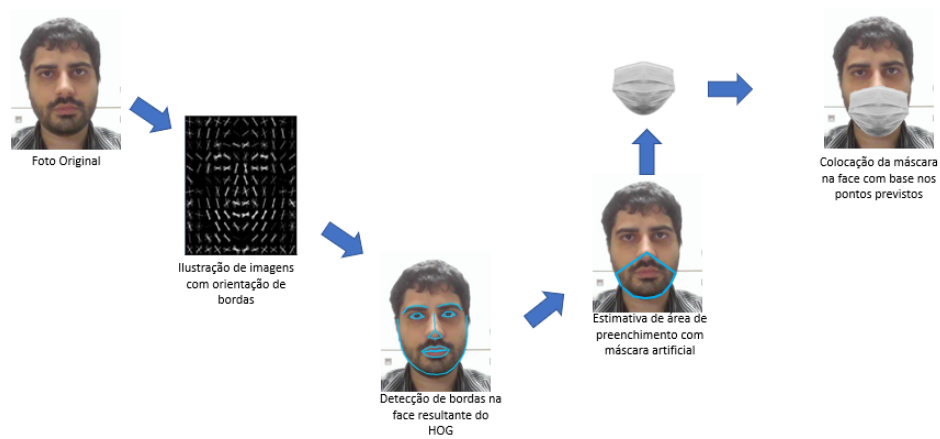


Figura 22 – Fluxo de mascaramento artificial de faces através da técnica *MaskTheFace*.
Fonte: Elaborado pelo Autor

A partir da detecção facial com o método HOG, conforme descrito em Chen et al. (2014), coordenadas geométricas dos rostos são marcadas de forma a explicitar quais parte devem estar cobertas por máscaras artificiais.

4 Resultados e Discussões

O capítulo vai falar separadamente dos resultados obtidos com os procedimentos indicados na seção 3.1 ao se trabalhar tanto com a base privada quanto as bases públicas. Inicialmente, será descrita a análise sobre a base privada e seus subconjuntos. Por fim, será feita a análise do fluxo proposto sobre as bases públicas.

4.1 Base de dados privadas

Dedicou-se um tempo inicial a entender a composição da base, seu padrão de arquivos bem como o processo de coleta. Nesse contexto, é dada uma prioridade para avaliar primeiro a capacidade de se identificar uma face dada um rosto que compõe essa base. Apresentadas as características e escolha de parâmetros para caracterização dessa base, por fim submete-se ao processo de *matching* para validação da aplicação da FaceNet.

4.1.1 Score de Detecção de Faces

Os scores apresentados nesta análise correspondem ao resultado das detecções por meio de padrões previamente apresentados a um classificador linear com características identificadas por meio do algoritmo HOG. Conforme mencionado na seção 2.3.2, o HOG foi treinado combinado com um classificador linear gerando uma pirâmide da imagem e um esquema de detecção janelada. O valor do Score, portanto, é obtido com base em outras duas métricas: S_{score} que corresponde ao valor retornado pelo janelamento para a detecção de objetos que correspondem aos padrões apresentados ao classificador, e W_{score} que corresponde ao valor do limiar de detecção armazenado no peso que retornou maior valor para a detecção. Assim, o valor *Score* pode ser interpretado pela seguinte equação 4.1:

$$Score = S_{score} - W_{score} \quad (4.1)$$

De forma que se estabelecermos um limiar igual a zero, então o *Score* dependerá somente do valor de S_{score} (ou seja da detecção do janelamento), por outro lado, se deixarmos o limiar mais permissivo, isto é, um intervalo aberto tendendo ao conjunto $(-\infty, +\infty)$ então mais detecções falsas serão apresentadas, justamente quando o S_{score} é muito pequeno o W_{score} é negativo, o retorno será então cada vez mais negativo quanto menor for o valor de S_{score} , o que indica o grau de confiança da detecção.

4.1.2 Resultados Preliminares

Uma das análises realizada sobre o conjunto diz respeito à qualidade da detecção de faces nas amostras. Ela afeta diretamente o reconhecimento efetivo entre uma amostra candidata e uma amostra registrada em base de dados. Uma detecção de má qualidade pode representar uma falsa detecção de face, contribuindo para o aumento no número de falsos positivos apresentados pelo sistema.

Neste aspecto, todo o conjunto foi analisado de acordo com o score de detecção da face em cada uma das amostras. Os casos cujo score de detecção foi inferior ou igual ao limiar de corte foram desconsiderados afim de evitar a contagem de falsas detecções. Em seguida, avaliamos os postos de coleta levantando estatísticas descritivas com base no score de detecção. Uma análise inicial sobre o conjunto de dados produziu um levantamento de estatísticas básicas quanto as amostras por posto de coleta. Neste sentido, observou-se que há uma distribuição aproximadamente balanceada entre os 204 postos de coleta, isto é, os postos possuem aproximadamente 100 imagens cada nesse conjunto.

Apresentamos na Tabela 3 uma listagem dos 20 postos de coleta cujo score de detecção médio apresentou os maiores resultados. Nesta tabela, são apresentados os scores mínimos e máximos para cada posto. A listagem é apresentada em ordem decrescente a partir do score médio. Observou-se que os postos de coleta A,B,e C apresentaram melhores casos de detecção, em média. Por outro lado, a Tabela 4 apresenta a listagem dos 20 postos de coleta com os menores scores de detecção apresentados.

4.1.3 Análises Estatísticas nas Bases de Dados

Nesta seção serão contempladas diferentes análises de duas bases privadas de faces. Alguns pontos de análise foram prioritariamente elencados a fim de avaliar o melhor cenário com que o reconhecimento facial teria melhor êxito, ou seja, para o caso da análise de uma base de faces quais são as características em uma imagem mais relevantes para que a extração das características seja feita de forma a ter um poder discriminatório mais efetivo. Essa questão resolvida trará como utilidade a informação de como deve ser montada uma próxima base facial para obtenção de melhores resultados.

4.1.3.1 Análise de Detecção Facial da Base Privada 1

Outro ponto a ser levantado é se seria interessante associar o score obtido com a resolução da imagem e descobrir se há uma resolução que implique em um score mais favorável ao problema de reconhecimento facial. Assim, a Figura 23 retrata a frequência com que os scores foram encontrados para o referido *dataset*.

É clara a percepção de que há uma quantidade de scores satisfatórios centralizados no valor 2, o que sugere que esse *dataset* possui um bom acervo de imagens com boa

Tabela 3 – Lista dos 20 postos de coleta que apresentaram os maiores scores de detecção, em média. A listagem está ordenada em forma decrescente pelo Score Médio de detecção da face.

Posto de Coleta	Score Médio	Score Mínimo	Score Máximo
Posto A	2,5312	1,3589	3,6915
Posto B	2,3216	0,3696	3,1704
Posto C	2,3208	0,6955	3,2582
Posto D	2,3147	0,9549	3,5348
Posto E	2,2987	0,5989	3,5709
Posto F	2,2950	1,2977	3,3035
Posto G	2,2841	1,0957	3,4964
Posto H	2,2822	0,9644	3,3822
Posto I	2,2807	1,1213	3,3454
Posto J	2,2783	1,1113	3,0491
Posto K	2,2731	1,0282	3,4061
Posto L	2,2588	0,8438	3,4143
Posto M	2,2587	0,7291	3,2258
Posto N	2,2584	1,4066	3,1566
Posto O	2,2575	0,8979	3,1576
Posto P	2,2571	1,0747	3,3503
Posto Q	2,2472	0,7769	3,3997
Posto R	2,2269	0,7528	3,1656
Posto S	2,2152	0,8642	3,2719
Posto T	2,2144	0,9985	3,2763

qualidade.

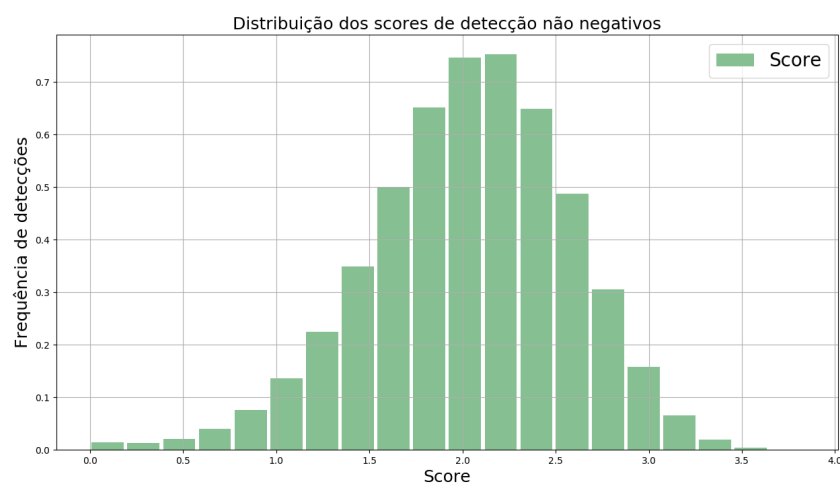


Figura 23 – Histograma dos Scores de todas as detecções no conjunto de dados. Esta figura ilustra o total de Scores obtidos com filtragem de Scores negativos.

A Figura 24 destaca a relação entre a resolução da imagem e o score obtido nessa resolução. Para isso, utilizou-se da média dos scores para cada dimensão a fim de obter

Tabela 4 – Lista dos 20 postos de coleta que apresentaram os menores scores de detecção, em média. A listagem está ordenada em forma crescente pelo Score Médio de detecção da face.

Posto de Coleta	Score Médio	Score Mínimo	Score Máximo
POSTO I	1,2840	0,3617	2,0886
POSTO II	1,4918	0,3001	2,8837
POSTO III	1,5426	0,4741	3,1512
POSTO IV	1,5694	0,4964	2,2949
POSTO V	1,5721	0,2628	3,0022
POSTO VI	1,5763	0,6094	2,6972
POSTO VII	1,6131	0,2748	3,0735
POSTO VIII	1,6536	0,2600	2,7589
POSTO IX	1,6682	0,4836	2,7725
POSTO X	1,6762	0,5806	2,5904
POSTO XI	1,6947	0,4219	2,6798
POSTO XII	1,6985	0,3733	2,8035
POSTO XIII	1,7059	0,4862	2,8567
POSTO XIV	1,7085	0,7998	2,8495
POSTO XV	1,7143	0,5618	2,7418
POSTO XVI	1,7144	0,4271	2,9517
POSTO XVII	1,7287	0,3060	2,7862
POSTO XVIII	1,7432	0,6788	2,9849
POSTO XIX	1,7505	0,4035	2,9891
POSTO XX	1,7512	0,5860	2,6980

um valor de referência. Observando a distribuição mais relevante do *dataset* (resoluções de 767x1028 e 472x630), nota-se que as imagens com dimensões 767x1028 tiveram um valor médio de score em 2.25 enquanto que as imagens de 472x630 tiveram score mais próximos de 2. Esse resultado corrobora com o gráfico ilustrado com a Figura 23 onde mostra a maior concentração do dado em torno desse valor.

Analisando o score por faixas de operação, a Figura 25 ilustra *boxplots* com diferentes filtros. Cada filtro é interpretado como um limiar para indicar como está a distribuição dos scores a partir de um nível 'T'. Observe que nos cenários onde $T < 2.5$, a concentração dos valores de score está entre 2 e 2.5, indicando novamente que existe uma boa quantidade de imagens com qualidades representativas para a aplicação de reconhecimento facial, e que seriam portanto imagens de referência para essa aplicação.

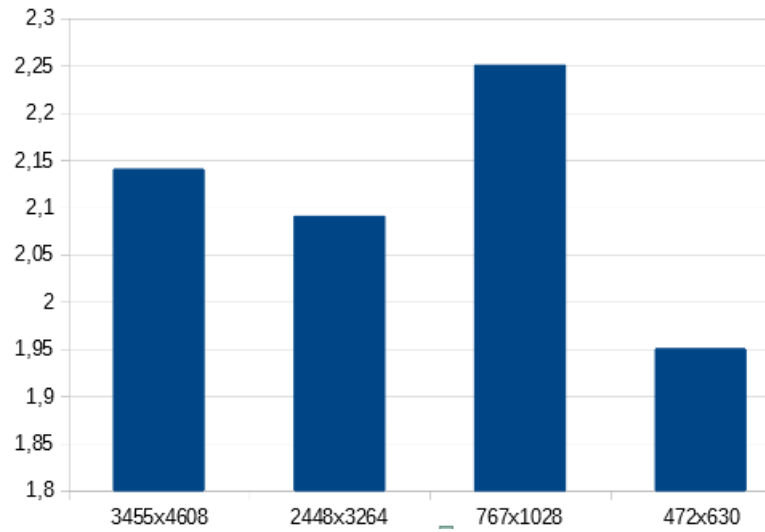


Figura 24 – Relação entre resolução da imagem e score obtido. O valor do score utilizado é o valor médio por dimensão.

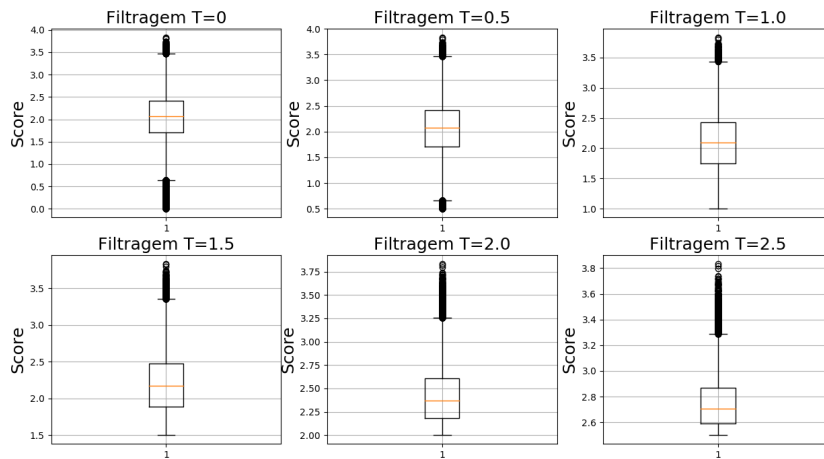


Figura 25 – Distribuição do conjunto de imagem quanto aos Scores de detecção por filtragem com $\text{Score} \geq T$.

4.1.3.2 Análise de Detecção Facial da Base Privada 2

O total de arquivos processados neste experimento somam um conjunto com 119485 amostras. Este mesmo total foi obtido quando analisamos o número de arquivos em que foram detectadas alguma face, independentemente do score de detecção, indicando que a base utilizada apresenta pelo menos uma ocorrência de detecção.

Neste contexto, vale destacar que nem todas as faces detectadas poderiam ser consideradas como faces válidas em um processo de reconhecimento facial. Para isto, avalia-se o score de detecção da face, uma vez que regiões da imagem também poderão induzir a detecções de faces falsas. A Figura 26 apresenta a distribuição em relação ao total de detecções nesta base, com a quantidade por Score. Nota-se que há uma alta quantidade

de faces detectadas com Score abaixo de zero, o que indica falsas detecções. Estas falsas detecções são caracterizadas por regiões da imagem que foram detectadas como possível face, mas que de fato não representam faces reais.

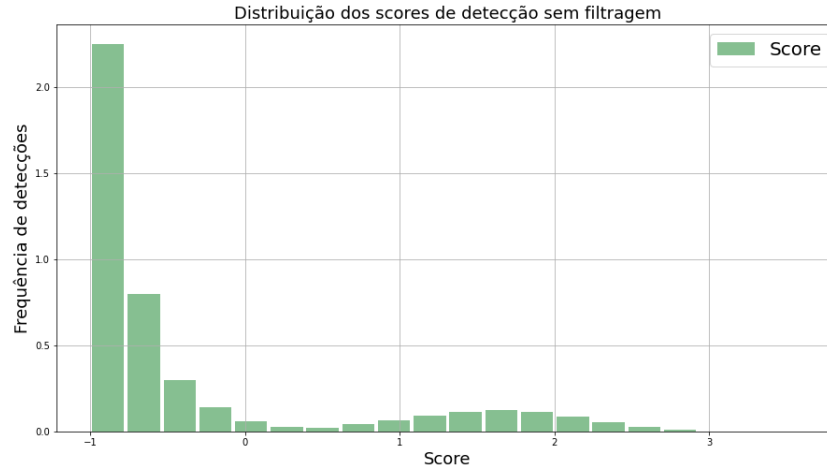


Figura 26 – Histograma dos Scores de todas as detecções no conjunto de dados. Esta figura ilustra o total de Scores obtidos sem filtragem de Score mínimo.

Neste sentido, outra análise possível é avaliar a distribuição do conjunto de dados com base em limiares de detecções afim de avaliar o conjunto de imagens sob a perspectiva da filtragem de detecções com base no Score. A Figura 27 apresenta a distribuição deste conjunto de dados quando aplicada filtragens com limitares T mínimos do Score de detecção aceitável. Observa-se que existem imagens com detecções de Score que superam o patamar de 3, o que representa excelentes detecções. Por outro, considerando Scores superiores a zero (Score T , $T=0$), observa-se que a mediana do conjunto está em torno de 1.5, indicando que pelo menos a metade do conjunto de faces representam detecções aceitáveis. Esta distribuição é ilustrada na Figura 28.

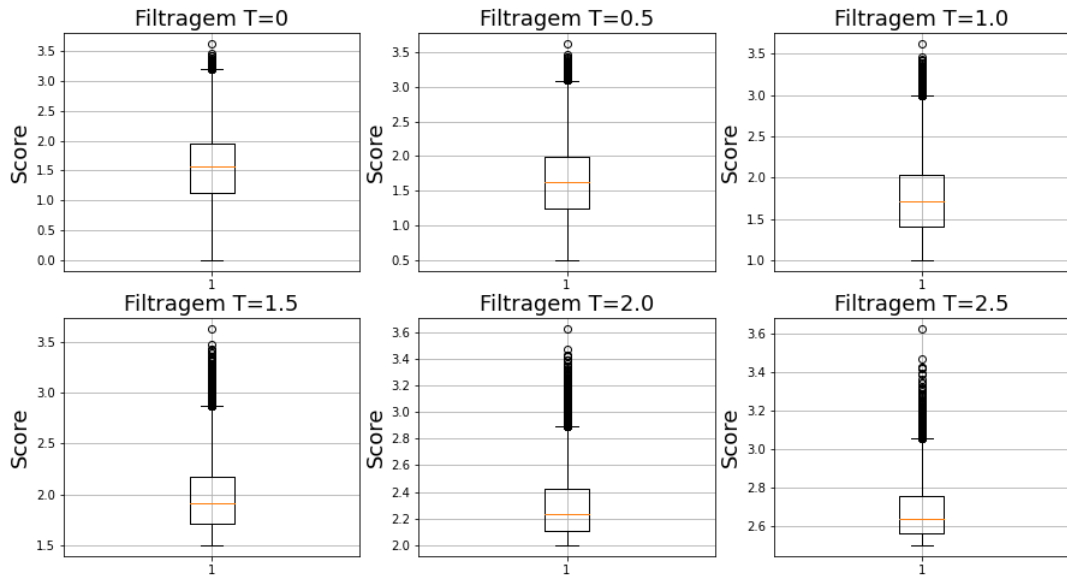


Figura 27 – Distribuição do conjunto de imagem quanto aos Scores de detecção por filtragem com $\text{Score} \geq T$.

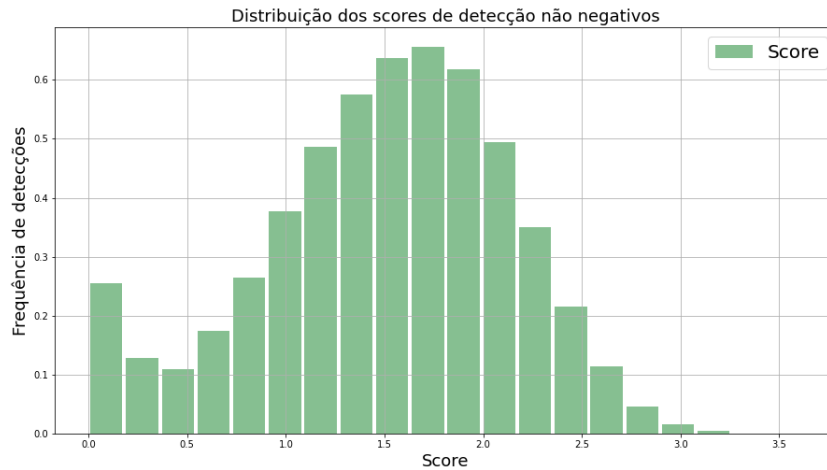


Figura 28 – Histograma dos Scores de todas as detecções no conjunto de dados. Esta figura ilustra o total de Scores obtidos com filtragem de Scores negativos.

Já a Tabela 5 apresenta uma variação do limiar T iniciando com valor 0 até 2.5, com passo de 0.5. Apresenta, portanto, os Scores mínimos e máximos, assim como médio de cada partição filtrada pelo limiar. Destaca-se que do total de 683.779 detecções, o que inclui também uma grande quantidade de falsas detecções, cerca de 18.70% representaram faces com Score superior a zero, reforçando ainda a necessidade de se aplicar uma filtragem, possibilitando uma redução de esforço computacional em aproximadamente 80% de todas as possíveis faces detectadas.

Além da análise de Score de detecção, realizou-se também um estudo quanto a proporção da face detectada em relação a área da imagem original. A Figura 27 apresenta a distribuição das amostras detectadas com base no percentual da imagem ocupada pela face.

Tabela 5 – Estatísticas quanto ao total de detecções quando aplicada a filtragem por limiar T do Score de detecção.

T	Total de Detecções	Médio	Mínimo	Máximo	Mediano
0	127906 (18.70%)	1.499	5.582e-06	3.620	1.560
0.5	117170 (17.13%)	1.619	0.500	3.620	1.625
0.1	102141 (14.93%)	1.740	1.000	3.620	1.713
1.5	68846 (10.06%)	1.968	1.500	3.620	1.916
2.0	28114 (04.11%)	2.289	2.000	3.620	2.238
2.5	4952(00.72%)	2.680	2.500	3.620	2.636

Ainda de acordo com a Figura 27, observa-se que a grande maioria das detecções ocupam até 10% do total da imagem, o que pode auxiliar em futuras compressões da imagem afim de reduzir o espaço de armazenamento necessário, preservando a região de interesse em cada arquivo.

4.1.4 Partições da Base Privada para Aplicação da FaceNet

Nessa subseção será explicado o processo de divisão da base de dados privada afim de se obter uma amostra de consulta que será submetida aos testes de reconhecimento facial, tanto para o problema sem e com oclusão.

4.1.4.1 Dataset de Consulta

O *dataset* de consulta abordado é dividido em pastas, denominadas nesse documento de subjects/folders. Cada *subject/folder* possui três ou mais pedidos de um mesmo indivíduo. O termo pedido consiste no registro facial e informações sociais do indivíduo. O intuito de mapear diferentes pedidos de um mesmo indivíduo é utilizar um dos pedidos como amostra para consulta tendo referência de acerto na identificação facial o retorno de outras amostras daquele *subject/folder* contidas na *dataset* de pesquisa. O *dataset* criado continha 415.625 imagens, divididas em 100.686 subjects/folders, aproximadamente 4 imagens/pedidos por indivíduo. Diante da vasta quantidade de imagens inicialmente selecionadas, para a geração de estatísticas e melhor controle do *dataset* de consulta foi aplicado mapeamentos do *dataset*. Os mapeamentos permitiram a limpeza do *dataset*, mais precisamente a respectiva retirada de amostras que atrapalhariam na geração dos resultados e que deveriam ser removidas no contexto da aplicação final.

4.1.4.2 Imagens em Tons de Cinza ou Imagens Preto e Branco

Realizada a análise das imagens em tons de cinza ou imagens preto e branco (PB), foi identificado que as imagens em tons de cinza, em sua maioria, são provenientes de digitalizações e registros de capturas de outras imagens. Diante disto, essas imagens preto e branco são definidas frequentemente por uma baixa qualidade, ou seja, fora dos padrões

adotados para a reconhecimento facial treinado. O padrão adotado para reconhecimento facial prioriza aspectos como iluminação normalizada, centralização e alinhamento facial na imagem, saturação e contraste em níveis que melhor representem a realidade.

Esses padrões são importantes pois permitem ao algoritmo uma melhor identificação dos *landmarks* faciais no indivíduo e uma consequente extração mais eficiente das características faciais. Dito isto, as imagens preto e branco foram retiradas do *dataset* de consulta a partir do seguinte procedimento: leitura da imagem, divisão e comparação dos canais da imagem, uma vez que os canais são idênticos (a diferença entre eles é 0) a imagem é caracterizada como tons de cinza e assim removida. Caso o *subject/folder* tenha menos que 2 imagens toda a pasta é separada para a volume II do *dataset*. Há também casos de manchas sobrepondo a face, pouca iluminação, faces que ocupam uma área muito pequena da imagem e imagens com desfoque. Para todos esses exemplos, as respectivas fotos também foram retiradas da amostragem.

4.1.4.3 Idade

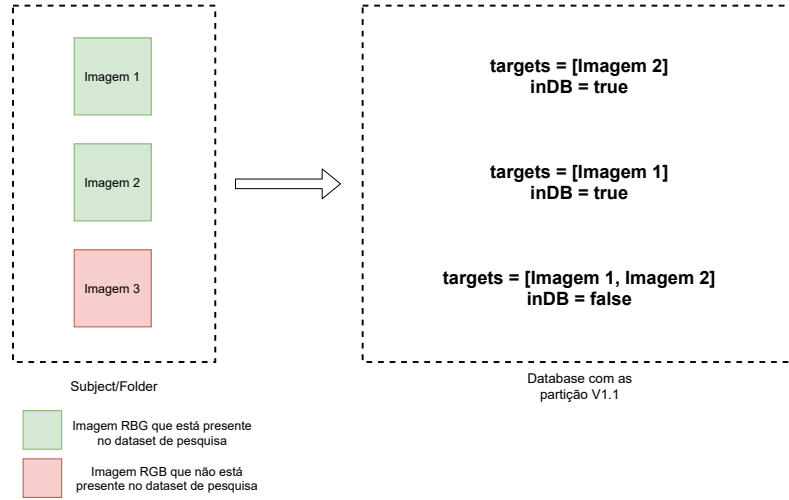
Uma das ações para homogeneizar a base de dados, até seguindo recomendação de boa utilização da própria FaceNet, removeu-se amostra cujo indivíduos eram de crianças ou adolescentes. Isso só foi possível pelo fato da idade de cada pessoa ser uma informação que estava atrelada ao cadastro da foto na base.

4.1.4.4 Qualidade de Detecção

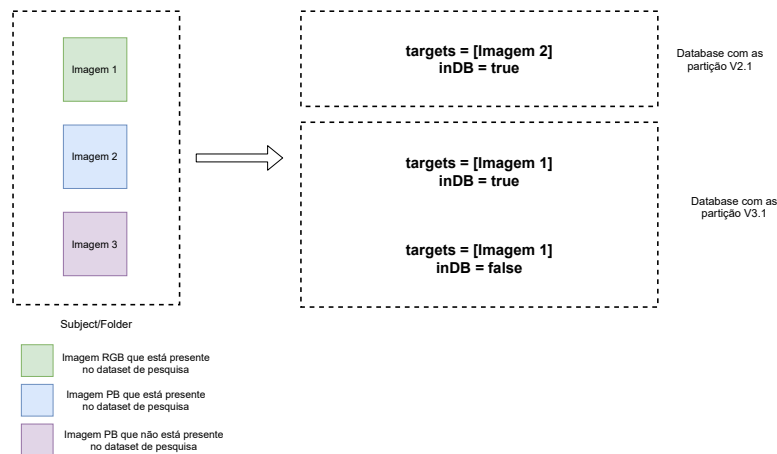
Aplicados os critérios 4.1.4.2 e 4.1.4.3, os *subject/folder* que ainda continham no mínimo 3 imagens foram submetidos à detecção e extração facial. Neste processo foi aplicado um limiar na qualidade da detecção facial da imagem. É feito a detecção da face e a partir disto indicada a qualidade daquela detecção, uma vez que não obteve nenhuma detecção com qualidade igual ou superior à 0.8 a imagem é descartada. Caso o *subject/folder* tenha menos que 2 imagens toda a pasta é separada para a volume II do *dataset*.

4.1.5 Partições

Partições são as versões dos *datasets* de consulta gerados. Cada versão do *dataset* de consulta possui um mapeamento diferente dos *targets*. *Targets* são os pedidos daquele *subject/folder* que estão presentes no *dataset* de pesquisa e que, caso retornado como um candidato pelo processo de identificação facial, é admitido como acerto. Diante disto, foram criadas três partições V1.1, V2.1, V3.1 e V-mask.

Figura 29 – Esquema de geração e organização da partição V1.1 com os *targets*.

A partição V1.1 consiste na aplicação sequencial de todos os critérios mencionados. De tal modo que foram geradas 13.322 amostras testáveis caracterizando o volume I do *dataset* de consulta. Uma amostra testável consiste em um pedido que contém no mínimo um *target* de acordo com a partição referente. As partições V2.1 e V3.1 consiste no uso volume II aplicando todos os critérios mencionados anteriormente com a adição de amostras em tons de cinza que passaram pelo critério de qualidade da detecção. Assim, as partições V2.1 e V3.1 visam a análise do comportamento do sistema de identificação facial nas atividades de buscar uma imagem colorida (RGB) a partir de uma imagem em tons de cinza e buscar uma imagem em tons de cinza a partir de uma imagem colorida. As partições V2.1 e V3.1 possuem 11.759 e 16.591 amostras testáveis, respectivamente.

Figura 30 – Esquema de geração e organização da partições V2.1 e V3.1 com os *targets*.

A partição V-mask deriva do volume I do *dataset* de consulta. Essa partição consiste no uso do algoritmo *MaskTheFace*, mencionado na seção 3.4, o qual permite aplicar máscaras cirúrgicas nas faces dos indivíduos presentes na imagem. Conforme já dito, o volume I é resultado da aplicação dos critérios citados exceto pela etapa filtragem de

qualidade da detecção. As imagens após a aplicação das máscaras passaram pelo processo de detecção e extração facial permitindo assim a geração da partição V-mask com a mesma estrutura de *target* da partição V1.1, mostrado na Figura 29. A partição V-mask contém 11.954 amostras testáveis.

4.2 Análise dos Resultados das Bases Públicas e Privadas.

A partir do banco de dados disponibilizado, uma análise preliminar das imagens foi realizada, possibilitando organizar e identificar as particularidades do conjunto de imagens. A partir das técnicas descritas no estudo, é possível aplicar uma técnica de detecção de faces que possibilita identificar *scores* de confiança da detecção.

Por outro lado, com base na análise por *score*, foi possível identificar amostras abaixo do limiar de corte e assim evitar falsas detecções. Além disso, estabelecer uma identificação dos postos de coleta com menores *scores* de detecção para que possam, eventualmente, se adaptarem em termos de tecnologia de aquisição de imagem, àqueles que apresentam melhores *scores*.

Neste sentido, o posto de coleta "Posto 1", por exemplo, apresentou os menores índices de ruído estimado além de bons *scores* de detecção de face, em média. O que indica ser um posto candidato a uma análise quanto aos equipamentos de aquisição de imagens, mas também quanto ao atual protocolo de captura utilizado, de forma a replicar nas outras unidades de coleta.

4.2.1 Partições da Base de Imagens Privada para Fins de *Matching*

A partir das imagens totais do base privada, optou-se por montar dois subconjuntos maiores que servirão como bases de comparação para faces *target*, a fim de se avaliar o efeito no resultado quando temos imagens de qualidades distintas na mesma base. Dessa forma, nomeiam-se dois *datasets*:

- Dataset 1: 7 milhões de imagens em tons de cinza e RGB.
- Dataset 2: 4.9 milhões de imagens, com exceção das imagens em tons de cinza.

4.2.2 Experimentos com o *dataset* 1

Tabela 6 – Resultados de acurácia obtidos pelo método NGT com os *datasets* de consulta V1.1, V2.1, V3.1 e V-mask usando como alvo o *dataset* com 7M de amostras.

Método de reconhecimento	<i>Dataset</i> consulta	Top-1	Top-3	Top-5	Top-10	Top-20	Top-50	Top-100
NGT	V1.1	0.61371	0.91022	0.92866	0.94341	0.95319	0.96409	0.97122
	V2.1	0.18739	0.30396	0.34310	0.39757	0.44690	0.51208	0.56107
	V3.1	0.10581	0.35733	0.40231	0.45631	0.50542	0.56568	0.60496
	V-mask	0.01405	0.02769	0.03698	0.05488	0.07562	0.10691	0.14179

4.2.3 Experimentos com o *dataset* 2

Tabela 7 – Resultados de acurácia obtidos pelo método NGT com os *datasets* de consulta V1.1, V2.1 e V-mask usando como alvo o *dataset* com 4.9M de amostras.

Método de reconhecimento	<i>Dataset</i> consulta	Top-1	Top-3	Top-5	Top-10	Top-20	Top-50	Top-100
NGT	V1.1	0.61756	0.91695	0.93587	0.95030	0.96160	0.97547	0.98317
	V2.1	0.27706	0.43722	0.48133	0.53717	0.59687	0.66855	0.72157
	V-mask	0.02242	0.04350	0.05822	0.08064	0.10473	0.15083	0.19634

Pode-se notar que somente a partição V1.1, cujas imagens tem melhor qualidade, a partir do top-3 os resultados de acurácia conseguem atingir patamares superiores a 90%. No caso do problema de máscaras, mesmo no top-100 o resultado não chega a 20% de acurácia.

4.2.4 Bases Públicas com Máscaras

Os resultados até aqui foram obtidos utilizando NGT, com os parâmetros de vizinhança iguais a 1,3,5,10,50 e 100, respectivamente, considerando a abordagem do mascaramento artificial. São também exibidos os gráficos relativos ao comportamento da variação do limiar com a taxa de acurácia obtida pela modelo, bem como as tabelas cujos limiares refletem em uma máxima acurácia.

Este estudo tem como base experimentos baseados em problemas relacionados a oclusão especificamente com uso de máscaras faciais. Durante a preparação da base de dados, na etapa de detecção, ocorre o descarte de fotos por não se conseguir detectar um rosto utilizando o HOG, conforme explícito na Tabela 8.

Tabela 8 – Resultados para a LFW

Dataset	Nº de imagens	Imagens com faces detectadas	Imagens com faces oclusas detectadas
LFW	13233	12323	9506
Yale Face	16128	9015	4747
Ar Faces	3121	2765	2291
CelebFaces A	202599	199200	162448

Fonte: Elaborado pelo autor.

Um ponto de atenção é que notou-se uma redução mais drástica no número de fotos com faces detectadas no caso do base Yale Faces. Contudo, isso poderia ser esperado, já que se trata de uma base cuja uma parte dos exemplares tem variados casos de oclusão por uso de adornos no rosto como óculos e cachecóis com uma frequência muito maior em comparação as outras bases.

A tabela 9 mostra como ficaram distribuídos os resultados conforme se modificou o método de reconhecimento, com o emprego das distâncias mencionadas na seção 3.2. Os percentuais de acurácia levam em consideração o subconjunto dos bancos cujas faces são detectáveis.

Tabela 9 – Resultados de acurácia obtidos pelo método NGT para o dataset LFW.

Método de reconhecimento	Top n	Distância	Acurácia
NGT - Distância Euclideana	Top-1	0.56802	0.67329
	Top-3	0.58252	0.79457
	Top-5	0.59701	0.83033
	Top-10	0.59701	0.86999
	Top-50	0.62599	0.93531
	Top-100	0.68396	0.94972
NGT - Distância Manhatthan	Top-1	5.19364	0.65194
	Top-3	5.19364	0.77101
	Top-5	5.32416	0.810034
	Top-10	5.3241	0.85936
	Top-50	5.7157	0.92994
	Top-100	6.1073	0.94709
NGT - Distância dos cossenos	Top-1	0.07879	0.65488
	Top-3	0.08197	0.77574
	Top-5	0.08516	0.81087
	Top-10	0.0883	0.85242
	Top-50	0.09790	0.920164
	Top-100	0.09790	0.93815

Tabela 10 – Resultados de acurácia obtidos pelo método NGT para o dataset Yale Faces.

Método de reconhecimento	<i>Top n</i>	Distância	Acurácia
NGT - Distância Euclideana	Top-1	0.5582	0.83989
	Top-3	0.5704	0.90983
	Top-5	0.5704	0.93111
	Top-10	0.95028	0.95028
	Top-50	0.59462	0.97809
	Top-100	0.59462	0.98293
NGT - Distância Manhatthan	Top-1	5.1937	0.82683
	Top-3	5.1937	0.90351
	Top-5	5.1937	0.92795
	Top-10	5.1937	0.94838
	Top-50	5.30497	0.97387
	Top-100	5.30497	0.97767
NGT - Distância dos cossenos	Top-1	0.07318	0.79355
	Top-3	0.08150	0.85864
	Top-5	0.08150	0.87908
	Top-10	0.08150	0.89635
	Top-50	0.08428	0.91784
	Top-100	0.08983	0.91952

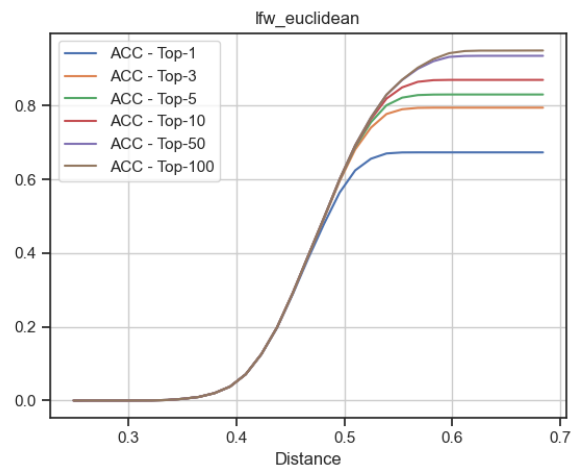
Tabela 11 – Resultados de acurácia obtidos pelo método NGT para o dataset AR Faces.

Método de reconhecimento	<i>Top n</i>	Distância	Acurácia
NGT - Distância Euclideana	Top-1	0.51406	0.72762
	Top-3	0.52875	0.81187
	Top-5	0.54344	0.84286
	Top-10	0.55813	0.88956
	Top-50	0.60220	0.95722
	Top-100	0.60220	0.97992
NGT - Distância Manhatthan	Top-1	4.8333	0.7237
	Top-3	4.8333	0.81099
	Top-5	4.8333	0.84242
	Top-10	5.3616	0.8856
	Top-50	5.3616	0.95896
	Top-100	5.8898	0.97817
NGT - Distância dos cossenos	Top-1	0.06778	0.73024
	Top-3	0.07110	0.80794
	Top-5	0.08438	0.84111
	Top-10	0.08438	0.88476
	Top-50	0.09102	0.9598
	Top-100	0.09102	0.97992

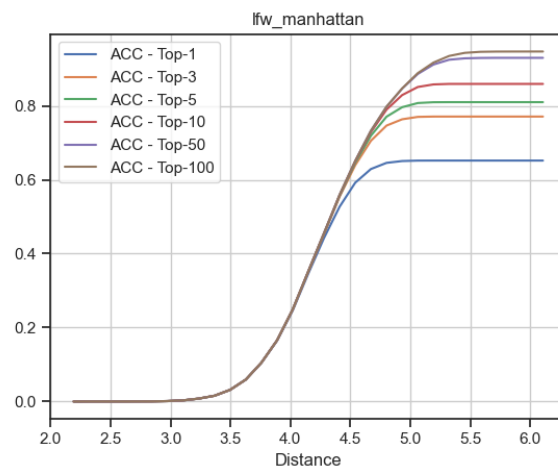
Tabela 12 – Resultados de acurácia obtidos pelo método NGT para o dataset CelebFaces A.

Método de reconhecimento	Top n	Distância	Acurácia
NGT - Distância Euclideana	Top-1	0.54718	0.38353
	Top-3	0.56418	0.52904
	Top-5	0.56418	0.58994
	Top-10	0.58118	0.6653
	Top-50	0.61517	0.80789
	Top-100	0.61517	0.85434
NGT - Distância Manhatthan	Top-1	4.8959	0.36508
	Top-3	5.0487	0.50847
	Top-5	5.0487	0.56985
	Top-10	5.0487	0.6474
	Top-50	5.5071	0.7967
	Top-100	5.5071	0.8454
NGT - Distância dos cossenos	Top-1	0.0752	0.3943
	Top-3	0.0786	0.5313
	Top-5	0.0786	0.5869
	Top-10	0.0855	0.6540
	Top-50	0.09248	0.7704
	Top-100	0.09248	0.8031

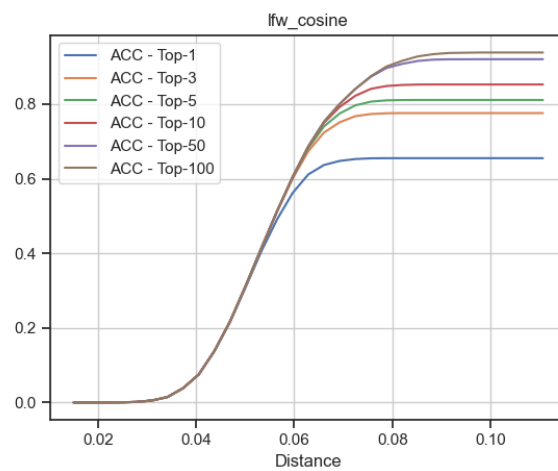
A partir das Figuras 31, 32, 33 e 34, é possível notar um comportamento padrão de atingimento de platô a medida em que o limiar da distância é aumentado. Esse é justamente o nível onde para fins de classificação a acurácia não é mais afetada independente da variação do limiar.



(a) Distância Euclideana

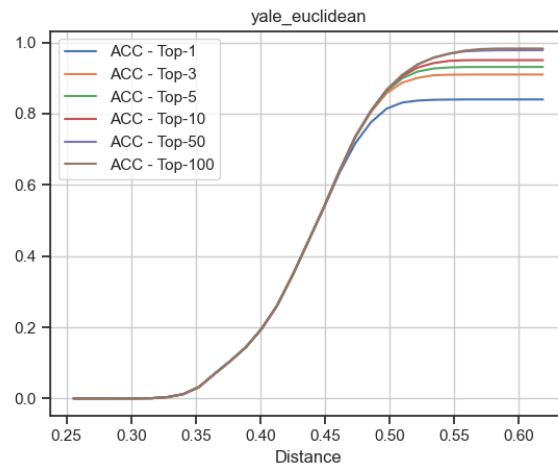


(b) Distância Manhatthan

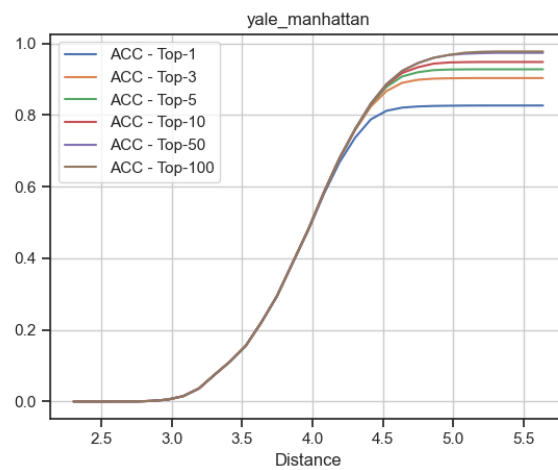


(c) Distância dos cossenos

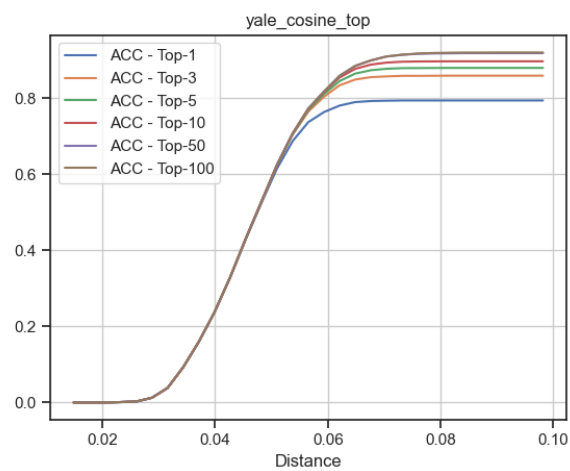
Figura 31 – Resultados para faces oclusas do dataset LFW



(a) Distância Euclideana

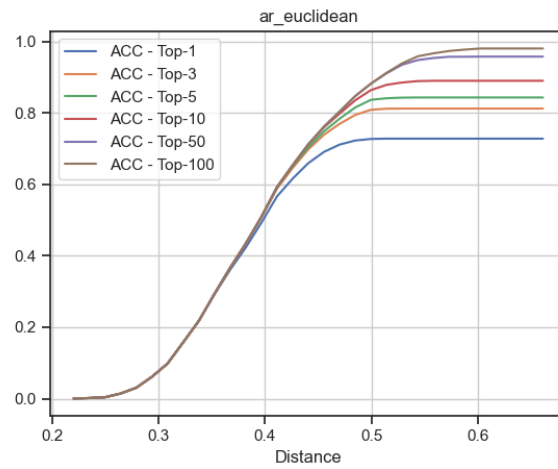


(b) Distância Manhatthan

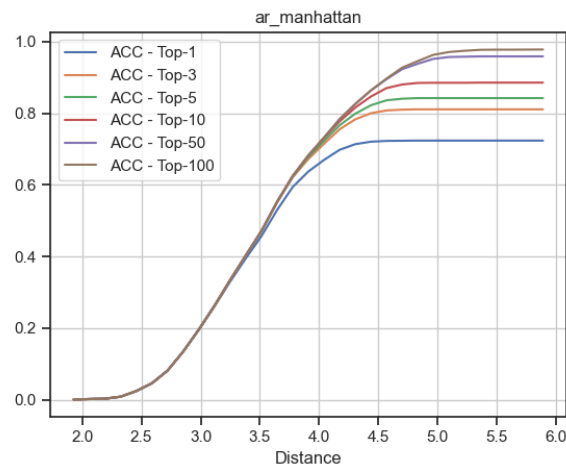


(c) Distância dos cossenos

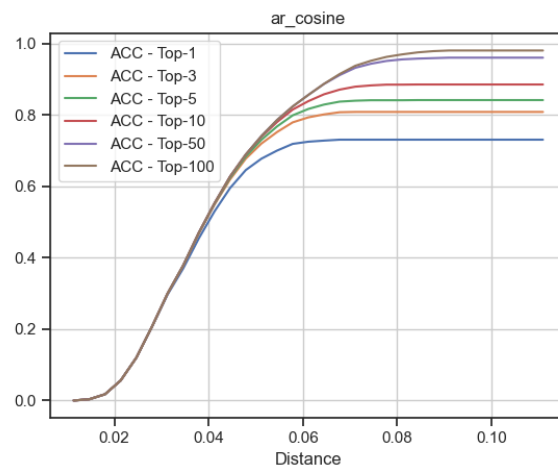
Figura 32 – Resultados para faces oclusas do dataset Yale Faces



(a) Distância Euclideana

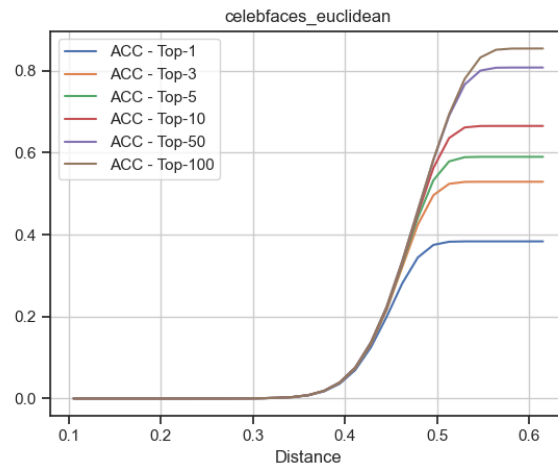


(b) Distância Manhatthan

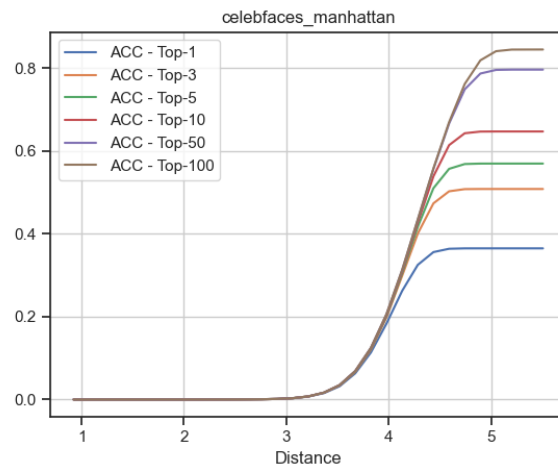


(c) Distância dos cossenos

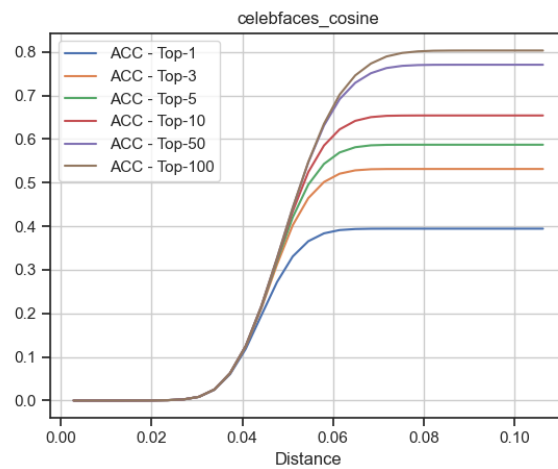
Figura 33 – Resultados para faces oclusas do dataset AR Faces



(a) Distância Euclideana



(b) Distância Manhattan



(c) Distância dos cossenos

Figura 34 – Resultados para faces oclusas do dataset CelebFaces A

5 Conclusões

Esta dissertação tem como principal contribuição prover a análise da eficácia da FaceNet em uma base com grande quantidade de fotos, no caso da base privada analisada, além de verificar o efeito do problema da oclusão para reconhecimento facial utilizando máscaras geradas automaticamente.

Demonstra-se que a etapa de detecção bem feita é fundamental para a análise crítica da base, exemplificado majoritariamente pelo estudo da base privada. Optou-se por utilizar o método HOG.

Considerando a base privada, resultados acima de 90% de acurácia foram obtidos. Contudo, salienta-se que isso se deu justamente na situação em que o subconjunto de fotos avaliadas eram justamente os que tinham uma qualidade boa. Ao tentar fazer o mesmo estudo utilizando de fotos com qualidade inferior o valor máximo de acurácia alcançado foi de cerca de 60%.

No que diz respeito ao problema da oclusão, no caso da base privada os resultados não passaram de 20%, possivelmente devido a limitação de mais exemplares de fotos de um mesmo indivíduo no banco. Em comparação com as bases públicas, onde o espaço amostral era bem menor e a disponibilidade de fotos para um mesmo indivíduos era maior, sugere-se que, quanto mais fotos de um indivíduo numa base de larga escalada, maiores as chances de encontrar essa pessoa mesmo com sua face parcialmente coberta. Isso é corroborado justamente na análise da maior base pública disponível, que é a CelebFaces A, com ordem de grande de 200 mil faces, onde a acurácia diminui em relação as outras bases justamente quando o espaço amostral em análise aumenta, indicando uma limitação de utilização da FaceNet para reconhecimento de faces oclusas em bases muito grande. Contudo, para pequenas aplicações de escopo limitado pode-se indicar que a técnica pode ter bons resultados.

Um ponto de atenção que relacionado as métricas de *matching*, no caso das bases públicas, apesar de haver uma leve vantagem quando se utilizou a distância Euclidiana, o comportamento dos três tipos de distância foi similar.

5.1 Trabalhos Futuros

Após a demonstração de resultados do método proposta aplicado, sugere-se que a linha de trabalhos futuros estejam focados em:

- Avaliação de outras redes neurais aplicadas aos problemas de reconhecimento e detecção facial, como Yolo e Detectron

-
- Verificar o efeito de outros métodos de detecção diferentes do HOG
 - A oclusão mencionada no trabalho foi feita utilizando-se apenas de um tipo de máscara, do tipo cirúrgica e cor de tons claros. Recomenda-se avaliar o efeito de máscaras de diferentes cores e formatos.
 - Verificar o efeito de outros tipos de oclusão por uso de acessórios no rosto partindo do mesmo princípio empregado neste trabalho.

Referências

- ABAZA, A.; HARRISON, M. A.; BOURLAI, T. Quality metrics for practical face recognition. Tsukuba, Japan, p. 3103–3107, Nov 2012. ISSN 1051-4651. Citado na página 33.
- AMOS, B. et al. Openface: A general-purpose face recognition library with mobile applications. *CMU School of Computer Science*, v. 6, n. 2, p. 20, 2016. Citado 2 vezes nas páginas 17 e 18.
- ANWAR, A.; RAYCHOWDHURY, A. *Masked Face Recognition for Secure Authentication*. 2020. Citado na página 42.
- AQUINO, C. P. et al. Avaliação do comportamento dos pontos fiduciais faciais durante o envelhecimento humano. Universidade Federal de Goiás, 2017. Citado 2 vezes nas páginas 7 e 31.
- BELHUMEUR, P.; HESPANHA, J.; KRIEGMAN, D. Eigenfaces vs. fisherfaces: recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Institute of Electrical and Electronics Engineers (IEEE), v. 19, n. 7, p. 711–720, jul. 1997. Disponível em: <<https://doi.org/10.1109/34.598228>>. Citado na página 26.
- BOOM, B. et al. The effect of image resolution on the performance of a face recognition system. Singapore, p. 1 – 6, 01 2007. Citado 2 vezes nas páginas 7 e 29.
- CARDARILLI, G. C. et al. *n*-dimensional approximation of euclidean distance. *IEEE Transactions on Circuits and Systems II: Express Briefs*, IEEE, v. 67, n. 3, p. 565–569, 2019. Citado na página 37.
- CHEN, J. et al. Facial expression recognition based on facial components detection and hog features. Yogyakarta, Indonésia, p. 884–888, 2014. Citado na página 43.
- CRUZ, J. E.; GUIMARÃES, L. N.; SHIGUEMORI, E. H. Um estudo da detecção automática de campos de futebol de imagens aéreas e orbitais utilizando SVM e descritores HOG. In: *XII Workshop de Computação Aplicada*. São José dos Campos: [s.n.], 2012. Citado na página 21.
- DALAL, N.; TRIGGS, B. Histograms of oriented gradients for human detection. San Diego, CA, EUA, 2005. Citado na página 21.
- ETEMAD, K.; CHELLAPPA, R. Discriminant analysis for recognition of human face images. In: *Audio- and Video-based Biometric Person Authentication*. Springer Berlin Heidelberg, 1997. p. 125–142. Disponível em: <<https://doi.org/10.1007/bfb0015988>>. Citado na página 26.
- FARFADE, S. S.; SABERIAN, M. J.; LI, L.-J. Multi-view face detection using deep convolutional neural networks. p. 643–650, 2015. Citado na página 23.

- FINLAYSON, G. D. Colour and illumination in computer vision. *Interface Focus*, The Royal Society, v. 8, n. 4, p. 20180008, jun. 2018. Disponível em: <<https://doi.org/10.1098/rsfs.2018.0008>>. Citado na página 29.
- FISHER, R. A. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, v. 7, n. 2, p. 179–188, 1936. Disponível em: <<https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-1809.1936.tb02137.x>>. Citado na página 26.
- FITOUSI, D. et al. Understanding the impact of face masks on the processing of facial identity, emotion, age, and gender. *Frontiers in Psychology*, v. 12, 2021. ISSN 1664-1078. Disponível em: <<https://www.frontiersin.org/articles/10.3389/fpsyg.2021.743793>>. Citado na página 14.
- FORCZMANSKI, P.; GEORGY, K.; SHCHEGOLEVA, N. An algorithm of face recognition under difficult lighting conditions. *Przegląd Elektrotechniczny*, v. 88, p. 201–204, 01 2012. Citado 2 vezes nas páginas 7 e 29.
- FREEMAN, M. R. W. T. *Orientation Histograms for Hand Gesture Recognition*. Cambridge, MA 02139, 1994. Disponível em: <<https://www.merl.com/publications/TR94-03/>>. Citado na página 19.
- GEORGHIADES, A.; BELHUMEUR, P.; KRIEGMAN, D. From few to many: illumination cone models for face recognition under variable lighting and pose. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 23, n. 6, p. 643–660, 2001. Citado na página 40.
- GHIASI, G.; FOWLKES, C. C. *Occlusion Coherence: Detecting and Localizing Occluded Faces*. 2015. Citado na página 28.
- HESELTINE, T. et al. Z.: Face recognition: A comparison of appearance-based approaches. p. 59–68, 01 2003. Citado 2 vezes nas páginas 7 e 24.
- HUANG, G. B. et al. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. 2008. Citado 2 vezes nas páginas 14 e 40.
- HUANG, G. B. et al. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. 2008. Citado na página 20.
- IWASAKI, M.; MIYAZAKI, D. *Optimization of Indexing Based on k-Nearest Neighbor Graph for Proximity Search in High-dimensional Data*. arXiv, 2018. Disponível em: <<https://arxiv.org/abs/1810.07355>>. Citado na página 38.
- JAIN, A. K.; LI, S. Z. *Handbook of face recognition*. [S.l.]: Springer, 2011. Citado na página 13.
- JEGOU, H.; DOUZE, M.; SCHMID, C. Product quantization for nearest neighbor search. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, v. 33, n. 1, p. 117–128, 2010. Citado 2 vezes nas páginas 33 e 34.
- JULINA, J. K. J.; SHARMILA, T. S. Facial recognition using histogram of gradients and support vector machines. In: *2017 International Conference on Computer, Communication and Signal Processing (ICCCSP)*. IEEE, 2017. Disponível em: <<https://doi.org/10.1109/icccsp.2017.7944082>>. Citado na página 24.

- JUNIOR, J. J. da S. *Redes neurais profundas para reconhecimento facial no contexto de segurança pública*. Dissertação (Mestrado) — Universidade Federal de Goiás, Goiânia, jul. 2020. Citado na página 13.
- KALAKUNTALA, P. *Performance Analysis of kNN Query Processing on large datasets using CUDA & Pthreads: comparing between CPU & GPU*. 2017. Citado na página 32.
- KAZEMI, V.; SULLIVAN, J. One millisecond face alignment with an ensemble of regression trees. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. Columbus, OH, EUA: IEEE, 2014. p. 1867–1874. Citado na página 13.
- KING, D. E. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, v. 10, n. Jul, p. 1755–1758, 2009. Citado na página 23.
- KING, D. E. Max-margin object detection. *arXiv preprint arXiv:1502.00046*, 2015. Citado na página 23.
- KORTLI, Y. et al. Face recognition systems: A survey. *Sensors*, v. 20, n. 2, 2020. ISSN 1424-8220. Disponível em: <<https://www.mdpi.com/1424-8220/20/2/342>>. Citado na página 13.
- KUANG, W.; BAUL, A. A real-time attendance system using deep-learning face recognition. 2020. Citado na página 18.
- LECUN, Y.; BENGIO, Y. et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, v. 3361, n. 10, p. 1995, 1995. Citado na página 22.
- LI, P. et al. *Face Recognition in Low Quality Images: A Survey*. 2018. Citado na página 30.
- LIU, Z. et al. Deep learning face attributes in the wild. In: *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*. USA: IEEE Computer Society, 2015. (ICCV '15), p. 3730–3738. ISBN 9781467383912. Disponível em: <<https://doi.org/10.1109/ICCV.2015.425>>. Citado na página 40.
- MALKAUTHEKAR, M. D. Analysis of euclidean distance and manhattan distance measure in face recognition. p. 503–507, 2013. Citado na página 37.
- MARTINEZ, A.; BENAVENTE, R. The ar face database. *Tech. Rep. 24 CVC Technical Report*, 01 1998. Citado na página 19.
- MARTINEZ, A. M. The ar face database. *CVC Technical Report 24*, 1998. Disponível em: <<https://ci.nii.ac.jp/naid/10011462458/en/>>. Citado na página 40.
- MI, J.-X. et al. Bilateral structure based matrix regression classification for face recognition. *Neurocomputing*, Elsevier, v. 348, p. 107–119, 2019. Citado na página 13.
- MUKHERJEE, D. et al. Energy efficient face recognition in mobile-fog environment. *procedia computer science*, Elsevier, v. 152, p. 274–281, 2019. Citado na página 17.
- MULLER, N.; MAGAIA, L.; HERBST, B. M. Singular value decomposition, eigenfaces, and 3d reconstructions. *SIAM review*, SIAM, v. 46, n. 3, p. 518–545, 2004. Citado na página 13.

- NUNES, L. F. d. M. Reconhecimento facial biométrico em nuvens de pontos tridimensionais. 2016. Citado na página 30.
- NYEIN, T.; OO, A. N. University classroom attendance system using facenet and support vector machine. In: IEEE. *2019 International conference on advanced information technologies (ICAIT)*. Yangon, Myanmar: IEEE, 2019. p. 171–176. Citado na página 18.
- PEARSON, K. Contributions to the mathematical theory of evolution. II. skew variation in homogeneous material. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, The Royal Society, v. 186, n. 0, p. 343–414, jan. 1895. Disponível em: <<https://doi.org/10.1098/rsta.1895.0010>>. Citado na página 24.
- PEARSON, K. LIII. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, Informa UK Limited, v. 2, n. 11, p. 559–572, nov. 1901. Disponível em: <<https://doi.org/10.1080/14786440109462720>>. Citado na página 25.
- PEIXOTO, S. A. et al. A high-efficiency energy and storage approach for iot applications of facial recognition. *Image and Vision Computing*, Elsevier, v. 96, p. 103899, 2020. Citado na página 17.
- PETERSON, L. E. K-nearest neighbor. *Scholarpedia*, v. 4, n. 2, p. 1883, 2009. Citado na página 32.
- QUIRITA, V. H. A. *Estudo de Métodos Automáticos de Reconhecimento Facial para Vídeo Monitoramento*. Tese (Doutorado) — PUC-Rio, 2014. Disponível em: <<https://doi.org/10.17771/PUCRio.acad.24340>>. Citado 2 vezes nas páginas 27 e 31.
- RANJAN, R.; PATEL, V. M.; CHELLAPPA, R. A deep pyramid deformable part model for face detection. IEEE, p. 1–8, 2015. Citado na página 23.
- RIBEIRO, F. M. et al. Detecção de pontos fiduciais sobre a face em tempo real. *sistema*, v. 200, p. 200, 2012. Citado na página 31.
- SALTON, K.; PRADO. *Face Recognition: Understanding LBPH Algorithm*. Towards Data Science, 2018. Disponível em: <<https://towardsdatascience.com/face-recognition-how-lbph-works-90ec258c3d6b>>. Citado 2 vezes nas páginas 7 e 25.
- SCHROFF, F.; KALENICHENKO, D.; PHILBIN, J. Facenet: A unified embedding for face recognition and clusterin. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. Boston, MA, EUA: IEEE, 2015. p. 815–823. Citado 7 vezes nas páginas 13, 14, 15, 17, 18, 19 e 32.
- SIROVICH, L.; KIRBY, M. Low-dimensional procedure for the characterization of human faces. *Journal of the Optical Society of America A*, The Optical Society, v. 4, n. 3, p. 519, mar. 1987. Disponível em: <<https://doi.org/10.1364/josaa.4.000519>>. Citado na página 24.
- SUMA, R. et al. Subjective evaluation of high dynamic range imaging for face matching. *IEEE Transactions on Emerging Topics in Computing*, IEEE, 2019. Citado na página 14.
- SUN, Y. et al. Deep learning face representation by joint identification-verification. p. 1988–1996, 2014. Citado 4 vezes nas páginas 7, 18, 26 e 27.

- SUN, Y.; WANG, X.; TANG, X. Deeply learned face representations are sparse, selective, and robust. p. 2892–2900, 2015. Citado na página 26.
- SWETS, D.; WENG, J. Using discriminant eigenfeatures for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Institute of Electrical and Electronics Engineers (IEEE), v. 18, n. 8, p. 831–836, 1996. Disponível em: <<https://doi.org/10.1109/34.531802>>. Citado na página 26.
- 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*. Columbus, OH, EUA: IEEE, 2014. p. 1701–1708. Citado 2 vezes nas páginas 7 e 26.
- TURK, M.; PENTLAND, A. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, MIT Press - Journals, v. 3, n. 1, p. 71–86, jan. 1991. Disponível em: <<https://doi.org/10.1162/jocn.1991.3.1.71>>. Citado na página 24.
- VOKHMINTCEV, A. et al. Face recognition based on a matching algorithm with recursive calculation of oriented gradient histograms. v. 93, n. 1, p. 37–41, 2016. Citado na página 14.
- WANG, D.; OTTO, C.; JAIN, A. K. Face search at scale. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, v. 39, n. 6, p. 1122–1136, 2016. Citado 3 vezes nas páginas 7, 33 e 34.
- WANG, J.; YUAN, Y.; YU, G. *Face Attention Network: An Effective Face Detector for the Occluded Faces*. 2017. Citado 2 vezes nas páginas 7 e 28.
- YANG, B. et al. Convolutional channel features for pedestrian, face and edge detection. CoRR. Citado na página 23.
- YANG, S. et al. From facial parts responses to face detection: A deep learning approach. p. 3676–3684, 2015. Citado na página 23.
- YI, D. et al. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014. Citado na página 33.
- ZHANG, C.; ZHANG, Z. Improving multiview face detection with multi-task deep convolutional neural networks. p. 1036–1041, 2014. Citado na página 23.
- ZHU, X.; RAMANAN, D. Face detection, pose estimation, and landmark localization in the wild. p. 2879–2886, 2012. Citado na página 13.