Detecção de Ataques Em Biometria Facial Utilizando Redes Neurais Convolucionais

Sandoval Verissimo de Sousa Neto

CENTRO DE INFORMÁTICA
UNIVERSIDADE FEDERAL DA PARAÍBA

João Pessoa, 2022

Sandoval Verissimo de Sousa Neto

Detecção de Ataques Em Biometria Facial Utilizando Redes Neurais Convolucionais

Dissertação de mestrado apresentada como requisito
para a obtenção do grau de Mestre pelo Centro de Informática
da Universidade Federal da Paraíba

Orientador: Leonardo Vidal Batista

Setembro de 2022

Catalogação na publicação
Seção de Catalogação e Classificação

S725d Sousa Neto, Sandoval Verissimo de.
Detecção de ataques em biometria facial utilizando redes neurais convolucionais / Sandoval Verissimo de Sousa Neto. - João Pessoa, 2022.
61 f. : il.

Orientação: Leonardo Vidal Batista.
Dissertação (Mestrado) - UFPB/CI.

1. Segurança contra acesso não autorizado. 2. Biometria facial. 3. Imagem digital. 4. Redes neurais. 5. Sistemas de anti-falsificação. I. Batista, Leonardo Vidal. II. Título.

UFPB/BC CDU 004.056.53(043)

Elaborado por Larissa Silva Oliveira de Mesquita - CRB-15/746

CENTRO DE INFORMÁTICA
UNIVERSIDADE FEDERAL DA PARAÍBA

Trabalho de Conclusão de Curso de Mestrado em Informáticaintitulado ***Detecção de Ataques Em Biometria Facial Utilizando Redes Neurais Convolucionais*** de autoria de Sandoval Verissimo de Sousa Neto, aprovada pela banca examinadora constituída pelos seguintes membros:

Prof. Dr. Leonardo Vidal Batista
Universidade Federal da Paraíba

Dr. João Jandyy Brasileiro Primo
Vsoft Tecnologia

Profa. Dra. Thais Gaudêncio Do Rego
Universidade Federal da Paraíba

João Pessoa, 17 de agosto de 2022

Centro de Informática, Universidade Federal da Paraíba
Rua dos Escoteiros, Mangabeira VII, João Pessoa, Paraíba, Brasil CEP: 58058-600
Fone: +55 (83) 3216 7093 / Fax: +55 (83) 3216 7117

DEDICATÓRIA

A todos aqueles que fizeram parte da minha vida durante esta jornada.

AGRADECIMENTOS

Gostaria de agradecer a orientação do professor Leonardo Vidal Batista, bem como a ajuda de todos aqueles que contribuíram passando seus conhecimentos a mim, em especial os membros da equipe da Vsoft, que me acompanharam desde o início dessa jornada, tornando assim possível alcançar os resultados obtidos neste trabalho.

Gostaria de agradecer também a minha família, apesar de tudo, sempre me incentivou a dar meu melhor, e sempre almejar por uma carreira de sucesso. A meu pai Sandoval Veríssimo de Sousa Filho, por ter me proporcionado uma boa educação, e ter cumprido seu papel de pai, sempre possibilitando o melhor para seu filho. A minha mãe Josenilda Alves de Sousa, que nunca me permitiu baixar a cabeça não importando as adversidades encontradas, e me serviu de porto seguro nos tempos mais difíceis.

Aos meus colegas que traçaram a maior parte dessa jornada comigo, onde enfrentamos dificuldades juntos, sempre superando-as e partindo para o próximo desafio.

Aos queridos amigos do coração, que encontrei uma vez que comecei a caminhar pela estrada que agora chega ao fim, comemorando os momentos de alegria, e não me permitindo titubear em face de adversidades encontradas.

Por fim gostaria de agradecer a Deus, por ter me garantido a fortitude de espírito necessária para chegar onde eu cheguei, bem como colocar todas as pessoas citadas acima em minha vida.

RESUMO

Com cada vez mais serviços sendo disponíveis virtualmente, métodos biométricos de autenticação que utilizam características como impressões digitais e face, são necessários para garantir uma melhor segurança para o usuário. A face de uma pessoa é um de seus traços biométricos mais importantes, principalmente devido a facilidade de uso, e com isso vem sendo cada vez mais estudada nos últimos anos. No entanto, conforme a utilização de métodos de autenticação por meio de biometria facial cresce, também aumentam as tentativas de ataques de impostores a esses sistemas. A grande facilidade de uso proporcionada pela biometria facial, também vem com a desvantagem de que devido ao grande uso de redes sociais pode ser mais fácil encontrar fotos ou vídeos de pessoas e utilizar assim esses registros para realizar ataques. Faz-se então necessário um sistema capaz de detectar se uma pessoa é genuína, ou se há uma foto ou vídeo de uma pessoa real tentando burlar um sistema. Essas aplicações são conhecidas como sistemas de anti-falsificação de face. Este trabalho propõe um método de detecção de falsificações utilizando redes neurais convolucionais. A transferência de aprendizado é utilizada para o treino do modelo. E o impacto de diferentes tipos de pré-processamento são estudados. Os testes são realizados em quatro bancos de imagens conhecidos na literatura (NUAA, MSU, Replay Attack, OULU). Os melhores resultados alcançam métricas melhores que alguns trabalhos da literatura, com uma taxa de erro igual inferior a $0,2\%$ no melhor experimento.

Palavras-chave: Anti-Falsificação, Face, Ataques de Apresentação, Detecção de Ataques, Classificação de Imagens, Redes neurais convolucionais profundas.

ABSTRACT

With more services becoming available online by the day, biometric authentication methods such as fingerprints and faces are necessary to provide better security for the user. A person's face is one of it's most critical biometric features, mainly due to the easiness of use, and so it has been increasingly studied in the last years. However, as the use of authentication methods with facial biometrics increases, so does the amount of attack attempts on these systems. The incredible ease of use of facial biometry also comes with the shortcoming that social media makes it may be easier to find photos and videos of someone and thus use its face to create attacks. Thus it is necessary a system that can detect if a person is real or if it is either a photo or video attack. These applications are known as Face-Anti-Spoofing systems. This work proposes a spoofing detection method using Convolutional neural networks. Transfer learning is used for training the model. The impact of different types of pre-processing tequiniques was studied. The experiments are made using four datasets widely known in the literature (NUAA, MSU, Replay Attack, OULU). The best results achieve better metrics than some works on literature. With an equal error rating lower than $0,2\%$ in the best experiment.

Key-words: Face Anti-Spoofing, Face, Presentation Attacks, Detection, Spoof Detection, Image Classification, Deep Convolutional Neral Network

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

CNN – *Convolutional Neural Network* (Redes Neurais Convolucionais)

Grad-CAM - *Gradient-weighted Class Activation Mapping* (Mapa de Ativações de Classes orientado por Gradiente)

NN - *Neural Network* (Rede Neural)

PDI - Processamento Digital de Imagens

RGB - *Red, Green, Blue* (Vermelho, Verde, Azul)

SSIM - *Structural Similarity Index Measure* (Medida de Índice de Similaridade Estrutural)

SVM – *Support Vector Machine* (Máquina de Vetores de Suporte)

Conteúdo

1 INTRODUÇÃO

A identificação biométrica busca analisar características físicas de uma pessoa, a fim de identificá-la de forma única. Desde os primeiros trabalhos sobre identificação automática de indivíduos, utilizando voz [14] ou face [24], passaram-se mais de 40 anos. A fim de melhorar quesitos de segurança em sistemas, e com o contínuo crescimento da área de computação, pesquisadores das áreas de visão computacional, processamento de imagem e reconhecimento de padrões se dedicaram-se à criação de melhores técnicas de reconhecimento biométrico. Entretanto, nos primeiros momentos, essas tecnologias estavam mais voltadas para aplicações de alta segurança, como análise forense, controle de fronteiras, entre outros.

Levando em consideração a possibilidade de identificar uma pessoa com base em sua biometria, vários sistemas passaram a incorporar biometria de maneira *on-line*. Por exemplo, sistemas bancários podem autorizar transações e a área de *e-commerce* pode requerer uma autenticação biométrica. Ainda assim, existe uma deficiência que é o fato de que, para se analisar traços biométricos, é necessário algum tipo de sensor ou dispositivo de captura. Para identificar uma pessoa pela voz, é necessário um microfone, para utilizar a face, uma câmera, e para identificar uma impressão digital, um sensor. Isso fazia com que a tecnologia biométrica, apesar de existir, não fosse tão viável há algumas décadas atrás.

Hoje em dia, com o avanço da tecnologia na área de eletrônica e computação em geral, essa identificação é utilizada frequentemente em sistemas no cotidiano. O exemplo mais comum é a habilidade de desbloquear a tela dos *smartfones* com a impressão digital ou a face. Esses dois traços biométricos são os mais utilizados pelos sistemas de identificação biométrica, e cada um tem suas peculiaridades. A impressão digital é única para cada pessoa, logo provê uma grande confiança no resultado. A face, tem a vantagem de ser uma característica biométrica que pode ser analisada de forma menos intrusiva, ou seja, sem a necessidade que o usuário do sistema colaborativamente apresente a face à uma câmera. Isso permite que ela seja utilizada tanto para sistemas de autorização biométrica, quanto sistemas de vigilância.

Enquanto de um lado, diversos avanços na disponibilidade e qualidade da verificação biométrica eram feitos ao longo dos anos, preocupações sobre possibilidades de fraudes também surgiam. Enquanto esses sistemas de identificação visam discernir pessoas com base em dados biométricos, questionamentos sobre a possibilidade de forjar essas características ganharam atenção. De fato, segundo o trabalho de Mohanty et al. [35], é possível simular dados biométricos e assim burlar sistemas. Assim iniciativas para estudar técnicas de identificar esses ataques de falsificações (também conhecidos como *Spoofs*), foram iniciadas.

Nas últimas duas décadas, esse tópico vem sendo mais estudado. Pesquisadores passaram a tentar desenvolver técnicas que podem burlar as vulnerabilidades de sistemas a fim de validá-los, e ao mesmo tempo encontrar maneiras de fazer com que sistemas possam se proteger desses ataques. O assunto até saiu do ramo acadêmico e é possível encontrar tutoriais de como forjar impressões biométricas na internet para leigos.

Enquanto é possível fraudar traços biométricos, obter uma falsificação de alguns é mais difícil do que outros. A impressão digital, por exemplo, precisa ser extraída de algum documento ou registro previamente deixado por uma pessoa, ou obtida por meio de algum objeto que foi tocado recentemente. A imagem da face de uma pessoa, por outro lado, é algo muito mais fácil de se obter, e em anos recentes, com o aumento da utilização das redes sociais, usuários estão praticamente entregando esse particular traço biométrico para outros utilizarem em possíveis ataques.

Com a recente onda de aplicações que passaram a utilizar aprendizagem de máquina e aplicar redes neurais convolucionais, ou CNNs (do inglês *Convolutinal Neural Network*) para resolução de problemas associados com processamento de imagem, faz sentido que essas técnicas também sejam empregadas em trabalhos sobre a detecção de falsificações. Entretanto, a maior parte dos trabalhos hoje em dia foca em obter bons resultados em bancos de imagens específicos encontrados na literatura [16], e frequentemente tem resultados ruins ao sair daquele padrão de imagens especifico [52].

1.1 Definição do Problema

A face, como um um traço biométrico, é extremamente vulnerável à realização de ataques, mesmo assim, sua facilidade de uso e avanços na área de tecnologia vem a tornando cada vez mais comum no cotidiano. Sendo assim, a necessidade de métodos que consigam discernir imagens com pessoas genuínas, de ataques maliciosos, é uma realidade. Isso combinado com o fato de que devido à recente pandemia mais e mais pessoas estão utilizando sistemas de autenticação biométrica para realizar atividades de casa, requer que tais soluções sejam genéricas e funcionem em casos reais diferentes de ambientes controlados e padronizados. Grande parte dos trabalhos mais recentes utilizam CNNs e focam na concepção de novos métodos que são capazes de obter resultados muito bons em um banco de imagens da literatura em específico. Entretanto os resultados frequentemente não generalizam bem para outros bancos. Num cenário real, diversas variáveis podem influenciar na precisão do algoritmo, tais como dispositivo de captura, iluminação e distância da face à tela. Enquanto isso, técnicas clássicas de visão computacional tentam fazer uso de características mais genéricas para tentar deixar suas soluções mais adaptáveis para diversos cenários, mas frequentemente seus resultados não são tão bons quanto as CNNs.

Este trabalho propõe um modelo de redes neurais convolucionais baseado em transferência de aprendizado. O modelo foi treinado e também foram feitos diversos experimentos com diferentes técnicas de pre-processamento. Os resultados desses pre-processamentos foram analisados e alguns pontos de discussão foram comentados.

1.2 Objetivo geral

O objetivo geral deste trabalho é propor um modelo baseado em redes convolucionais profundas, utilizando transferência de aprendizado para classificar imagens de pessoas e discernir se a pessoa é genuína, ou se é uma foto ou vídeo de uma pessoa sendo exibido.

1.3 Objetivos específicos

Os objetivos específicos são:

1. Analisar os principais bancos de imagens disponíveis publicamente para o problema de anti-falsificação de face;
2. Desenvolver um modelo baseado em redes neurais convolucionais para classificar imagem entre real ou falsa;
3. Analisar o impacto de diversas formas de pré-processamento.

1.4 Trabalhos Relacionados

Segundo o exposto por Ming et al. [34] trabalhos relacionados a anti-falsificação de faces podem ser divididos primariamente em 3 categorias: Métodos Baseados em Vivacidade, Métodos Baseados em Textura e Métodos Baseados em Formas 3D. Há também a possibilidade de combinar mais de um desses métodos no que se chama de Métodos Baseados em Múltiplas informações.

1.4.1 Métodos baseados em vivacidade

Os métodos baseados em vivacidade buscam detectar variações numa sequência de imagens, que podem indicar a vivacidade de uma pessoa. Esses movimentos são ações como piscada de olhos, mudanças de expressões faciais e até micro variações na pulsação sanguínea. Esses tipos de métodos ainda se dividem em duas subcategorias, os intrusivos e não intrusivos. Os métodos intrusivos requisitam uma certa ação do usuário, este então deve realizar a ação correspondente, num sistema de comando e resposta. Como por exemplo, o trabalho em Kollreider et al. [26], que faz uso de uma máquina de vetores de

suporte, ou SVM [48] (do inglês *Support Vector Machine*). Já os não intrusivos podem ser executados sem que o usuário perceba, pois não exigem interação ativa de sua parte. É possível citar Li et al. [29] que fazem uso de características baseadas no domínio da frequência, Aggarwal e Nandhakumar [3], e Azarbayejani et al. [6], que fazem uso de mapas 3D para tentar estimar movimentos na cabeça, Kollreider et al. [25] , que fazem uso do fluxo de linhas ópticas e Pan et al. [38], que utilizam piscadas de olhos. Em geral, métodos baseados em vivacidade funcionam bem para ataques com imagens, mas podem facilmente ser enganados com vídeos. O problema dos vídeos é amenizado quando se usa um método intrusivo de comando e resposta, onde os comandos são dados de maneira aleatória. Ainda assim, é possível gravar uma série de respostas e tentar escolher o certo. Além disso, esses métodos não funcionam em uma única imagem pois requerem um fluxo de vídeo.

1.4.2 Métodos Baseados em Textura

Métodos baseados em textura são provavelmente os mais usados para o problema de anti-falsificação de faces. Esses métodos não exigem nenhum tipo de interação ativa com o usuário, assim são inerentemente não intrusivos. Esse tipo de método pode ser aplicado em apenas uma imagem, eliminando uma limitação dos métodos baseados em vivacidade. Métodos baseados em textura estáticos são aqueles que levam em consideração uma única imagem para realizar uma classificação. Alguns trabalhos que podem ser mencionados são: Diferença da reflectância da luz [29], e sua variante que usa filtros por Diferença de Gaussiana, ou DoG [43] (do inglês *Difference Of Gaussians)*, Equalização de histograma adaptativa limitada com contraste [40], padrões binários locais, ou LBP [36] (do inglês *Local Binary Patterns*), ondas de Gabor [33] e histograma de gradientes orientados, ou HOG [13] (do inglês *Histogram of Oriented Gradients*). Já os métodos de textura dinâmicos utilizam a informação temporal de uma sequência de imagens. Trabalhos envolvem métodos como: LBP [15], Histograma de fluxos ópticos orientados, ou HOOF [7] (do inglês *Histogram of Oriented Optical Flows*), Análise de ruído residual de Fourier em vídeo [12]. Outros trabalhos mais recentes fazem uso de CNNs, um tipo de arquitetura de rede neural que é comumente utilizado no processamento de imagens, mas devido a ser algo mais próximo deste trabalho, uma seção vai ser dedicada para estes métodos mais a frente.

1.4.3 Métodos baseados em características 3D

Métodos baseados em informações 3D visam diferenciar uma tela ou foto de uma pessoa real utilizando a informação 3D da imagem. Essas abordagens focam geralmente em reconstruir a informação 3D a partir de uma imagem 2D [49], ou na estimação de

profundidade da imagem. Esses métodos são em geral mais facilmente realizados com uma câmera 3D. Entretanto esse tipo de câmera não é de fácil disponibilidade para o consumidor geral, logo outros métodos precisaram ser desenvolvidos, como por exemplo o trabalho de Atoum et al. [5], que faz uso de um pseudo mapa de profundidades numa sequencia de imagens.

1.4.4 Métodos baseados em CNNs

O primeiro trabalho a envolver CNN que propôs um método para resolver o problema de anti-falsificação de faces foi de Yang et al. [52], e fez uso de uma arquitetura AlexNet [27], com um SVM ao final da rede para fazer a classificação. Neste trabalho os autores confirmaram uma hipótese anterior, que dizia que aumentar a área do retângulo da face detectada para incluir mais do fundo da imagem foi benéfico para os resultados do modelo até um certo ponto. Em [39], Patel et al. propuseram uma solução utilizando a mesma rede, porém sem o SVM, assim todo o processo de classificação era feito pela rede convolucional. Esse trabalho propôs um sistema de votação com base no resultado de duas redes idênticas, uma treinada em imagens sem nenhum tipo de alinhamento e outra treinada com as faces alinhadas. O trabalho de Li et al. [30] propôs um ajuste fino numa arquitetura VGG previamente treinada, com imagens do problema de anti-falsificação de face. Esta decisão também foi tomada na metodologia deste trabalho, entretanto há o diferencial que o trabalho citado ainda inclui uma análise de componentes principais (PCA, do inglês *Principal Component Analysis*), seguida de um SVM ao final da rede para fazer a classificação, o que diferencia os dois trabalhos. Em [20], George e Marcel mostram um modelo treinado utilizando duas funções de perda calculadas no último mapa de características da rede e no resultado de classificação. As funções de perda eram calculadas num nível por pixel e mostraram resultados promissores.

1.4.5 Contribuições principais deste trabalho

Finalmente, tendo apresentado alguns dos trabalhos relacionados é possível destacar as principais contribuições deste trabalho, que o diferencia dos outros citados. Em particular, destacam-se a questão da análise com o pré-processamento de subamostragem para remover imagens muito similares do treinamento; a realização de experimentos em quatro bancos conhecidos da literatura e mais um banco privado, cujas imagens tem uma maior variação; experimentos feitos tanto com bancos de dados isolados, quanto entre diferentes bancos de dados, para testar a capacidade de generalização dos modelos; análise de diferentes pre-processamentos e como eles impactam na acurácia do modelo baseado em transferência de aprendizado.

1.5 Estrutura do Trabalho

Este trabalho foi dividido em mais 4 capítulos: Capítulo 2; Conceitos Gerais, que apresenta as definições dos principais métodos usados neste trabalho; Capítulo 3; Materiais e Metodologia, que mostra informações sobre o ambiente de desenvolvimento, as bases de dados e a descrição do método proposto neste trabalho; Capítulo 4; Apresentação e Análise dos Resultados, que exibe todos os resultados obtidos e algumas considerações sobre eles; por fim, Capítulo 5; Trabalhos Futuros e Conclusão, vai brevemente sumarizar alguns dos pontos mais importantes discutidos no trabalho e comentar sobre possíveis melhorias e trabalhos futuros.

2 CONCEITOS GERAIS

Neste capítulo, são apresentados os principais conceitos teóricos utilizados na produção deste trabalho, que disserta sobre conceitos gerais relacionados a redes neurais, redes neurais convolucionais, imagens e Processamento Digital de Imagens.

2.1 Imagem Digital

A definição comum para imagem é que se trata de uma representação de algo ou alguém, porém na computação, essa definição vaga é melhor explicada de outra maneira. Uma imagem digital (monocromática), na área de computação, pode ser interpretada como uma matriz de valores, onde cada um deles tem um valor finito que representa a intensidade do tom do cinza de um pixel na tela [21]. Esses pixels são agrupados numa forma bidimensional, frequentemente interpretada como uma matriz 2D ou uma função bidimensional *I(u,v)*.

Para que seja possível processar uma imagem digital, é necessário capturá-la do mundo real com um sensor, e realizar dois passos: amostrar a imagem e quantizar a imagem. O processo de amostragem se refere ao intervalo de tempo em que um sinal é capturado T, dado em medida de tempo ou espaço, dependendo do sinal. A frequência de amostragem f é dada pelo inverso de T e representa quantas amostras são extraídas em um certo espaço, ou período de tempo. Já a quantização está relacionada aos conjunto de valores $Q = \{-nq, ..., -2q, -q, 0, q, 2q, ..., mq\}$, que cada amostra de um sinal vai armazenar em memória. Esses valores são finitos e discretos, onde q se refere a um passo de quantização e m e n são números inteiros, positivos e não nulos. A Figura 1 mostra uma representação visual.

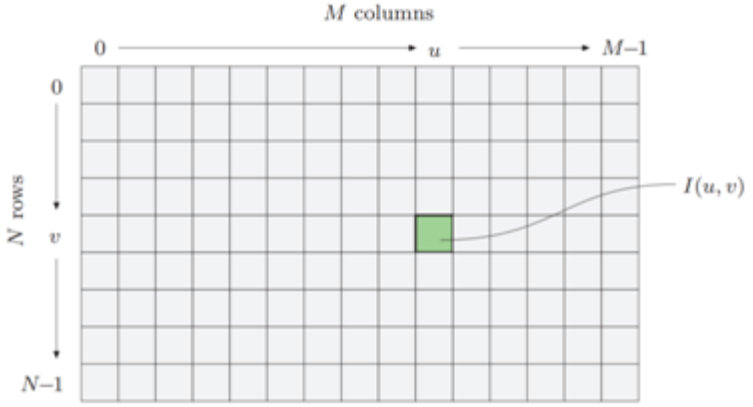


Figura 1: Representação gráfica de uma imagem. Onde a intensidade (I) de um pixel depende de sua posução (u,v) em relação a imagem. Fonte: [44].

Na Figura 1 M e N representam as dimensões da imagem (número de colunas e

linhas, respectivamente), quanto maior a resolução da imagem, maior o número total de pixels. O valor de um pixel que se encontra na u-ésima linha e na v-ésima coluna é dado como $I(u,v)$. Os possíveis valores de I variam de acordo com o conjunto Q definido no processo de quantização. O padrão mais comum encontrado em imagens cotidianas e que vai ser utilizado neste trabalho é que 0 representa o preto, enquanto 255 representa o branco.

Para representar cores numa imagem, é comum que se utilizem 3 valores por pixel, um para vermelho, um para verde e um para azul (sistema de cores RGB, do inglês *Red, Green, Blue*). Essa convenção é dada pois essas cores maximizam as diferenças entre as respostas das células cone nos olhos humanos [22]. Monitores e dispositivos que exibem essas imagens frequentemente também adotaram essa representação. Quando se trata um pixel como um conjunto de valores RGB, ele passa a ser representado por um vetor de 3 componentes, onde cada um é a representação da intensidade do pixel em uma cor. Neste trabalho será utilizado um pixel como um vetor RGB e a Figura 2 mostra uma representação gráfica de uma imagem RGB.

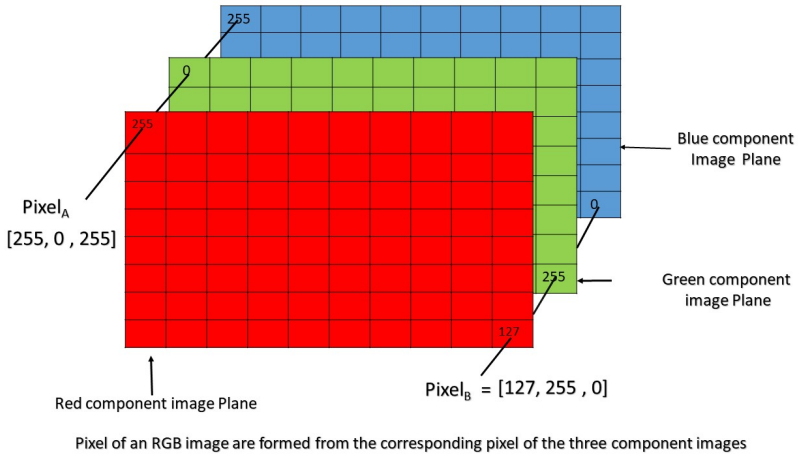


Figura 2: Representação gráfica de uma imagem RGB. Onde cada camada representa a intensidade em um dos tons de cor. Fonte: [2].

2.1.1 Processamento Digital de Imagens

Tendo entendido como uma imagem é representada no contexto de computação, é interessante entender o processo de alteração dessas imagens. Processamento digital de imagens (PDI) é uma subcategoria do processamento de sinais digitais, e se refere a utilização de um computador para processar imagens por meio de um algoritmo. Algoritmos mais complexos de PDI permitem aplicações voltadas para problemas como: Classificação, Extração de Características, Análise de padrões, entre outras. Existem ainda duas áreas muito próximas de PDI, são elas: Visão computacional, que trata de extrair informação útil de imagens, e Computação gráfica, que realiza o inverso e utiliza dados para gerar

imagens. A principal diferença dessas áreas em relação ao processamento digital de imagens é que PDI é relacionada a modificação de uma imagem de entrada gerando uma imagem de saída. A Figura 3 mostra uma explicação gráfica das três áreas.

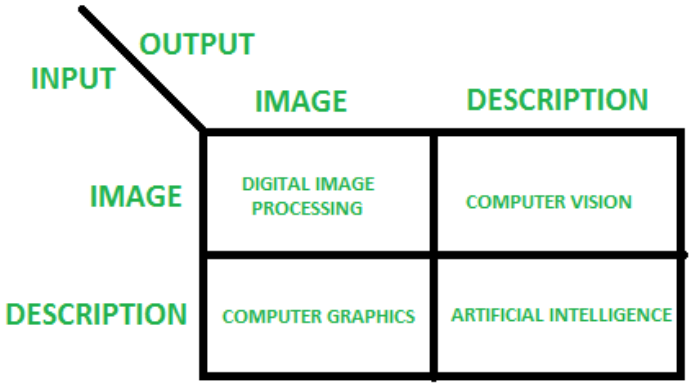


Figura 3: Representação das áreas próximas a processamento digital de imagens. Fonte: [1].

2.2 Redes Neurais

Redes neurais, ou NNs (do inglês *Neural Networks*), são um modelo de aprendizagem de máquina, e são o principal componente de algoritmos de aprendizagem profunda ou *deep learning*. Seu nome e estruturas são inspirados no cérebro de animais, e seu funcionamento tenta imitar as interações que os neurônios humanos realizam [19].

Redes neurais são compostas de camadas de nós, e geralmente contêm uma camada de entrada, diversas camadas intermediárias e uma camada de saída. Cada um dos nós da rede (um neurônio artificial) se comunica com outro e tem uma função de ativação associada. Se a resposta de cada neurônio individual for suficiente para ter uma resposta de determinada função de ativação, o neurônio ativa e envia informações para a próxima camada da rede, caso contrário, os dados não são propagados. Redes neurais dependem dos dados de treinamento para aprender e melhorar sua precisão ao longo do tempo. Uma vez treinada, uma rede pode ser capaz de realizar uma decisão muito rapidamente. A Figura 4 mostra um esquema simplificado de uma arquitetura de rede neural com uma camada de entrada, uma camada intermediária (ou camada escondida), e uma camada de saída.

Para melhor entender como as redes neurais funcionam é possível imaginar cada nó individual como um modelo de regressão linear, composto de um dado de entrada, pesos, um limiar, e uma saída. Quando um dado de entrada é definido, os pesos dos neurônios determinam a importância de cada variável. Todas as entradas dos nós são

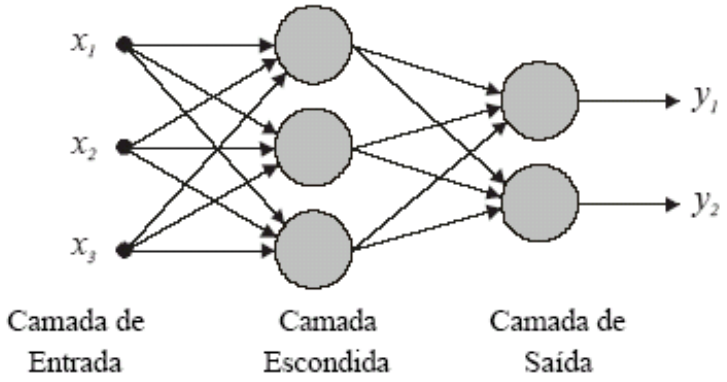


Figura 4: Representação simples de uma rede neural. Fonte: [37].

então multiplicados pelos seus pesos, e então somados. Em seguida, o resultado passa por uma função de ativação, que determina se a saída é relevante o suficiente para ser passada para a próxima camada ou não.

Segundo Zell [54], de forma matemática, um neurônio j recebendo uma entrada $p_j(t)$ de um neurônio predecessor é composto dos seguintes componentes:

- Uma ativação $a_j(t)$, é o estado do neurônio que varia com um parâmetro discreto.
- Um limiar opcional θ_j, que é fixo a menos que seja alterado durante o treinamento.
- Uma função de ativação f, que computa uma nova ativação num tempo $t+1$ de um $a_j(t)$, θ_j e uma nova entrada $p_j(t)$, gerando a seguinte relação:

$$a_j(t+1) = f(a_j(t), p_j(t), \theta_j)$$

.

- Uma função de saída $f_{out}(a_j(t))$, que calcula a saída com base na ativação dada por:

$$o_j(t) = f_{out}(a_j(t))$$

Para a primeira camada da rede, a camada de entrada, não existem predecessores, assim ela serve como interface de entrada de dados para a rede. Similarmente, a última camada não tem uma sucessora, assim seu resultado é considerado a saída da rede neural.

2.2.1 Redes Neurais Convolucionais

Redes neurais convolucionais, frequentemente referidas como CNNs, são um subtipo de redes neurais comumente utilizada na análise e processamento de imagens. Elas

se baseiam numa arquitetura de pesos compartilhados de máscaras de convolução. Essas máscaras deslizam sobre uma entrada e produzem respostas chamadas de mapas de características [47]. Para construir uma rede neural convolucional existem três tipos principais de camadas: Camadas Convolucionais, Camadas de Subamostragem e Camadas Densas.

A camada convolucional é o ponto distinto desse tipo de arquitetura de redes neurais. Seus elementos são filtros que possuem a capacidade de aprender. Os filtros frequentemente são pequenos em largura e altura (geralmente 3x3 ou 5x5), mas atuam por todo volume da entrada. Durante a execução da rede, esses filtros deslizam (convoluem) sobre a entrada. À medida que essas convoluções ocorrem, um mapa de ativação 2D é gerado. Esse mapa dá uma resposta para o filtro em cada posição espacial. A rede vai então aprender filtros que ativam ao encontrar certas características, que podem ser simples como bordas, manchas, cores, ou complexas como rodas de veículos.

A camada de subamostragem é geralmente utilizada entre um conjunto de camadas de convolução. Essa camada serve para reduzir a representação espacial das entradas, reduzindo assim a quantidade de parâmetros treináveis das redes, e auxiliando no controle de possíveis superajustes. Enquanto existe mais de uma forma de realizar a operação de subamostragem, geralmente se usa o valor máximo de uma vizinhança, devido a sua agilidade. O passo de subamostragem dita o quanto as dimensões da entrada serão reduzidas. Comumente os passos são números inteiros e iguais nas duas dimensões, sendo 2x2 o mais comum.

As camadas densas também são utilizadas em redes neurais convolucionais, frequentemente sendo referidas como camadas totalmente conectadas. O motivo dessa nomenclatura se dá pois as camadas densas frequentemente vem ao fim de todas as camadas convolucionais, tendo conexões com todos os neurônios da camada anterior. Uma vez que a resolução espacial das entradas está suficientemente pequena, é comum adicionar camadas densas para realizar os cálculos finais antes da saída da rede.

Uma das grandes vantagens das CNNs quando comparadas as redes convencionais está quando comparamos às entradas das redes. Por exemplo, para um problema clássico como o da identificação de dígitos [28], um dos primeiros bancos de imagens disponível é o MNIST [18]. Nele as imagens têm uma dimensão de apenas 28x28 pixels, e apenas 1 canal de cor. Caso seja necessário processar uma imagem pequena como essa numa rede comum, o vetor de entrada da rede vai ser de tamanho 784. Enquanto isso pode ser manejável, numa imagem com 3 canais de cores, o tamanho de entrada subiria para 2352. Arquiteturas mais comuns conseguem tratar imagens em que a resolução de entrada é acima de 200x200x3, o que resultaria em um vetor de tamanho 120000 para uma rede neural comum.

3 MATERIAIS E METODOLOGIA

3.1 Ambiente de desenvolvimento

Para a realização deste trabalho foi escolhida a linguagem de programação Python 3.7. Junto a ela foram utilizados os módulos OpenCV e NumPy para o processamento de imagens, o módulo Jupyter Lab para codificação do projeto, a API de alto nível para aprendizagem de máquina Keras, Tensorflow-GPU, como fundação para o treinamento dos modelos. Os experimentos foram realizados em uma máquina com processador Intel Core i7-8700 de 3.20Ghz, que possui 48GB de memória RAM, e uma placa de vídeo Nvidia Titan X. O sistema operacional do computador é o Windows 10 com 64 bits.

3.2 Bases de dados

Este trabalho faz uso de 4 bancos de imagens publicamente disponíveis e amplamente utilizados em trabalhos da área de anti-falsificação de face, são eles o NUAA, MSU-MFSD, Replay Attack, OULU-NPU. Em adição a estes bancos, também foi utilizado um banco privado criado pela empresa Vsoft Tecnologia. Este último tem cenários muito mais descontrolados que os bancos da literatura, em relação a aspectos como iluminação, fundo e variação de pose, sendo assim mais desafiador. A Figura 5 exibe algumas amostras de imagens reais e impostoras (casos de uma foto ou vídeo ataque) de cada banco.

Vendo as imagens, é possível notar que, diferenciar uma face real de um ataque não é uma tarefa trivial até mesmo para os humanos. Dos bancos de imagens utilizados neste trabalho, os bancos NUAA e Vsoft são compostos por imagens, enquanto os bancos MSU-MFSD, Replay Attack e OULU-NPU são compostos por pequenos vídeos de menos de 15 segundos cada. Para estes, foram extraídos cada um dos quadros dos vídeos, assim transformando todos os dados para imagens.

A seguir será dada um breve descrição de cada um dos bancos de imagens, suas características e como eles se diferenciam uns dos outros.

3.2.1 NUAA

O banco de imagens NUAA [43] foi o primeiro banco para o problema de anti-falsificação de face a ser publicamente disponível. O banco de imagens possui aproximadamente 12000 imagens, de 16 indivíduos. Por padrão as partições estão divididas apenas em treino e teste, e há sobreposição de identidades nas duas partições. Os ataques nesse banco são realizados apenas por meio de fotos de pessoas reais impressas em papéis. O banco de imagens foi criado em três sessões de captura, onde cada sessão varia as condições de iluminação e as informações de fundo das imagens. Dito isso, nem todas

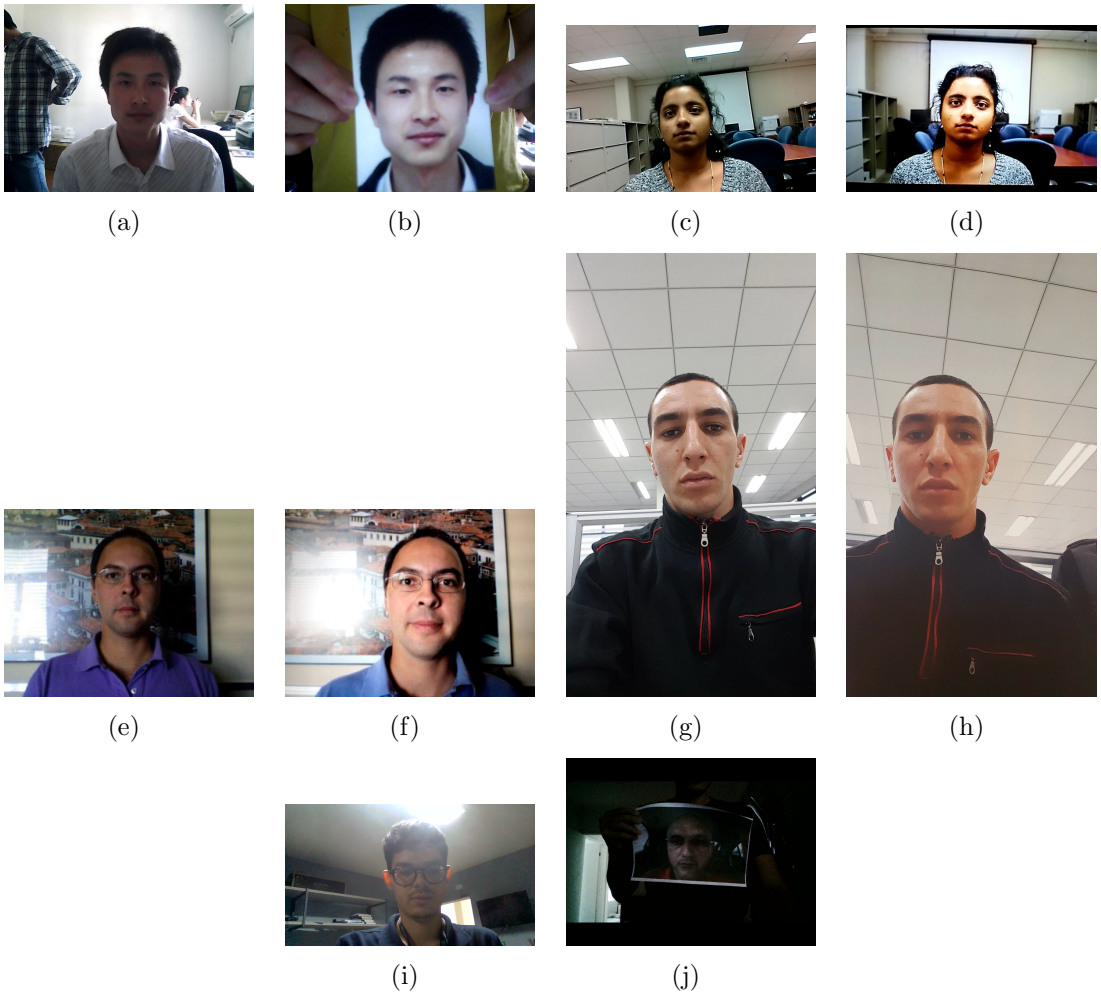


Figura 5: Amostras de imagens reais e de ataque dos bancos : NUAA (a) real | (b) ataque; MSU-MFSD (c) real | (d) ataque; Replay Attack (e) real | (f) ataque; OULU-NPU (g) real | (h) ataque; Vsoft (i) real | (j) ataque. Fonte: Autoria Própria

as identidades existem em cada sessão de captura, e também não existe nenhum tipo de padronização em relação a quantas imagens de cada pessoa foram capturadas em cada sessão. O dispositivo de captura das imagens reais e de ataque não é detalhado, mas os autores do banco dizem que foi utilizada uma webcam barata.

3.2.2 Replay Attack

O Replay attack [10] é um dos bancos de imagens mais conhecidos e utilizados para o problema de anti-falsificação de face. Ele é um banco baseado em vídeos, que possui um total de 1200 gravações. Desse total, 200 são de pessoas genuínas e 1000 são de ataques. Os vídeos vêm divididos em 3 partições, treino, desenvolvimento (validação) e teste, com 15, 15 e 20 identidades em cada partição, respectivamente. Não existe nenhuma identidade que é vista em outras partições, e os vídeos reais e de ataque têm duração aproximada de

15 e 9,5 segundos, respectivamente. As sessões de captura foram feitas em dois ambientes distintos, um controlado, onde as pessoas se encontram em frente a um fundo uniforme, e um sem controle, onde há um quadro atrás das pessoas refletindo uma luz. Os ataques nesse banco de imagens foram feitos com fotos impressas em papel, fotos exibidas em display e vídeos exibidos em display. Após o processo de extração de quadros, 75000 imagens reais e 235000 imagens de ataques foram obtidas.

3.2.3 MSU-MFSD

Com 280 vídeos sendo 70 deles de pessoas reais e 210 deles de ataques, o banco de imagens MSU-MFSD [51] é outro banco baseado em vídeos publicamente disponível na internet. Neste banco existem 35 identidades divididas em 15 para treino e 20 para teste. A taxa de captura dos quadros dos vídeos e a duração variam entre os vídeos. Após o processo de extração dos quadros de vídeo, existem aproximadamente 19000 imagens de pessoas reais e 58000 imagens de ataque. Similarmente ao banco Replay Attack, o MSU-MFSD possui ataques realizados com fotos impressas, fotos em displays e vídeos. Esse banco foi o primeiro a introduzir dispositivos móveis como meios de reprodução dos ataques (previamente, iPads eram utilizados).

3.2.4 OULU-NPU

Um dos bancos de imagens mais novos, o OULU-NPU [55] foi disponibilizado em 2017. O banco de imagens é baseado em vídeos, e dispõe de 55 identidades divididas em 20 para treino, 15 para validação e 20 para teste. Há um total de 4950 vídeos neste banco, sendo 990 de acessos reais e o restante de ataques. Ao final do processo de extração de quadros, 130000 imagens genuínas de 530000 imagens de ataque foram obtidas. O número de imagens neste banco é maior que todos os outros bancos combinados. Os dispositivos de captura deste banco são aparelhos móveis, e os ataques são feitos com impressões em papel, fotos e vídeos reproduzidos em displays.

3.2.5 Dataset Vsoft

O banco de imagens da Vsoft foi criado com a finalidade de ser utilizado para o desenvolvimento de um sistema focado em identificar casos reais. O banco de imagens é baseado em imagens similarmente ao banco NUAA, e possui cerca de 20000 imagens, e dividido em três partições: treino, validação e teste. Como as imagens foram obtidas de cenários práticos, o número de identidades diferentes é incerto. Os dispositivos de captura variam entre dispositivos móveis e webcams. Os ataques foram feitos com fotos impressas em papel comum e fotográfico, e a exibição de fotos em displays de dispositivos

móveis. Na tentativa de simular casos práticos, as imagens deste banco são extremamente não controladas, com diversas variações em iluminação, proximidade das pessoas até a câmera, resolução das imagens, dispositivos de captura, entre outros fatores. Neste banco de dados há aproximadamente 10% mais imagens reais que imagens de ataques.

3.3 Método Proposto

A proposta deste trabalho é o desenvolvimento de um modelo baseado em redes neurais convolucionais capaz de classificar corretamente imagens entre duas classes: Falsificações (ou ataques) e Reais. Para alcançar tal objetivo, diversos passos intermediários foram executados e estudados. As metodologias utilizadas, bem como a decisão de suas escolhas, serão descritas nas seções a seguir.

3.3.1 Transferência de aprendizado

A transferência de aprendizado é uma técnica utilizada em aprendizagem de máquina em que um modelo previamente treinado é utilizado como um ponto de partida para um modelo a ser treinado numa nova tarefa. Realizar o treinamento de uma rede neural convolucional do zero pode ser um processo longo. Dependendo da quantidade de dados a se utilizar, o treinamento de uma CNN robusta pode demorar dias.

Como relatado por Li et al. [31], abordagens anteriores treinavam modelos do zero, entretanto, tais abordagens dão margem a existência de um problema comum, o superajuste. Os bancos de imagens para o problema de anti-falsificação de face são muito consistentes, e muito bem controlados, podendo levar redes treinadas do zero a se superajustar e aprender essas características comuns, prejudicando sua precisão em cenários genéricos.

Outra possível solução para este problema, que não envolve a transferência de aprendizado, é criar redes muito simples, pois isso ajuda a reduzir a possibilidade de superajuste. Essa abordagem, entretanto, pode acarretar num modelo que não é capaz de discernir com tanta precisão. Para este trabalho, optou-se por seguir a abordagem de transferência de aprendizado numa rede com a arquitetura VGG16, após alguns experimentos preliminares com diferentes arquiteturas.

3.3.2 Arquitetura VGG16

A VGG16 [42] foi inicialmente proposta para o problema de classificação no banco de imagens ImageNet[17], que consiste de mais de 14 milhões de imagens divididas em 1000 classes. O trabalho original da VGG foi capaz de obter 92% de acurácia top-5.

O modelo demorou semanas para ser treinado utilizando a placa de vídeo Nvidea Titan Black. A arquitetura é composta de 6 blocos, sendo 5 convolucionais, que consistem de um par de convoluções seguidas de uma camada de subamostragem, e o último bloco sendo um bloco denso de 3 camadas, levando até a camada de classificação. A Figura 6 mostra a arquitetura da VGG16. A escolha da VGG16 se deu em experimentos preliminares onde foram avaliados os modelos MobilenetV2, Inception 3, VGG16, e VGG19, com a VGG16 obtendo os melhores resultados.

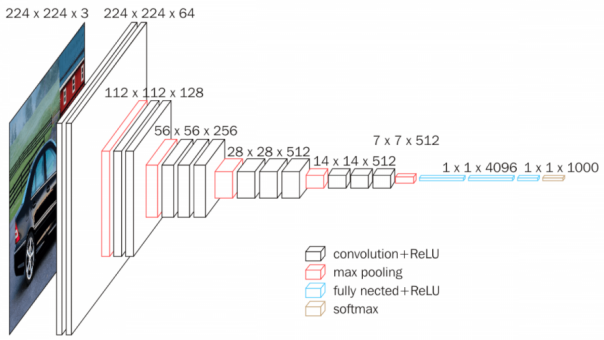


Figura 6: Arquitetura da VGG16. Fonte [46]

No caso deste trabalho, a última camada foi alterada para fornecer dois valores de saída. Esses valores estão no intervalo $[0, 1]$ e sua soma resulta em 1. Esses valores representam a confiança da predição da rede em cada classe (Ataque ou Genuíno).

3.4 Pré-processamento de dados

Nesta pesquisa quatro diferentes técnicas de pré-processamento foram empregadas: Recorte, Alinhamento, Subamostragem e Variação de Brilho. O intuito do pré-processamento é de modificar os dados de entrada, a fim de padronizar ou explorar cenários diferentes dos contidos nos bancos de imagens originais.

3.4.1 Recorte

A operação de recorte é utilizada para realçar regiões de interesse da imagem, por meio do descarte de outras regiões menos importantes. No domínio de anti-falsificação de face, isso é comumente realizado para restringir as imagens de entrada à face dos indivíduos, padronizando os dados recebidos pela rede. Com menos variações nas entradas, os filtros espaciais aprendidos pelas CNNs podem focar em aprender características voltadas para uma determinada região.

O primeiro pré-processamento realizado foi o recorte. Primeiramente foi utilizado um detector baseado no módulo Dlib disponível para Python. Em caso de imagens com faces que não foram detectadas, essas foram descartadas dos experimentos. Em seguida,

as faces foram recortadas das imagens originais conforme o resultado da detecção. Essa abordagem entretanto se mostrou problemática, pois devido às diferentes resoluções dos dispositivos, e à distancia das faces até a camera, algumas imagens acabavam por ficar ampliadas ou reduzidas, quando redimensionadas para a entrada da rede (224x224). Um exemplo de imagem de entrada da rede para esse experimento segue abaixo na Figura 7.

Figura 7: Exemplo de imagem de entrada para experimento do recorte. Fonte: Autoria Própria

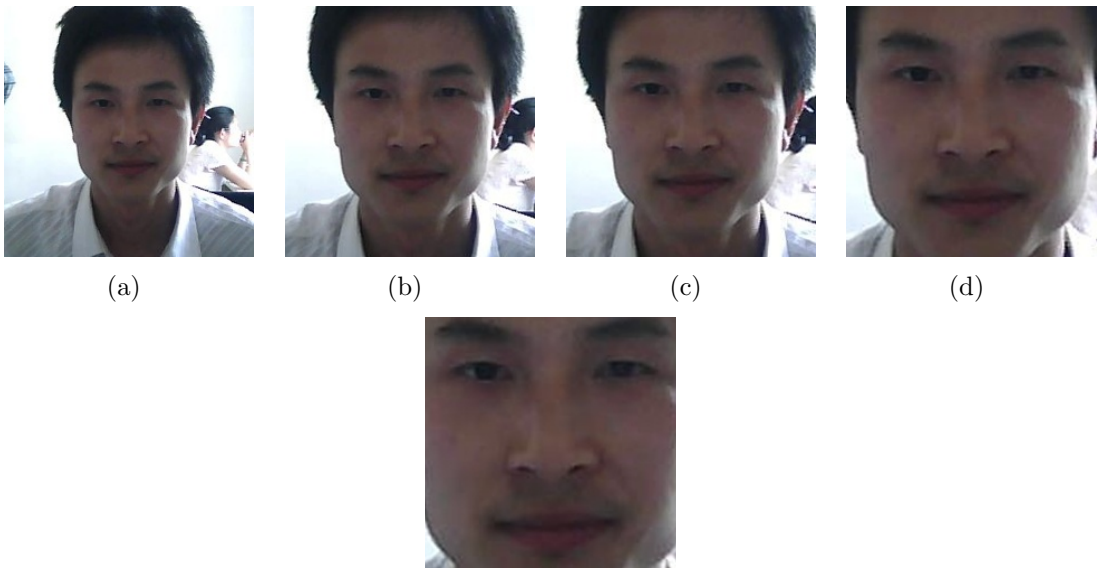
3.4.2 Alinhamento Geométrico

Em seguida o alinhamento geométrico das faces foi feito. Essa abordagem é mais robusta quando comparada apenas com o recorte. Ela envolve utilizar um ponto da imagem, aqui o ponto escolhido foi o ponto médio entre os dois olhos, e alinhar a face em relação a ele, resultando em uma imagem de tamanho conhecido, no caso 224x224. Isso ocorre com a utilização de operações de translação, rotação e escala.

Tanto o recorte quanto o alinhamento geométrico, entretanto, tem um ponto a ser considerado. A informação de fundo que é perdida pode ser relevante para o processo de identificação de um ataque. Isso acontece pois algumas imagens de ataque podem ser facilmente detectadas com base em artefatos como dedos de pessoas, e bordas de displays nas telas. Realizar essas operações apenas na região retangular da face pode então acarretar na perda de informação relevante para o problema.

Para estudar esse problema foram definidos 5 alinhamentos diferentes: 125, 100, 75, 67, 50 pixels entre os olhos. A distância em pixels entre os olhos é considerada para o alinhamento pois, uma vez que a imagem tem um tamanho fixo de 224x224, imagens onde a distância entre os olhos é maior tem menos espaço para as informações de fundo. Intuitivamente, quanto menor a distância entre os olhos em pixels, mais da imagem está

disponível para conter informações de fundo. A Figura 8 abaixo mostra um exemplo de imagem de cada alinhamento utilizado.



(a) (b) (c) (d)

(e)

Figura 8: Amostras de imagens em diferentes alinhamentos baseado na distancia entre os olhos medida em pixels : (a) 50 pixels; (b) 67 pixels; (c) 75 pixels; (d) 100 pixels; (e) 125 pixels. Fonte: Autoria Própria.

3.4.3 Subamostragem

Devido a alta taxa de quadros por segundo nos dispositivos de captura, a variação entre os quadros dos bancos de imagens baseadas em vídeos é muito pequena. O processo de extração de quadros sequenciais traz muita informação redundante, que além de não introduzir diversidade nos dados de treinamento da rede, desbalanceia os bancos de imagens. Isso pode implicar na CNN otimizando-se para acertar os resultados em um banco de imagens que constitui uma maior porcentagem da partição de treinamento, em detrimento de outro banco que não tem tantas imagens. A Tabela 1 mostra o número total de imagens para cada banco, bem como seu percentual em relação ao número total de imagens.

É possível observar que caso uma subamostragem nos dados de treinamento não seja empregada, a rede pode ficar enviesada em acertar apenas imagens de dois bancos mais proeminentes.

Para a fazer a subamostragem, uma maneira simples de realizar é extrair um quadro a cada 'X' quadros consecutivos, onde 'X' é um número arbitrário. Entretanto, essa abordagem não leva em consideração a variação entre os quadros, e pode acabar perdendo importantes variações.

Tabela 1: Tamanho de cada banco de imagens após o processo de extração de quadros.

Banco	Número de Imagens	% Do Total
NUAA	12614	1,16%
MSU-MFSD	77898	7,19%
Replay Attack	309998	28,63%
OULU-NPU	661432	61,09%
Vsoft	20718	1,91%

Outro método, o escolhido para esta pesquisa, é o de fazer uso de algum algoritmo que quantiza a diferença entre as imagens, e utilizar esse resultado para assim escolher as imagens mais diferentes. Nesse caso optou-se por utilizar o algoritmo de Medida do Índice de Similaridade Estrutural, ou SSIM (do inglês *Structural Similarity Index Measure*) [50]. O SSIM foi escolhido pois além de sua facilidade de implementação, ele leva em consideração as diferenças numa proximidade espacial na imagem, calculando a similaridade em diversas janelas menores.

Uma vez com o ranqueamento de quais imagens são as mais diferentes, são extraídas imagens de cada indivíduo a fim de não ocorrer desbalanceamento dentro de cada banco de imagens. Em seguida optou-se por manter 12000 imagens de cada banco, este valor é escolhido pois o menor banco disponível possui aproximadamente 12000 imagens. Assim as disparidades de tamanho entre os bancos devem diminuir. Vale notar que os bancos de imagens Vsoft e NUAA, por serem baseados em imagem e têm uma variedade maior do que os baseados em vídeos, e por serem os menores bancos de imagens, não foram sub-amostrados.

Como o único tipo de preprocessamento feito nesse experimento foi na quantidade de imagens, nenhuma modificação nas imagens foi feita, assim elas são idênticas às imagens originais dos bancos. Segue um exemplo de imagem na Figura 9.

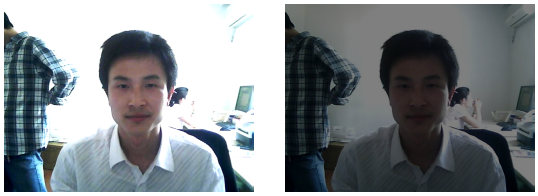
3.4.4 Variação de Brilho

A primeira análise de resultados mostrou que para o banco Vsoft muito erros ocorriam em imagens onde a iluminação é precária (muito escura ou muito clara). Como este banco se baseia em cenários práticos, as condições de iluminação variam muito. Para tentar combater esse problema, realizou-se um pré-processamento de variação de brilho. O

Figura 9: Exemplo de imagens de entrada variadas para experimento da subamostragem.

propósito dessa medida é fazer com que as imagens de treinamento e de validação incluam mais casos com diferentes iluminações, sem alterar o conjunto de testes, mantendo assim os resultados consistentes.

A variação de brilho é feita utilizando um brilho multiplicativo simples, que multiplica o valor da imagem por um número aleatório dentro de um intervalo. Para este trabalho, inicialmente o intervalo de variação de brilho escolhido foi [0,6-1,4]. A Figura 10 abaixo mostra a diferença de uma imagem com variação de brilho maior e menor que 1, respectivamente.



(a) (b)

Figura 10: Amostras de imagens com diferentes variações de brilho. (a) imagem com brilho aumentado; (b) imagem com brilho diminuído. Fonte: Autoria Própria.

3.5 Parâmetros de treinamento

Mesmo realizando a transferência de aprendizado, diversos hiperpârametros de treinamento precisam ser avaliados, como: Otimizador, Taxa de Aprendizagem, Tamanho de Batch. Além disso, também foi adicionada uma camada densa antes da camada de classificação, assim essas duas camadas serão as únicas da rede a serem treinadas do zero. Após experimentos preliminares, os parâmetros escolhidos seguem na Tabela 2.

Tabela 2: Hiperparâmetros de treinamento.

Hiperparametro	Valor
Otimizador	Adamax
Tamanho de Batch	16
Taxa de Aprendizagem	0,001
Unidades Densas	128
Dropout	0,3 Neurônios desativados
Epochs	10
Inicialização de Kernel	Glorot Uniforme
Função de Perda	Entropia Cruzada de Classificação

3.6 Métrica de avaliação

Os resultados dos experimentos deste trabalho foram medidos utilizando principalmente a métrica de Taxa de Erros Iguais, ou EER (do inglês *Equal Error Rate*. Essa medida é dada quando a taxa de Falsa Aceitação, ou FAR (do inglês *False Acceptance Rate*), e a taxa de Falsa Rejeição, ou FRR (do inglês *Fase Rejection Rate*), são iguais. A Figura 11 mostra uma representação gráfica do EER dado uma curva de FAR e uma curva de FRR.

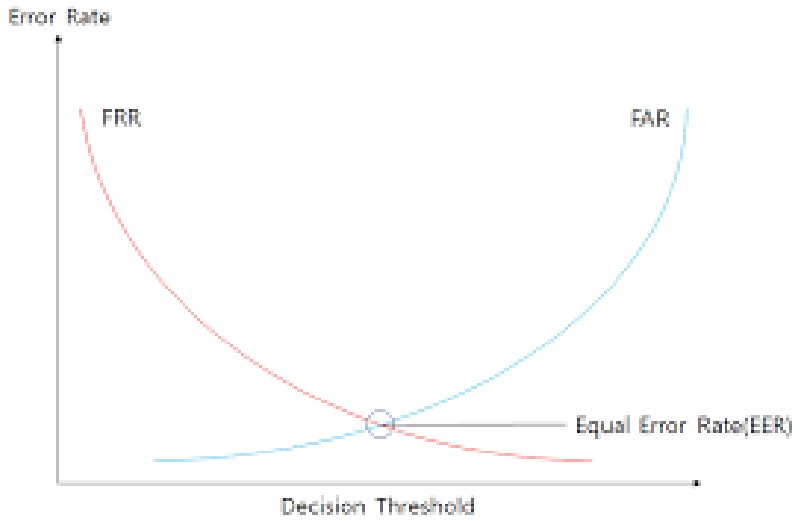


Figura 11: Representação gráfica do ponto de erro igual. A curva azul é a taxa de falsa aceitação, enquanto a vermelha é a taxa de falsa rejeição, variando para um limiar de decisão. Fonte: [11]

3.7 Visão Geral do Trabalho

A fim de melhorar esclarecer o fluxo desse trabalho, a Figura 12 ilustra a visão geral das etapas desenvolvidas.

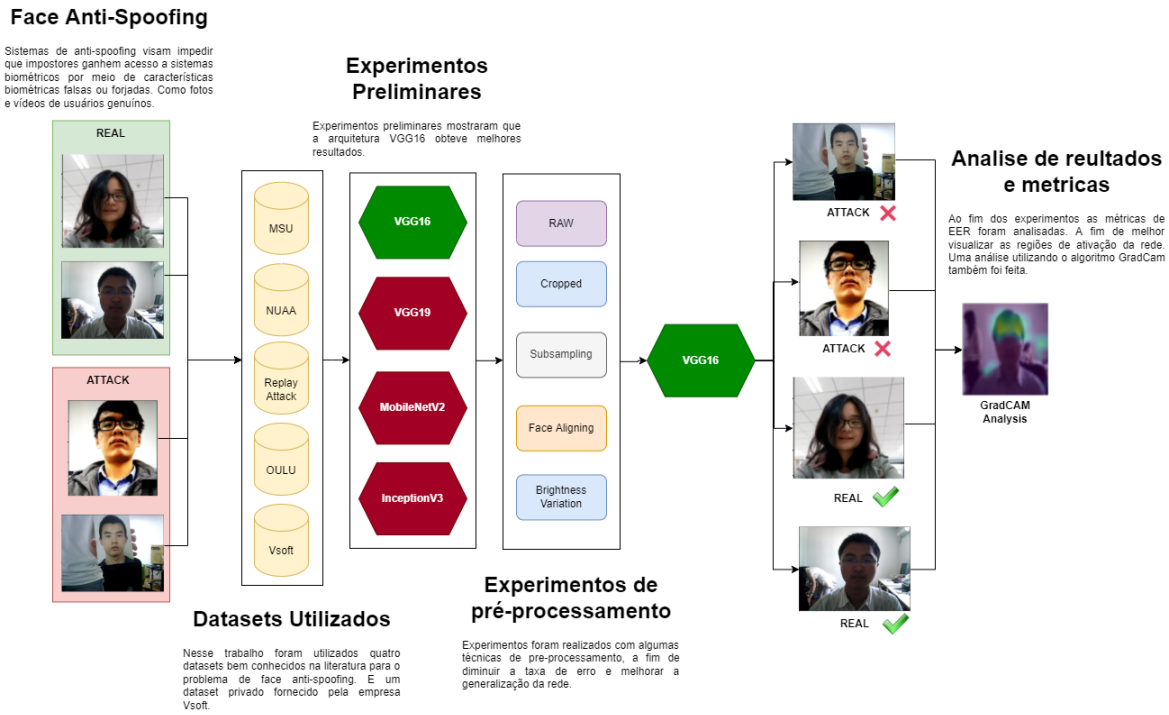


Figura 12: Visão geral do trabalho. Fonte: Autoria Própria.

4 APRESENTAÇÃO E ANÁLISE DOS RESULTADOS

Como dito na seção 3.3.2, experimentos preliminares mostraram que a arquitetura VGG16 obteve melhor resultado. A seguir serão mostrados os impactos de cada pré-processamento. Vale ressaltar novamente que nenhuma das imagens de teste foram utilizadas durante o processo de treinamento e validação.

4.1 Experimento do Recorte

O experimento do recorte foi utilizado para identificar se as informações presentes na face dos indivíduos são suficientes ou melhores do que a utilização das imagens completas. Os resultados de EER de ambos os cenários são mostrados na Tabela 3.

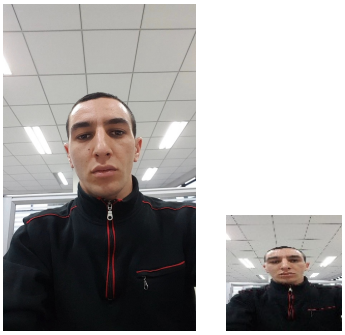
Tabela 3: Resultados para os experimentos entre imagens originais e banco com recorte das faces, com modelos treinados nos 3 bancos de imagens.

Banco de Teste	EER Original	EER Recorte
`NUAA`	$13,38\%$	$4,56\%$
`MSU`	$5,84\%$	$4,59\%$
`Replay Attack`	$4,06\%$	$5,70\%$
`OULU`	$8,59\%$	$14,47\%$
`Vsoft`	$2,13\%$	$7,08\%$

Como é possível visualizar, para os bancos de imagens Replay Attack, OULU e Vsoft os resultados para o modelo treinado nas imagens recortadas foi pior. Entretanto, para os outros dois bancos de imagens, o recorte foi benéfico. Isso pode ser um indício de um problema, pois como as imagens do Replay Attack e OULU são a grande maioria nesse experimento, a rede pode estar classificando as imagens apenas com base no fundo dessas imagens. Enquanto isso, seria um bom sinal para obter uma boa métrica em um banco específico da literatura, tal fator não se traduz em habilidade de generalização.

Quando o recorte ocorre, a rede é forçada a aprender características apenas da região da face, enquanto isso faz com que seu erro aumente em alguns bancos de imagens, essas características podem ser generalizadas para os outros bancos, resultando numa diminuição do erro. Por outro lado, é possível que as informações de fundo contenham características que facilitem a identificação de um ataque, o que pode levar a um aumento na taxa de erro, caso o recorte na face seja feito.

Entretanto, algo que foi percebido na análise de erros é que, ao redimensionar as imagens, distorções causadas por achatamentos em um dos sentidos eram criadas em muitas imagens, para o modelo treinado nas imagens originais. Isso ocorre pois a entrada da rede é 224x224 pixels, e as imagens originais têm diferentes resoluções e diferentes razões entre largura e altura. Isso pode ser visto na Figura 13.



(a) (b)

Figura 13: Exemplo de deformação ocorrido numa imagem original (a) causado pelo redimensionamento, gerando a imagem (b). Fonte: Autoria Própria.

Enquanto o recorte é uma medida que pode auxiliar nesse problema, uma outra análise de imagens identificou problemas com o processamento de imagens em diferentes distâncias da câmera. Faces muito distantes ocupam um espaço pequeno na imagem e precisam ser expandidas para 224x224. Esse processo diminui a qualidade da imagem. Tal problema é particularmente frequente no bando de imagens Replay Attack, pois este possui uma resolução muito baixa.

4.2 Experimento da Subamostragem

Em paralelo ao experimento anterior, foi analisado o impacto da subamostragem. Neste teste, pretende-se descobrir quanto a subamostragem afeta os resultados e analisar a relação de compromisso entre mais dados de treinamento, e maior generalização. Os resultados são mostrados na Tabela 4.

Como é possível analisar nos resultados, enquanto EER obtido foi mais alto no banco de imagens do OULU, os bancos menores que esse tiveram um EER mais baixo. Pode-se inferir que essa diferença nos resultados se dá pois agora que o OULU não corresponde a grande maioria das imagens de treino, a rede foi capaz de aprender características dos outros bancos, melhorando assim sua capacidade de generalização.

Dados os resultados deste experimento, passou-se a adotar a estratégia de utilizar bancos de imagens sub-amostrados ao longo deste trabalho.

Tabela 4: Resultados para os experimentos entre imagens originais e banco sub-amostrado, com modelos treinados nos 3 bancos de imagens.

Banco de Teste	EER Original	EER Subamostragem
NUAA	13,38%	4,09%
MSU	5,84%	2,86%
Replay Attack	4,06%	2,81%
OULU	8,59%	12,68%
Vsoft	2,13%	1,69%

4.3 Experimentos com alinhamento

Tendo visto que os experimentos com o recorte promoveram alguns pontos de discussão, os experimentos de alinhamento geométrico da face buscaram analisar os resultados de um método de pré-processamento mais refinado. Nesse experimento os modelos foram treinados e testados em um único banco individualmente. Os resultados podem ser vistos na Tabela 5 abaixo.

Tabela 5: Resultados para os experimentos entre imagens originais e banco alinhado em diferentes distancias entre os olhos.

Banco de Teste	EER-125	EER-100	EER-75	EER-67	EER-50
NUAA	0,97%	2,15%	3,09%	1,81%	0,18%
MSU	9,43%	9,33%	4,95%	13,6%	19,09%
Replay Attack	3,87%	4,71%	7,32%	7,19%	11,72%
OULU	8,33%	8,18%	14,41%	13,65%	18,59%
Vsoft	5,88%	6,64%	6,15%	4,87%	3,25%

Esses resultados mostram que cada banco de imagens tem suas particularidades em relação a qual alinhamento é melhor. Sendo assim, não foi possível encontrar um alinhamento único que funcione melhor para todos os bancos. Os bancos Replay Attack e OULU tiveram uma taxa de erro menor nas imagens onde a distância entre os olhos era maior. Já para os bancos de imagens Vsoft e NUAA, imagens que tinham uma menor distância entre os olhos, e consequentemente mais informação de fundo, promoveram a

menor taxa de erro.

4.4 Experimentos de variação de brilho

A fim de tentar introduzir uma maior variedade de cenários no treinamento, a variação de brilho foi realizada. Com imagens mais desafiadoras (iluminação mais variada) na partição de treino, espera-se que a rede tenha uma maior facilidade em classificar casos mais comuns. A Tabela 6 mostra os resultados para o experimento de variação de brilho, e a comparação com o melhor resultado do teste de alinhamento.

Tabela 6: Resultados para os experimentos entre imagens alinhadas e com variações de brilho no treinamento.

Banco de Teste	EER Variação de Brilho	EER Melhor Alinhamento
`NUAA`	$0,36\%$	$0,18\%$
`MSU`	$14,03\%$	$4,95\%$
`Replay Attack`	$4,41\%$	$3,87\%$
`OULU`	$9,68\%$	$8,33\%$
`Vsoft`	$3,45\%$	$3,25\%$

É possível notar que nenhum banco apresentou uma melhoria nos resultados, e isso pode ser o indício de um comportamento problemático. Uma vez que a introdução de uma maior variedade de imagens no treinamento causa uma perda de precisão do modelo, isso pode significar que as imagens de teste não refletem a diversidade encontrada no treinamento, e consequentemente, não representam bem os cenários genéricos do mundo real.

4.5 Análise de Imagens

Além de realizar a análise de métricas, também foram observados os resultados da passagem do Grad-CAM [41] sobre as imagens corretas e de erros. Abaixo segue uma breve análise sobre alguns tópicos de discussão.

4.5.1 NUAA

Como mostrado na sessão anterior, o resultado neste banco de imagens apresentou o menor EER ($0,18\%$), ainda assim, a análise das imagens provê algumas possíveis repostas

para os poucos erros encontrados. Primeiramente as Figuras 14 e 15 mostram o resultado do Grad-CAM para algumas amostras de acertos da rede.

Figura 14: Análise do Grad-CAM de algumas amostras reais corretamente classificadas no banco NUAA. Fonte: Autoria Própria.

É possível observar que enquanto nas imagens de pessoas reais as ativações da rede focam muito mais nas regiões da face, as imagens de ataque seguem um padrão diferente. Devido ao fato de que os ataques são por meio de fotos em papel, as regiões de ativação da rede são mais próximas às bordas, e fortemente visíveis nos dedos das pessoas segurando as fotos. Isso também relaciona a menor distância entre os olhos com o bom resultado, pois, uma menor distância entre os olhos implica mais fundo visível. Nas Figuras 16 e 17 segue a análise das imagens erroneamente classificadas.

É possível notar que para as imagens de ataque, apesar de que a rede ativou majoritariamente regiões ao redor da face, essas ativações não foram o suficiente para classificar a imagem como um ataque. Também vale notar que nessas imagens não aparecem dedos segurando as fotos das pessoas. Em contrapartida, as imagens reais que foram classificadas como ataque são de uma mulher que posicionou os dedos na borda da foto capturada, o que pode ter levado a uma confusão por parte da rede.

Figura 15: Análise do Grad-CAM de algumas amostras de ataque corretamente classificadas no banco NUAA. Fonte: Autoria Própria.

Figura 16: Análise do Grad-CAM das amostras de ataque erroneamente classificadas no banco NUAA. Fonte: Autoria Própria.

Figura 17: Análise do Grad-CAM das amostras reais erroneamente classificadas no banco NUAA. Fonte: Autoria Própria.

4.5.2 MSU

A análise de erros do banco de imagens do MSU mostrou um comportamento diferente do banco NUAA. Como é possível ver nas Figuras 18 e 19, que mostram os acertos da rede, as imagens de pessoas reais ativam majoritariamente as regiões de fundo, enquanto as imagens de ataque ativam as faces das pessoas.

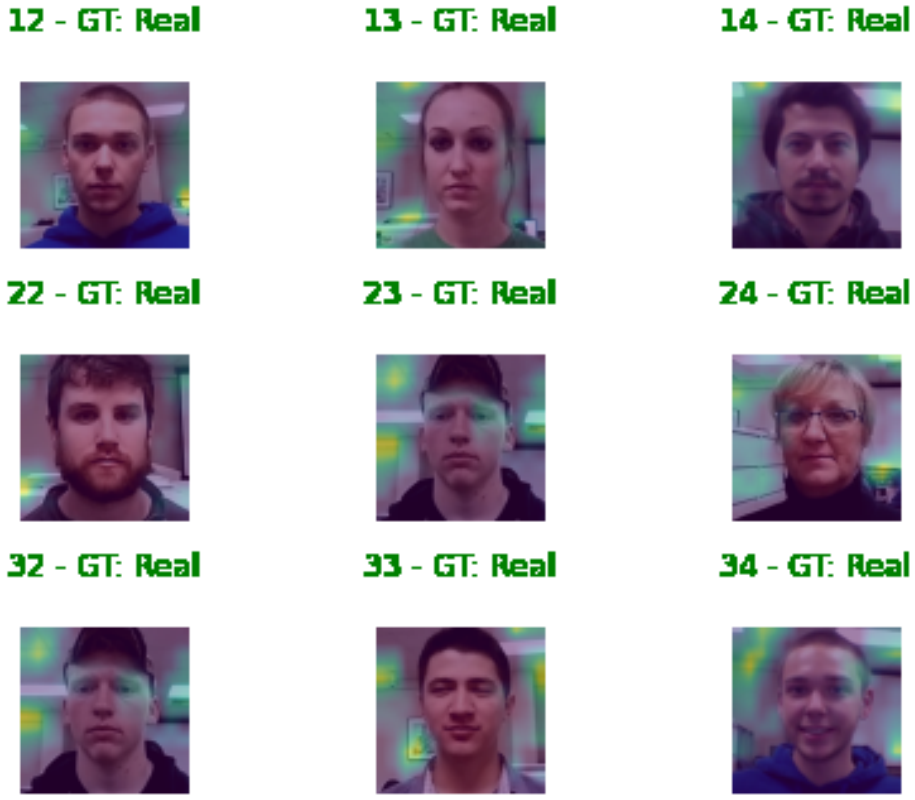


Figura 18: Análise do Grad-CAM de algumas amostras reais corretamente classificadas no banco MSU. Fonte: Autoria Própria.

É possível ainda notar que para as imagens de pessoas reais, a ativação da rede parece maior em objetos que podem melhor refletir a iluminação do ambiente como a tela branca do projetor, ou emissores de luz como as lâmpadas do teto. Esse comportamento também é visualizado nas imagens que a rede não acertou, porém como são erros, o comportamento é o inverso, evidenciado nas Figuras 20 e 21.

Como pode ser observado, em várias das imagens de ataque, regiões da face foram ativadas, mas não foi o suficiente para discernir essas imagens como ataques. Igualmente, as imagens reais ativaram as regiões ao redor da face, mas com menos intensidade do que as que foram corretamente classificadas.

Figura 19: Análise do Grad-CAM de algumas amostras de ataque corretamente classificadas no banco MSU. Fonte: Autoria Própria.

4.5.3 OULU

O resultado da análise do banco de imagens OULU, em geral, é muito similar a do banco MSU. É possível ver que para as imagens reais que a rede acerta, as regiões que causam ativação na rede são as de fundo, ligeiramente ao redor da face, enquanto para os ataques, a face está sendo o ponto mais focado pela rede, como pode ser visto nas Figuras 22 e 23.

Apesar de mostrar resultados parecidos, é possível notar que enquanto as imagens de pessoas reais focam em ativações ao redor da face, essas em sua maioria ocorrem na parte de cima, como nas sobrancelhas e testa. Já nas imagens de ataque, a maior parte das ativações aparenta estar na região do queixo e bochechas da face. Para as imagens em que houve erro por parte da rede, o comportamento também é similar, mas com as regiões ativadas invertidas, como mostrado abaixo nas Figuras 24 e 25.

Novamente, algumas imagens ainda possuem traços de ativações em regiões que são consistentes com as regiões ativadas em imagens classificadas corretamente, mas essas ativações não foram o suficiente para classificar uma imagem de acordo com o gabarito. É possível notar que nas imagens reais, apesar de ter a região do queixo ativada, como as imagens de ataque, essas regiões se estendem e muitas vezes chegam na parte central da

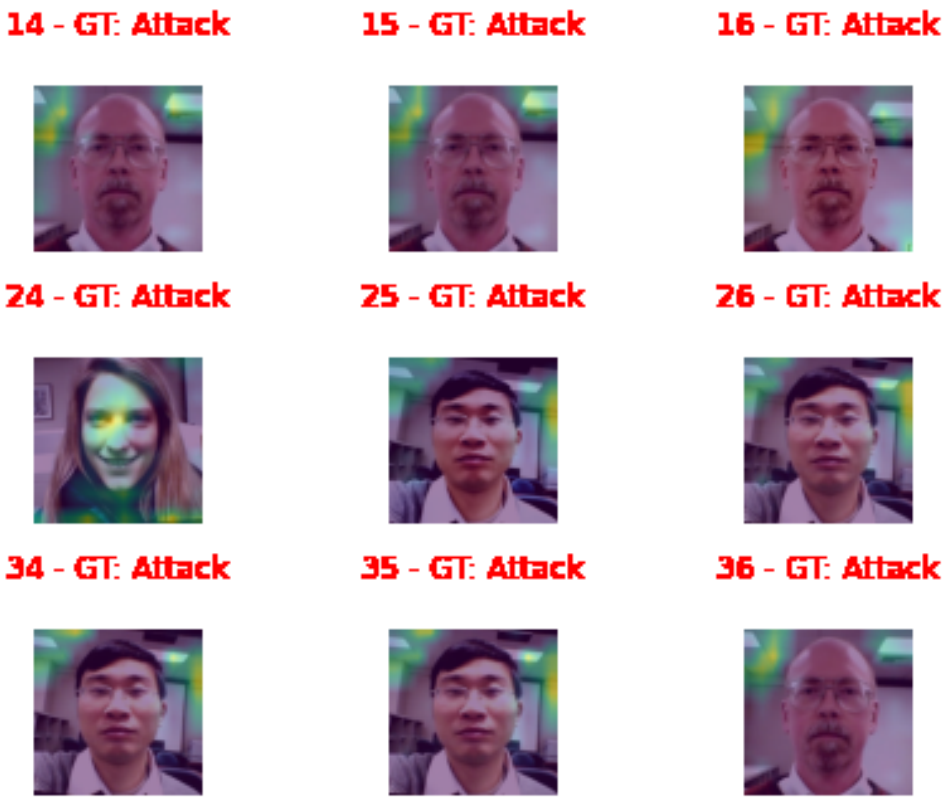


Figura 20: Análise do Grad-CAM de algumas amostras de ataque erroneamente classificadas no banco MSU. Fonte: Autoria Própria.

testa.

4.5.4 Replay Attack

Diferentemente dos outros bancos, o melhor resultado do Replay Attack foi no banco, que menos mostrava informação de fundo, assim todas as regiões de importância que geraram uma ativação nas camadas da rede estão na face da pessoa. As Figuras 26 e 27 mostram o resultado do Grad-CAM.

Nas imagens reais é possível notar que as regiões de ativação se espalham pela face, em alguns casos elas se concentram em uma linha horizontal da região da boca, mas em geral, diversas regiões da face ativam as camadas da rede. Para as imagens de ataque, por outro lado, as regiões de ativação estão concentradas nas áreas do nariz, abaixo e acima dos olhos. Uma possível hipótese que explica esse comportamento é o fato que essas regiões tendem a possuir maior oleosidade na face, o que contribui para reflexos. Nas Figuras 28 e 29 seguem as imagens de análise em alguns dos erros da rede.

Notavelmente as imagens de pessoas reais têm regiões de ativação mais concentradas nas partes que ativam as regiões das imagens de ataque corretamente classificadas. Em adição a isso, as imagens de ataques classificadas como reais não possuem, em sua

Figura 21: Análise do Grad-CAM de algumas amostras de reais erroneamente classificadas no banco MSU. Fonte: Autoria Própria.

maioria, regiões de ativação no nariz e ao redor dos olhos.

4.5.5 Vsoft

Nesse banco de imagens, o melhor resultado foi do experimento que possuía mais informações de fundo. As Figuras 30 e 31 mostram o resultado do Grad-CAM no banco.

É possível ver que nos acertos da rede para ambas as classes, Ataque ou Real, as regiões de ativação tendem a se localizar ao redor da face, mais especificamente nas regiões da testa, camisa e fundo nas imagens reais. Já para as imagens impostoras, a rede possui uma maior ativação nas bordas de displays. Vale notar que, em diversas imagens de ataque corretamente avaliadas, a rede não teve muitas ativações. Abaixo, nas Figuras 32 e 33, seguem as imagens de análise em alguns dos erros da rede.

Para as imagens reais que a rede classifica como ataques, é possível ver que as faces estão próximas da câmera, ou que a rede teve ativações na parte inferior da face dos indivíduos, como nariz e boca. Já para as imagens de ataque que foram dadas como reais, pode se notar os padrões de ativações nos fundos ou nas testas das faces.

Figura 22: Análise do Grad-CAM de algumas amostras reais corretamente classificadas no banco OULU. Fonte: Autoria Própria.

4.6 Comparações com outros trabalhos

A fim de melhor representar os resultados obtidos nesse trabalho, a Tabela 7 mostra a comparação de EER entre alguns trabalhos da literatura e os melhores resultados deste trabalho obtidos em cada banco.

O banco de imagens Vsoft não possui comparações pois este é um banco proprietário e não está disponível em literatura. É possível notar que enquanto os resultados desse trabalho não foram estado da arte em todos os bancos, eles ainda foram competitivamente próximos aos resultados encontrados na literatura.

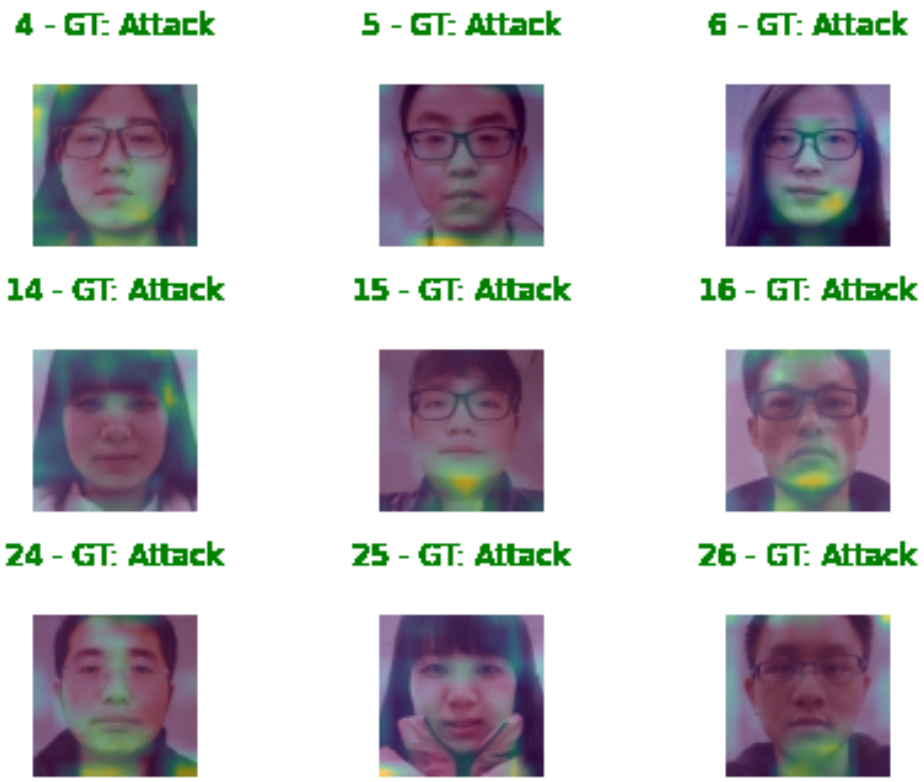


Figura 23: Análise do Grad-CAM de algumas amostras de ataque corretamente classificadas no banco OULU. Fonte: Autoria Própria.

Figura 24: Análise do Grad-CAM de algumas amostras de ataque erroneamente classificadas no banco OULU. Fonte: Autoria Própria.

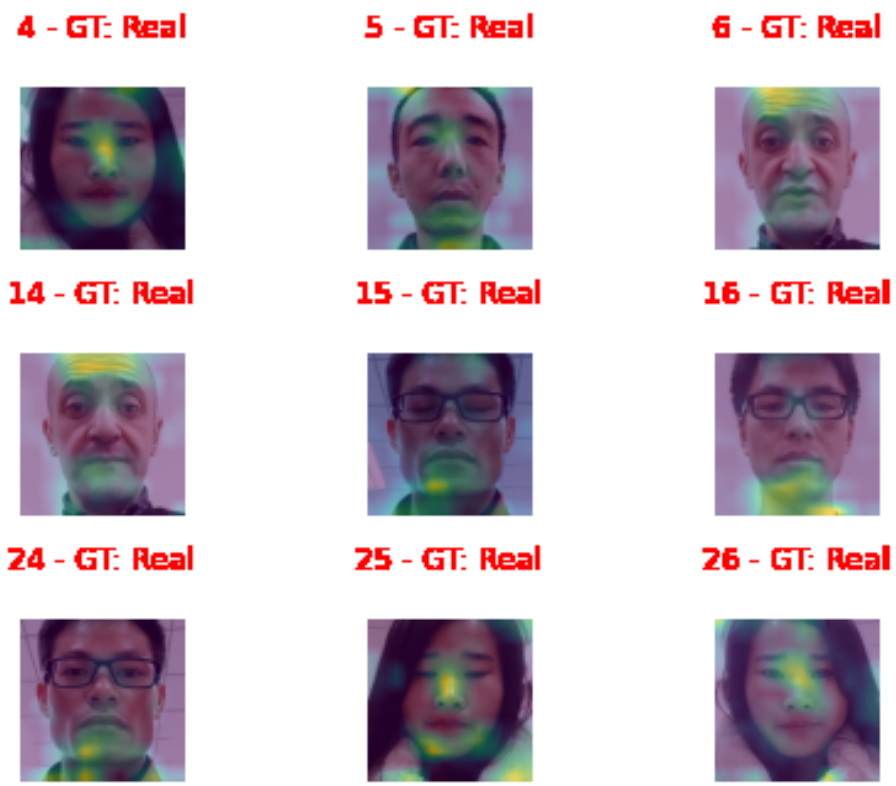


Figura 25: Análise do Grad-CAM de algumas amostras de reais erroneamente classificadas no banco OULU. Fonte: Autoria Própria.

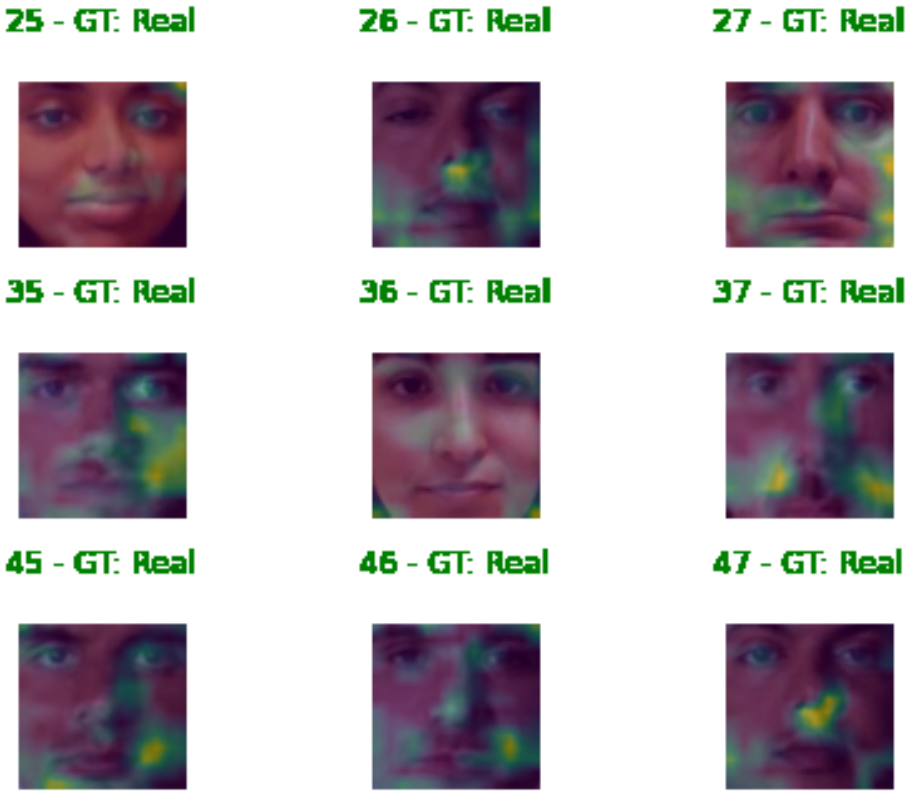


Figura 26: Análise do Grad-CAM de algumas amostras reais corretamente classificadas no banco Replay Attack. Fonte: Autoria Própria.

Figura 27: Análise do Grad-CAM de algumas amostras de ataque corretamente classificadas no banco Replay Attack. Fonte: Autoria Própria.

Figura 28: Análise do Grad-CAM de algumas amostras de ataque erroneamente classificadas no banco Replay Attack. Fonte: Autoria Própria.

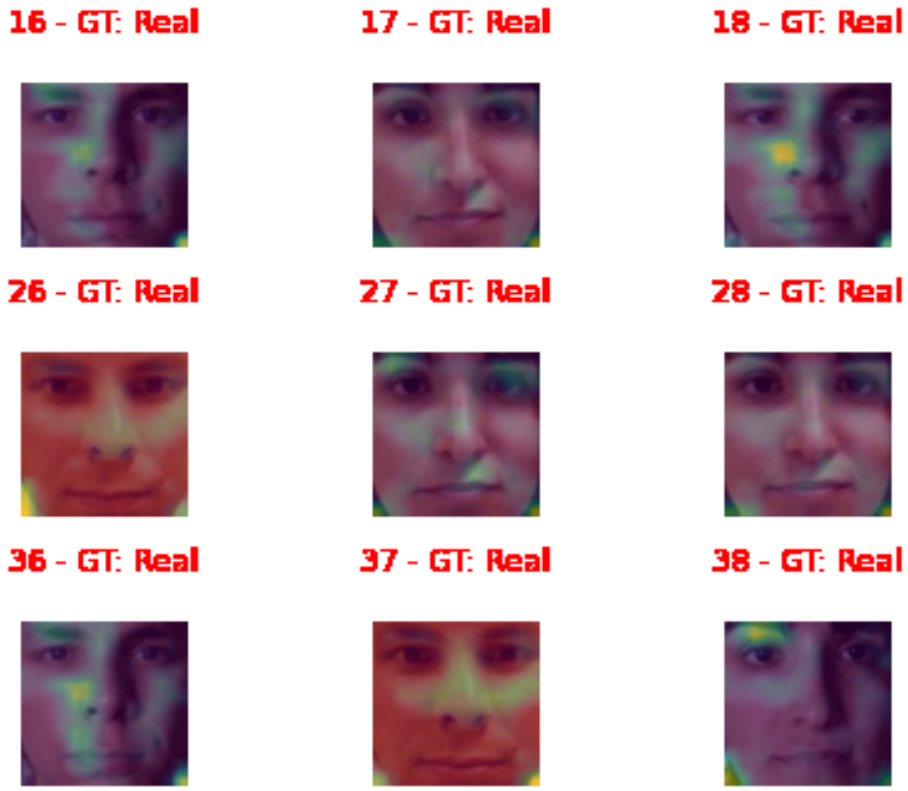


Figura 29: Análise do Grad-CAM de algumas amostras de reais erroneamente classificadas no banco Replay Attack. Fonte: Autoria Própria.

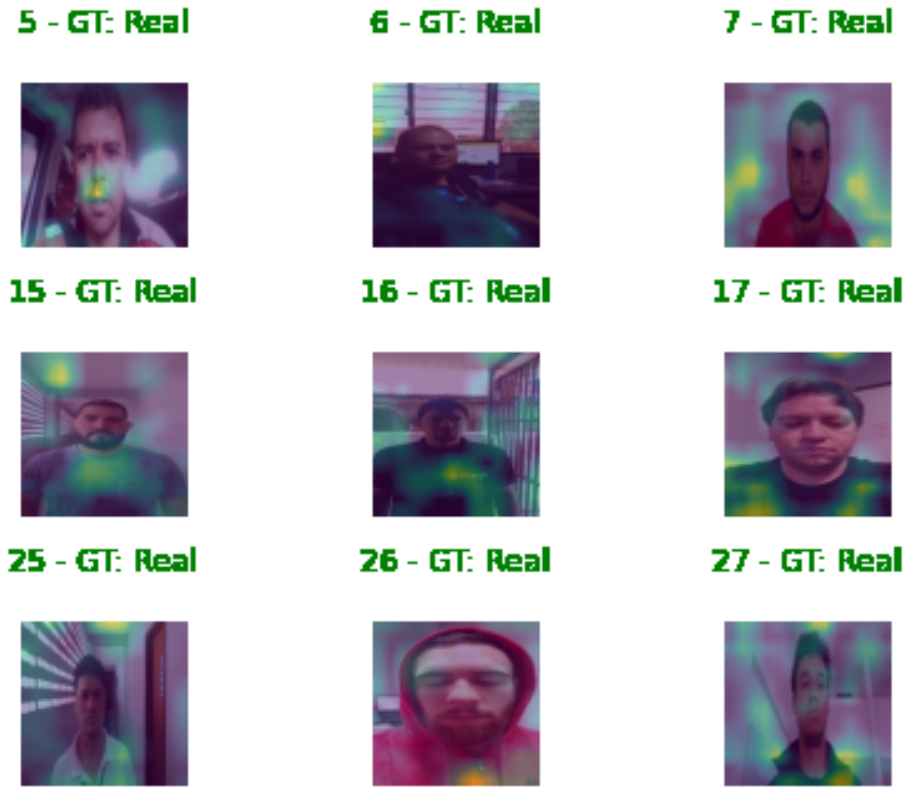


Figura 30: Análise do Grad-CAM de algumas amostras reais corretamente classificadas no banco Vsoft. Fonte: Autoria Própria.

Figura 31: Análise do Grad-CAM de algumas amostras de ataque corretamente classificadas no banco Vsoft. Fonte: Autoria Própria.

Figura 32: Análise do Grad-CAM de algumas amostras de ataque erroneamente classificadas no banco Vsoft. Fonte: Autoria Própria.

Figura 33: Análise do Grad-CAM de algumas amostras de reais erroneamente classificadas no banco Vsoft. Fonte: Autoria Própria.

Tabela 7: Comparação de resultados com outros trabalhos para os bancos de literatura utilizando a menor taxa de erro no experimento de alinhamento.

Dataset	Method	EER%
NUAA	LBP+Gabor Wavelets+HOG [32]	1,10%
NUAA	LBP+LPQ+HOG [53]	1,90%
NUAA	MLPQ-TOP [4]	1,10%
NUAA	**Esse Trabalho**	0,18%
Replay Attack	Fine-Tuned-VGGFace [30]	8,40%
Replay Attack	DPCNN [30]	2,90%
Replay Attack	Patch Based CNN [5]	2,50%
Replay Attack	**Esse Trabalho**	2,81%
MSU	Color LBP [9]	10,80%
MSU	GFA-CNN [45]	7,50%
MSU	**Esse Trabalho**	2,86%
OULU-NPU	DeepPixBis [20]	6,00%
OULU-NPU	FaceDs [23]	4,30%
OULU-NPU	CPqD [8]	6,90%
OULU-NPU	**Esse Trabalho**	8,33%
Vsoft	**Esse Trabalho**	3,25%

5 CONCLUSÕES E TRABALHOS FUTUROS

Este trabalho explorou algumas formas de pre-processamento de imagens para o problema de detecção de ataques em biometria facial, e analisou como essas diferenças impactaram a taxa de erro do modelo. Os melhores pre-processamentos testados variaram para cada banco de imagens utilizado, sendo assim, não foi possível concluir que algum dos pre-processamentos testados é o melhor para todas as situações. Ainda assim, mesmo que os resultados obtidos não tenham sido estado da arte em todos os bancos, eles foram competitivos. Os objetivos do trabalho foram alcançados, os bancos mais utilizados em literatura foram analisados, o modelo de classificação de imagens que faz a distinção entre uma imagem real ou de ataque foi criado, os impactos de diversos tipos de pre-processamento foram experimentados e os resultados foram analisados com os devidos questionamentos levantados.

Alguns comentários adicionais podem ser feitos em relação a dois bancos de dados que tiveram resultados mais destoantes entre si. O primeiro deles é o banco NUAA, em que este trabalho obteve um EER menor que resultados da literatura. É possível supor que um fator que influenciou nesses resultados é o fato de que os ataques deste banco de imagens são compostos primariamente de pessoas segurando imagens com as faces utilizadas nos ataques em papel. Além de um ataque simples, não foi tomado um cuidado por parte dos criadores desse banco de imagens em esconder as bordas do papel, nem as mãos das pessoas que estão realizando o ataque. Sendo assim, não é errado dizer que o banco de imagens NUAA possui os ataques mais simples de todos os bancos de imagens explorados nesse trabalho. Outro fator que pode ter influenciado no bom resultado do NUAA em alguns experimentos é que ele não sofreu nenhuma subamostragem, por já ser o banco com menor número de imagens, e não ser baseado em vídeos.

Por outro lado, os experimentos com o banco de imagens OULU obtiveram resultados piores que os trabalhos mostrados na literatura. Pode-se inferir que a subamostragem do banco atuou como um fator de importância nesse resultado, uma vez que o banco de imagens, em seu tamanho original, é o que possui mais imagens. As partições de treino e validação foram reduzidas para menos de 10% de seu valor original. Além disso, esse banco já possui uma maior variedade de ataques, incluindo o uso de *displays* de aparelhos móveis e *tablets* com telas em alta resolução.

Como trabalhos futuros pode-se destacar um estudo mais aprofundado sobre a capacidade de generalização dos bancos e de outras arquiteturas de rede neural. É possível que modelos menores que a VGG16 permitam mais generalização. Também pode-se realizar experimentos com arquiteturas treinadas do zero, ou seja, sem a utilização de transferência de aprendizado. Além disso, mais estudos em relação a quantidade de imagens utilizadas no experimento de subamostragem podem se provar eficientes, como

por exemplo uma redução baseada num percentual de cada banco original, ao invés de uma subamostragem para um valor fixo.

Outra linha de pesquisa que pode ser estudada é a utilização de fluxos de vídeo, ao invés de realizar uma classificação utilizando uma única imagem, pois utilizando uma sequência de quadros de um vídeo, é possível utilizar outras técnicas que fazem uso da estimação de um mapa de profundidade da face. Entretanto essa técnica não poderia ser utilizada em todas as situações em que o modelo apresentado neste trabalho poderia, uma vez que não seria aplicável para uma única imagem.

Ainda é possível tentar outras técnicas de pré-processamento para filtrar imagens de qualidade baixa, atribuindo um terceiro rótulo, além de Ataque ou Real. Essa técnica é empregada por alguns sistemas de detecção de ataques em biometria facial comerciais, que quando recebem uma imagem de baixa qualidade, impedem a entrada do usuário sem necessariamente declarar que uma tentativa de ataque foi efetuada. Para fazer isso, é necessário antes definir o que seria uma imagem de qualidade baixa, e como quantizar essa medida de qualidade. Alguns trabalhos já foram iniciados nessa linha de pesquisa, mas não envolvem os bancos utilizados neste trabalho.

REFERÊNCIAS

[1] Digital image processing basics, 2 2018.

[2] Matlab — rgb image representation, 6 2018.

[3] JK Aggarwal and N Nandhakumar. On the computation of motion from sequences of images-a review. *Proceedings of the IEEE*, 76(8):917–935, 1988.

[4] Shervin Rahimzadeh Arashloo, Josef Kittler, and William Christmas. Face spoofing detection based on multiple descriptor fusion using multiscale dynamic binarized statistical image features. *IEEE Transactions on Information Forensics and Security*, 10(11):2396–2407, 2015.

[5] Yousef Atoum, Yaojie Liu, Amin Jourabloo, and Xiaoming Liu. Face anti-spoofing using patch and depth-based cnns. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pages 319–328. IEEE, 2017.

[6] Ali Azarbayejani, Thad Starner, Bradley Horowitz, and Alex Pentland. Visually controlled graphics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(6):602–605, 1993.

[7] Samarth Bharadwaj, Tejas I Dhamecha, Mayank Vatsa, and Richa Singh. Computationally efficient face spoofing detection with motion magnification. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 105–110, 2013.

[8] Zinelabdine Boulkenafet, Jukka Komulainen, Zahid Akhtar, Azeddine Benlamoudi, Djamel Samai, Salah Eddine Bekhouche, Abdelkrim Ouafi, Fadi Dornaika, Abdelmalik Taleb-Ahmed, Le Qin, et al. A competition on generalized software-based face presentation attack detection in mobile scenarios. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pages 688–696. IEEE, 2017.

[9] Zinelabidine Boulkenafet, Jukka Komulainen, and Abdenour Hadid. Face anti-spoofing based on color texture analysis. In *2015 IEEE international conference on image processing (ICIP)*, pages 2636–2640. IEEE, 2015.

[10] Ivana Chingovska, André Anjos, and Sébastien Marcel. On the effectiveness of local binary patterns in face anti-spoofing. In *2012 BIOSIG-proceedings of the international conference of biometrics special interest group (BIOSIG)*, pages 1–7. IEEE, 2012.

[11] Maro Choi, Shincheol Lee, Minjae Jo, and Ji Sun Shin. Keystroke dynamics-based authentication using unique keypad. *Sensors*, 21(6):2242, 2021.

[12] Allan da Silva Pinto, Helio Pedrini, William Schwartz, and Anderson Rocha. Video-based face spoofing detection through visual rhythm analysis. In *2012 25th SIBGRAPI Conference on Graphics, Patterns and Images*, pages 221–228. IEEE, 2012.

[13] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. Ieee, 2005.

[14] Ken H Davis, R Biddulph, and Stephen Balashek. Automatic recognition of spoken digits. *The Journal of the Acoustical Society of America*, 24:637–642, 1952.

[15] Tiago de Freitas Pereira, André Anjos, José Mario De Martino, and Sébastien Marcel. Lbp- top based countermeasure against face spoofing attacks. In *Asian Conference on Computer Vision*, pages 121–132. Springer, 2012.

[16] Tiago de Freitas Pereira, André Anjos, José Mario De Martino, and Sébastien Marcel. Can face anti-spoofing countermeasures work in a real world scenario? pages 1–8, 2013.

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[18] Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[19] IBM Cloud Education. Neural networks, 8 2020.

[20] Anjith George and Sébastien Marcel. Deep pixel-wise binary supervision for face presentation attack detection. In *2019 International Conference on Biometrics (ICB)*, pages 1–8. IEEE, 2019.

[21] Rafael C Gonzalez and Richard E Woods. Digital image processing, hoboken, 2018.

[22] Robert William Gainer Hunt. *The reproduction of colour*, volume 4. Wiley Online Library, 1995.

[23] Amin Jourabloo, Yaojie Liu, and Xiaoming Liu. Face de-spoofing: Anti-spoofing via noise modeling. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11217 LNCS:297–315, 2018.

[24] Michael David Kelly. *Visual identification of people by computer*. Department of Computer Science, Stanford University., 1970.

[25] Klaus Kollreider, Hartwig Fronthaler, and Josef Bigun. Evaluating liveness by face images and the structure tensor. In *Fourth IEEE Workshop on Automatic Identification Advanced Technologies (AutoID'05)*, pages 75–80. IEEE, 2005.

[26] Klaus Kollreider, Hartwig Fronthaler, Maycel Isaac Faraj, and Josef Bigun. Real-time face detection and motion analysis with application in "liveness" assessment. *IEEE Transactions on Information Forensics and Security*, 2(3):548–558, 2007.

[27] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[28] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[29] Jiangwei Li, Yunhong Wang, Tieniu Tan, and Anil K Jain. Live face detection based on the analysis of fourier spectra. In *Biometric technology for human identification*, volume 5404, pages 296–303. International Society for Optics and Photonics, 2004.

[30] Lei Li, Xiaoyi Feng, Zinelabidine Boulkenafet, Zhaoqiang Xia, Mingming Li, and Abdenour Hadid. An original face anti-spoofing approach using partial convolutional neural network. In *2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–6. IEEE, 2016.

[31] Lei Li, Xiaoyi Feng, Zinelabidine Boulkenafet, Zhaoqiang Xia, Mingming Li, and Abdenour Hadid. An original face anti-spoofing approach using partial convolutional neural network. *2016 6th International Conference on Image Processing Theory, Tools and Applications, IPTA 2016*, 2017.

[32] Jukka Määttä, Abdenour Hadid, and Matti Pietikäinen. Face spoofing detection from single images using texture and local shape analysis. *IET biometrics*, 1(1):3–10, 2012.

[33] Bangalore S Manjunath and Wei-Ying Ma. Texture features for browsing and retrieval of image data. *IEEE Transactions on pattern analysis and machine intelligence*, 18(8):837–842, 1996.

[34] Zuheng Ming, Muriel Visani, Muhammad Muzzamil Luqman, and Jean Christophe Burie. A survey on anti-spoofing methods for face recognition with rgb cameras of generic consumer devices. *arXiv*, 2020.

[35] Pranab Mohanty, Sudeep Sarkar, and Rangachar Kasturi. From scores to face templates: a model-based approach. *IEEE transactions on pattern analysis and machine intelligence*, 29:2065–2078, 2007.

[36] Timo Ojala, Matti Pietikainen, and Topi Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 24(7):971–987, 2002.

[37] Robson José de Oliveira. Uso de redes neurais artificiais na avaliação funcional de estradas florestais. 2008.

[38] Gang Pan, Lin Sun, Zhaohui Wu, and Shihong Lao. Eyeblink-based anti-spoofing in face recognition from a generic webcamera. In *2007 IEEE 11th international conference on computer vision*, pages 1–8. IEEE, 2007.

[39] Keyurkumar Patel, Hu Han, and Anil K Jain. Cross-database face antispoofing with robust feature representation. In *Chinese Conference on Biometric Recognition*, pages 611–619. Springer, 2016.

[40] Bruno Peixoto, Carolina Michelassi, and Anderson Rocha. Face liveness detection under bad illumination conditions. In *2011 18th IEEE International Conference on Image Processing*, pages 3557–3560. IEEE, 2011.

[41] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, Oct 2019.

[42] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[43] Xiaoyang Tan, Yi Li, Jun Liu, and Lin Jiang. Face liveness detection from a single image with sparse low rank bilinear discriminative model. In *European Conference on Computer Vision*, pages 504–517. Springer, 2010.

[44] Pitchaya Thipkham. Image processing class (egbe443) 1.2 — digital image, 11 2018.

[45] Xiaoguang Tu, Zheng Ma, Jian Zhao, Guodong Du, Mei Xie, and Jiashi Feng. Learning generalizable and identity-discriminative representations for face anti-spoofing. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(5):1–19, 2020.

[46] Muneeb ul Hassan. Vgg16 – convolutional network for classification and detection, 11 2018.

[47] CS231 Stanford University. Cs231n convolutional neural networks for visual recognition.

[48] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

[49] Tao Wang, Jianwei Yang, Zhen Lei, Shengcai Liao, and Stan Z Li. Face liveness detection using 3d structure recovered from a single camera. In *2013 international conference on biometrics (ICB)*, pages 1–6. IEEE, 2013.

[50] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[51] Di Wen, Hu Han, and Anil K Jain. Face spoof detection with image distortion analysis. *IEEE Transactions on Information Forensics and Security*, 10(4):746–761, 2015.

[52] Jianwei Yang, Zhen Lei, and Stan Z Li. Learn convolutional neural network for face anti-spoofing. 6 2014.

[53] Jianwei Yang, Zhen Lei, Shengcai Liao, and Stan Z Li. Face liveness detection with component dependent descriptor. In *2013 International Conference on Biometrics (ICB)*, pages 1–6. IEEE, 2013.

[54] Andreas Zell. *Simulation neuronaler netze*, volume 1. Addison-Wesley Bonn, 1994.

[55] Boulkenafet Zinelabidine, Komulainen Jukka, Lei Li, Xiao Feng, and Abdenour Hadid. Oulunpu: a mobile face presentation attack database with real-world variations. In *Proc. IEEE Int. Conf. on Identity, Security and Behavior Analysis, ISBA*, pages 1–7, 2017.