Universidade Federal da Bahia
Instituto de Matemática

Programa de Pós-Graduação em Ciência da Computação

REMOÇÃO DE EXPRESSÕES FACIAIS EM IMAGENS 3D PARA FINS DE RECONHECIMENTO BIOMÉTRICO

Lucas Amparo Barbosa

DISSERTAÇÃO DE MESTRADO

Salvador
17 de outubro de 2018

LUCAS AMPARO BARBOSA

REMOÇÃO DE EXPRESSÕES FACIAIS EM IMAGENS 3D PARA FINS DE RECONHECIMENTO BIOMÉTRICO

Esta Dissertação de Mestrado foi apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal da Bahia, como requisito parcial para obtenção do grau de Mestre em Ciência da Computação.

Orientador: Prof. Dr. Maurício Pamplona Segundo

Salvador
17 de outubro de 2018

Ficha catalográfica elaborada pelo Sistema Universitário de Bibliotecas (SIBI/UFBA),
com os dados fornecidos pelo(a) autor(a).

Barbosa, Lucas Amparo
Remoção de Expressões Faciais em Imagens 3D para fins de reconhecimento biométrico / Lucas Amparo Barbosa. -- Salvador, 2018.
27 f. : il

Orientador: Maurício Pamplona Segundo.
Dissertação (Mestrado - Mestrado em Ciência da Computação) -- Universidade Federal da Bahia, Universidade Federal da Bahia, 2018.

1. Deep Learning. 2. Reconhecimento Facial. 3. Imagens 3D. I. Pamplona Segundo, Maurício. II. Título.

TERMO DE APROVAÇÃO

LUCAS AMPARO BARBOSA

REMOÇÃO DE EXPRESSÕES FACIAIS EM IMAGENS 3D PARA FINS DE RECONHECIMENTO BIOMÉTRICO

Esta Dissertação de Mestrado foi julgada adequada à obtenção do título de Mestre em Ciência da Computação e aprovada em sua forma final pelo Programa de Pós-Graduação em Ciência da Computação da Universidade Federal da Bahia.

Salvador, 17 de OUTUBRO de 2018

Prof. Dr. Maurício Pamplona Segundo
Universidade Federal da Bahia

Prof. Dr. Luciano Rebouças de Oliveira
Universidade Federal da Bahia

Prof. Dr. Alexandre da Costa e Silva Franco
Instituto Federal de Educação, Ciência e Tecnologia da Bahia

Dedico este escrito ao meu grupo de pesquisa. Do caos, criou-se a ordem. E da zoeira, fez-se a ciência. Obrigado a todos.

AGRADECIMENTOS

Aos meus pais, Tony e Ilma, pelo apoio incondicional;

A Deus, por ter me tornado capaz de superar esse desafio;

A minha linda namorada, Thamires, por ouvir minhas lamúrias e ser o meu apoio mais próximo quando foi necessário;

Ao meu orientador, Prof. Maurício, pela paciência e pela tranquilidade na transmissão dos conhecimentos;

Ao meu caríssimo colega Gabriel Dahia, pelo auxílio nas horas do aperto e, às vezes, desespero.

Aos laços que fiz aqui, que perdurarão por longos anos;

Aos laços que trouxe comigo, que continuem sempre ao meu lado.

Gostaria de agradecer também a Fapesb pela bolsa concedida e a UFBA, em especial o laboratório iVision, pela infraestrutura disponibilizada para minha pesquisa.

A persistência é o menor caminho do êxito

—CHARLES CHAPLIN (Desconhecido)

RESUMO

A pesquisa realizada apresenta um modelo de rede neural *encoder-decoder* para remover deformações causadas por expressões faciais em imagens 3D. Este modelo recebe uma imagem 3D da face com ou sem expressões como entrada e gera uma face neutra como saída. O objetivo não é obter um resultado realístico e sim melhorar a precisão de sistemas de reconhecimento facial 3D. Para realizar isso, foi proposto o uso de uma função de custo baseada em um sistema de reconhecimento durante o processo de treinamento para que a rede aprendesse a manter informações inerentes à identidade do indivíduo na saída. Os experimentos usando a base de dados Bosphorus 3D mostraram que a técnica foi bem sucedida em reduzir a diferença entre imagens do mesmo indivíduo afetadas por diferentes expressões faciais e ampliar a distância entre os valores das intraclasses e interclasses. Eles também mostram que nossas imagens neutras geradas sinteticamente melhoram os resultados de quatro métodos de reconhecimento, atingindo assim o objetivo original.

Palavras-chave: *Deep Learning.* Reconhecimento Facial. Imagens 3D.

ABSTRACT

We present an encoder-decoder neural network to remove deformations caused by expressions from 3D face images. It receives a 3D face with or without expressions as input and outputs its neutral form. Our objective is not to obtain the most realistic results but to enhance the accuracy of 3D face recognition systems. To this end, we propose using a recognition-based loss function during training so that our network can learn to maintain important identity cues in the output. Our experiments using the Bosphorus 3D Face Database show that our approach successfully reduces the difference between face images from the same subject affected by different expressions and increases the gap between intraclass and interclass difference values. They also show that our synthetic neutral images improved the results of four different well-known face recognition methods, thus accomplishing the original objective.

Keywords: Deep Learning. Facial Recognition. 3D Images.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

ASM Active Shape Models . . . 2
CMC Cumulative Match Characteristic . . . 18
CNN Convolutional Neural Network . . . 2
EER Equal Error Rate . . . 18
GAN Generative Adversarial Network . . . 23
GPU Graphics Processor Unit . . . 13
ICP Iterative Closest Points . . . 7
LBP Local Binary Pattern . . . 17
LBPH Local Binary Pattern Histogram . . . 17
LDA Linear Discriminant Analysis . . . 17
OpenCV Open Source Computer Vision Library . . . 13
PCA Principal Components Analysis . . . 17
PCL PointCloud Library . . . 13
PDF Probability Density Function . . . 16
RAM Random Access Memory . . . 13
RBF Radial Basis Function . . . 2
RMSE Root Mean Square Error . . . 16
ROC Receiver Operating Characteristic . . . 18

Capítulo

1

INTRODUÇÃO

O rosto é amplamente utilizado como fonte de informação para reconhecer indivíduos, seja em um contexto real — alguém procurando outra pessoa na multidão, por exemplo — ou virtual — um sistema automatizado de reconhecimento facial. Esse processo é não-intrusivo e pode ter uma variedade de utilizações, desde sistemas simples que utilizem uma aquisição controlada até os de larga-escala baseados em câmeras de segurança (ZHAO et al., 2003). A tarefa de reconhecimento pode ser feita através de diferentes métodos (TURK; PENTLAND, 1991; BELHUMEUR et al., 1997; SCHROFF et al., 2015; AHONEN et al., 2004), utilizando imagens 2D (cor e/ou infravermelho) e/ou 3D. Independente da modalidade de aquisição, todas elas compartilham um problema: humanos usam expressões faciais para comunicação, que modifica a forma do rosto, tornando-o consideravelmente diferente da sua respectiva versão neutra (FRITH, 2009). As diferenças causadas pelas deformações podem ser vistas na Figura 1.1. Isto pode fazer com que duas capturas faciais de pessoas diferentes sejam mais parecidas para um sistema de reconhecimento do que duas imagens da mesma pessoa com expressões diferentes, e isso pode ser visto na Figura 1.2.

Figura 1.1 Exemplo de diferentes expressões faciais de um mesmo indivíduo. Fonte: BillionPhotos.com (bit.ly/2OVLPu8).

O caminho mais direto para solucionar o problema é focar em regiões da face que são menos afetadas pelas expressões, como as áreas ao redor dos olhos (CAMPOS et al., 2000)

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

Figura 1.2 Demonstração da diferença entre imagens por meio de mapas de calor. As imagens (a) e (c) são do mesmo indivíduo e a diferença entre elas é mostrada na imagem (b). As imagens (c) e (e) são de pessoas diferentes, e a imagem (d) apresenta a diferença entre elas. Como pode ser observado, expressões faciais podem tornar duas imagens do mesmo indivíduo menos similares do que duas imagens neutras de indivíduos diferentes.

ou do nariz (CHANG et al., 2006; EMAMBAKHSH; EVANS, 2017). Uma generalização dessa ideia pressupõe que uma pequena vizinhança em torno de qualquer ponto da face seja localmente rígida e, consequentemente, menos propensa a variações de expressão, o que permite a criação de abordagens que utilizam comparações baseadas em partes do rosto (LI et al., 2015; ELAIWAT et al., 2014). Criar uma representação invariante as expressões também é uma alternativa, que encontrou algum sucesso quando utilizada em imagens 3D (BERRETTI et al., 2010; KAKADIARIS et al., 2007). Nenhuma dessas soluções, porém, procura entender o efeito das variações causadas pelas expressões faciais, pois são essencialmente mecanismos para evitar o problema.

Lidar com expressões significa modelar deformações até certo ponto que permita a simulação ou remoção das mesmas. Simulação de expressões é mais indicada para sistemas controlados em que existe a garantia de faces neutras cadastradas. Assim, uma face cadastrada pode ser deformada de acordo com um modelo pré-computado para ser comparada com uma imagem de entrada com variação de expressões (LU; JAIN, 2008). Enquanto isso, a remoção da expressão não tem esse requisito, podendo eliminar as deformações causadas pelas expressões faciais de todas as imagens antes de compará-las entre si. Isso é uma grande vantagem sobre a primeira opção quando considerado um cenário de identificação, em que uma imagem de entrada é comparada contra várias imagens cadastradas, porque a galeria pode ser pré-processada para a remoção de expressões faciais, mas não para a simulação das mesmas.

Como expressões são, em sua maioria, mudanças na forma do rosto, os estudos iniciais focaram em imagens 3D. Pan et al. (2010) aprenderam como inferir o resíduo da expressão de uma imagem não-neutra usando um modelo de regressão baseado em Radial Basis Function (RBF). Com essa abordagem, eles passaram a subtrair o resíduo da imagem de entrada para reconstruir a forma neutra. Agianpuye e Minoi (2014) eliminaram a necessidade desta segunda etapa modelando a identidade e o resíduo causado por expressões em conjunto através de um modelo deformável baseado em Active Shape Models (ASM).

Com a ascensão das Redes Neurais de Convolução (Convolutional Neural Network (CNN)), estudos recentes foram bem-sucedidos em utilizar aprendizado generativo ad-

versarial para obter resultados realísticos (muito próximos das imagens reais) para essa tarefa em imagens 2D (DING et al., 2018; SONG et al., 2017; DING et al., 2018). Mesmo assim, os mais recentes trabalhos baseados em CNN usando imagens 3D focam na tarefa de reconhecer tipos de expressões (JAN et al., 2018; YANG; YIN, 2017), mas não no reconhecimento facial robusto a expressões.

Neste trabalho, foi proposto o uso de CNN para remover expressões faciais de imagens 3D especificamente para o fins de reconhecimento. Realismo nas imagens de saída não é um requisito, uma vez que a precisão do reconhecimento é o principal alvo. Para alcançar esse objetivo:

- apresentamos um modelo de rede neural adaptado do trabalho de Badrinarayanan, Kendall e Cipolla (2015) para executar o mapeamento de uma imagem facial 3D não-neutra para a sua versão neutra correspondente;
- propusemos, como principal contribuição, usar uma função de correção baseada em reconhecimento para regularizar o processo de treino buscando manter ou aumentar a discriminalidade durante a remoção;
- comparamos o efeito desta remoção na similaridade das faces 3D com o estado-da-arte;
- testamos os resultados da remoção de expressões em métodos consagrados de reconhecimento facial para mostrar o potencial de melhora na precisão destes sistemas.

Capítulo 2

MÉTODO PROPOSTO

Neste trabalho propomos o uso de uma CNN para, dada uma imagem facial 3D com ou sem expressões, se obtenha a face neutra do indivíduo como resultado. Considerando os avanços no aprendizado de máquina para visão computacional usando representações 2D (RADFORD et al., 2015; KRIZHEVSKY et al., 2012; BADRINARAYANAN; KENDALL; CIPOLLA, 2015; LIU et al., 2018), optamos por evitar representações volumétricas e usar projeções ortogonais 2D das imagens 3D de face. Essa escolha também traz outros benefícios: considerável redução no número de parâmetros do modelo; manutenção das formas da imagem 3D original; e manutenção da vi

Para realizar o objetivo mencionado acima, foi executada uma normalização da pose das faces 3D alinhando-as com um modelo de face média. Então, foram projetadas no espaço 2D e utilizada a rede neural para remover qualquer deformação causada por expressões faciais. Esse processo é ilustrado pela Figura 2.1. Detalhes sobre cada uma das etapas são fornecidos nas Seções 2.1-2.3.

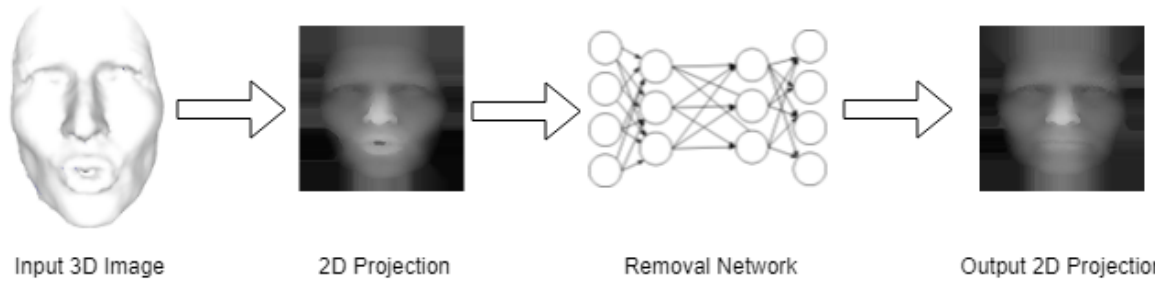

Figura 2.1 Diagrama do *pipeline* proposto para remoção de expressões faciais. As faces 3D de entrada são convertidas em projeções 2D, que são utilizadas para alimentar uma rede *encoder-decoder* para produzir a imagem neutra correspondente.

2.1 AQUISIÇÃO

A base de dados Bosphorus 3D Face Database (SAVRAN et al., 2008) (a partir de agora, simplesmente chamada de Bosphorus) foi utilizada como fonte de imagens 3D para este trabalho. A Bosphorus contém imagens faciais de 105 indivíduos, onde um terço delas são atores profissionais. As imagens foram capturadas em um sistema de luz estruturada e tem uma média de 36 mil pontos. Foram utilizadas apenas imagens praticamente frontais, totalizando 299 faces neutras e 2189 com variações de expressão. Exemplos dessas imagens podem ser vistos na Figura 2.2.

Para executar uma avaliação justa da abordagem proposta, a base foi dividida em conjuntos de treino, validação e teste de modo que um indivíduo só esteja presente em um único conjunto. Para isso, foram selecionados aleatoriamente metade dos indivíduos para treino (53 pessoas), um quarto para validação (26 pessoas) e um quarto para teste (26 pessoas). Para avaliar o poder de generalização da rede, essa separação foi feita de quatro formas diferentes e todas elas submetidas ao processo de treinamento e avaliação do reconhecimento. Em todos os casos, um indivíduo só está presente em um único conjunto.

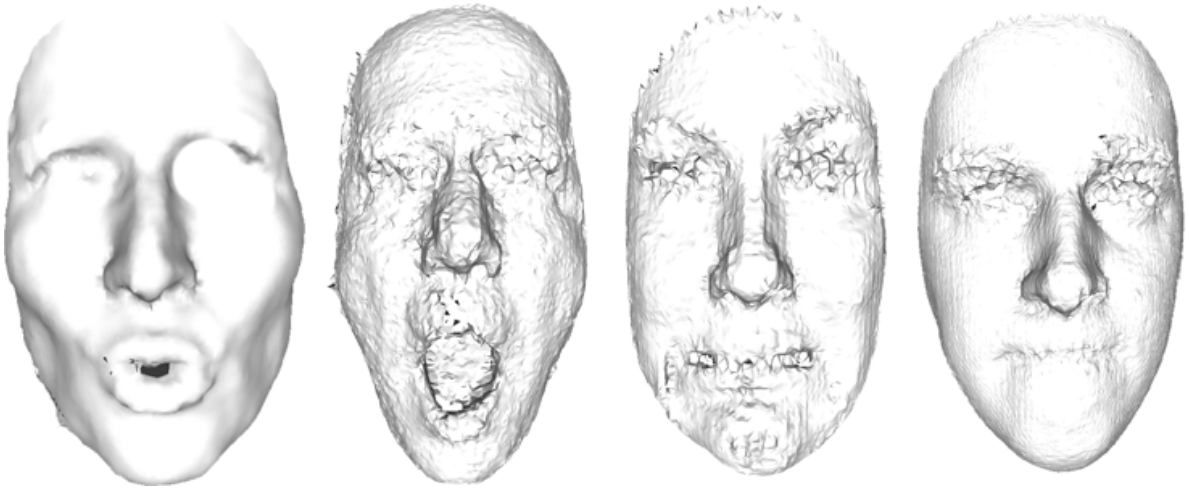

Figura 2.2 Exemplos de imagens faciais da Bosphorus apresentando diferentes expressões faciais. Diferentes artefatos de captura podem ser observados, como pontas, buracos e ondulações.

2.2 NORMALIZAÇÃO

O processo de normalização das imagens é extremamente importante, pois diminui variações na imagem causadas pela pose do indivíduo. Para um melhor treinamento da rede neural, quanto menos variações que não sejam causadas por expressões estiverem presentes nas imagens de treino, mais robusto é o mapeamento que permite a remoção das deformações causadas por expressões faciais presentes na imagem de entrada.

2.2.1 Modelo de face média

Para minimizar a variação entre os exemplos de treinamento, precisamos que estes estejam alinhados em um mesmo sistema de coordenadas. Para executar esse alinhamento, precisamos de uma referência para a posição esperada. Para isso criamos um modelo médio da face, que foi gerado a partir de todas as imagens neutras do conjunto de treinamento da Bosphorus seguindo os passos definidos no Algoritmo 1 e ilustrados na Figura 2.3.

Algoritmo 1: Pseudo-código do cálculo da face média

Entrada: Conjunto C de imagens neutras de treino
Saída: Face média M
Escolha aleatoriamente uma imagem I $\in$ C;
Alinhe as imagens restantes em C com a imagem I utilizando o algoritmo Iterative Closest Points (ICP);
para *cada ponto* $P \in I$ **faça**
    Encontre o ponto mais próximo de P em cada imagem restante de C;
    Calcule a média desses pontos;
    Adicione esse ponto médio à imagem M;
fim
Calcule o centro de massa de M;
Translade o centro de massa até a origem.

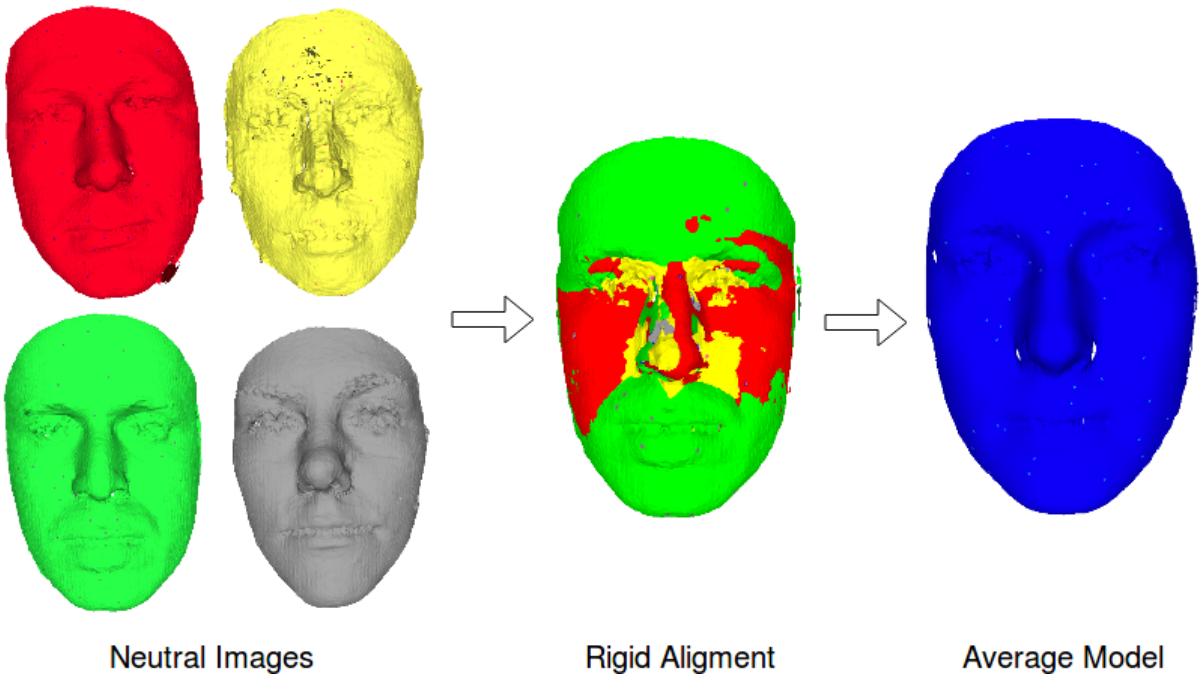

Figura 2.3 Modelo de face média construído após o alinhamento das faces neutras do conjunto de treino e cálculo da média das mesmas. A mistura de cores na segunda etapa representa o alinhamento das faces

2.2.2 Alinhamento rígido e projeção 2D

Para normalizar a pose da imagem 3D de entrada, alinhamos a mesma ao modelo de face média utilizando o ICP (BESL; MCKAY, 1992). Após o alinhamento, a imagem 3D é mapeada em uma projeção ortogonal 2D de dimensões 128×128 aplicando-se as seguintes equações para cada ponto 3D x_i, y_i, z_i:

$$r = \lfloor 64 + y_i \rceil \tag{2.1}$$

$$c = \lfloor 64 + x_i \rceil \tag{2.2}$$

$$I(r, c) = 127 + 3z_i \tag{2.3}$$

onde r e c são as linhas e colunas do pixel onde o ponto será projetado, $\lfloor a \rceil$ é o inteiro mais próximo de a, e $I(r, c)$ é a intensidade do pixel na imagem projetada I.

Por fim, os buracos presentes na projeção (pixels sem valor definido) são preenchidos com base nos valores dos vizinhos mais próximos. A Figura 2.4 apresenta um exemplo dos passos do processo e seus resultados.

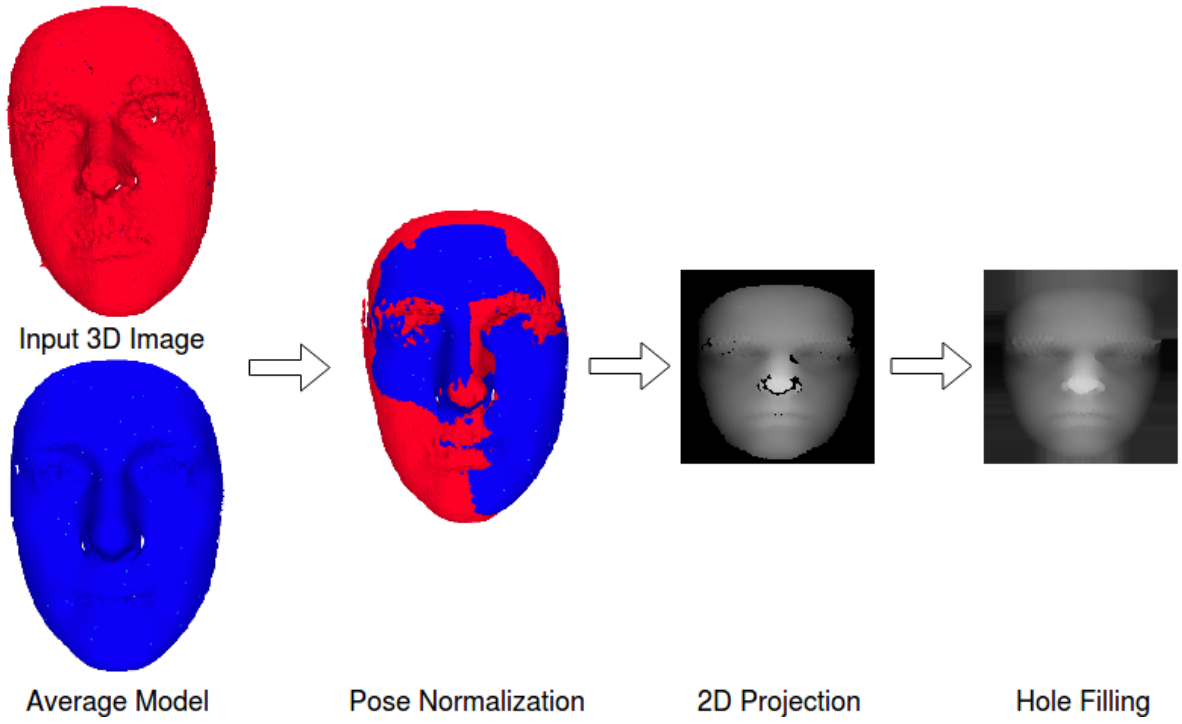

Figura 2.4 Projeção ortogonal da face 3D em uma imagem 2D. A face de entrada é alinhada a um modelo de face média pré-computado para padronizar a pose. Após isso, as coordenadas X e Y são mapeadas em valores de linha e coluna, respectivamente, enquanto o eixo Z é representado através da intensidade do pixel. Pixels vazios na projeção resultante são preenchidos através dos valores de seus vizinhos válidos mais próximos.

2.3 MODELO DE REDE NEURAL

O modelo proposto é uma adaptação da arquitetura proposta por Badrinarayanan, Kendall e Cipolla (2015), consistindo em um *encoder-decoder* convolucional para executar a regressão de uma projeção ortogonal 2D de uma face 3D de entrada em outra projeção correspondendo a mesma face sem expressões faciais. Essa arquitetura alcançou algum sucesso na literatura executando tarefas similares, como remoção de ruídos de imagens (PATHIRAGE et al., 2015) e segmentação semântica (BADRINARAYANAN; KENDALL; CIPOLLA, 2015).

A rede é constituída por duas partes:

- Encoder: O *encoder*, ou codificador, é a primeira parte da rede. Composto por dois blocos de duas camadas convolucionais e uma camada de *max-pooling* e três camadas convolucionais posteriores a estes blocos, tem como objetivo transformar a imagem de entrada em um conjunto de valores que armazenem as características do indivíduo com o menor impacto das expressões e com a maior capacidade de descrição possível. A entrada desta camada é criada pelas etapas anteriores. Sua saída é um vetor de características em um espaço latente com dimensões $4 \times 4 \times 64$. Se comparado com a arquitetura proposta por Badrinarayanan *et al.* (BADRINARAYANAN; KENDALL; CIPOLLA, 2015), a última camada de convolução da proposta era, originalmente, uma camada de *max-pooling*. Essa modificação foi feita para evitar uma redução exagerada do vetor de características.

- Decoder: O *decoder*, ou decodificador, é simétrico ao codificador, substituindo o *max-pooling* por um processo de *upsampling* e as matrizes de convoluções pelas suas transpostas. Outra diferença entre a arquitetura de Badrinarayanan, Kendall e Cipolla (2015) e esta rede é a ausência de função de ativação na última camada da rede, pois a ativação linear nos permite treinar o modelo com a função de correção L_2 (GOODFELLOW; BENGIO; COURVILLE, 2016).

A descrição completa da rede, com detalhes de sua configuração, se encontra na Tabela 2.1.

Com isso em mente, o problema é definido como a otimização dos parâmetros da rede θ, minimizando a função de custo L_2 computada entre a saída da rede $\hat{\mathbf{y}}$ e a imagem neutra correspondente $\mathbf{y}$. Também foi considerada uma função de custo alternativa para ajudar a manter a identidade da imagem de entrada. Para isso, foi utilizado a distância Euclidiana entre os descritores extraídos por uma implementação pública inspirada na Facenet (SCHROFF et al., 2015), uma CNN treinada para imagens 2D e considerada o estado-da-arte. Por simplicidade, desde ponto em diante chamaremos esta CNN de Facenet e a função de correção alternativa de *recognition loss*. Mais especificamente, o *recognition loss* é definido como:

$$\mathcal{L}_r = \alpha \|h(\hat{\mathbf{y}}) - h(\mathbf{y})\| + \|h(\hat{\mathbf{y}}) - h(\mathbf{x})\| \tag{2.4}$$

onde $h(\mathbf{a})$ é o descritor construído pela Facenet para $\mathbf{a}$.

Tabela 2.1 Arquitetura da Rede Neural para remoção de Expressões Faciais. Ela recebe como entrada uma imagem 128x128 criada a partir de uma projeção ortogonal da face com ou sem expressões e tem como saída outra projeção da face sintética neutra de mesmas dimensões.

	#	Type	Input	Filter	Stride	Output
Encoder	1	Convolutional + LReLU	128×128×1	7×7×1×64	2	64×64×64
	2	Convolutional + LReLU	64×64×64	7×7×64×64	1	64×64×64
	3	Max Pooling	64×64×64	2×2	2	32×32×64
	4	Convolutional + LReLU	32×32×64	7×7×64×64	2	16×16×64
	5	Convolutional + LReLU	16×16×64	7×7×64×64	1	16×16×64
	6	Max Pooling	16×16×64	2×2	2	8×8×64
	7	Convolutional + LReLU	8×8×64	7×7×64×64	2	4×4×64
	8	Convolutional + LReLU	4×4×64	7×7×64×64	1	4×4×64
	9	Convolutional + LReLU	4×4×64	7×7×64×64	1	4×4×64
Decoder	10	Deconvolutional + LReLU	4×4×64	7×7×64×64	1	4×4×64
	11	Deconvolutional + LReLU	4×4×64	7×7×64×64	1	4×4×64
	12	Deconvolutional + LReLU	4×4×64	7×7×64×64	2	8×8×64
	13	Upsampling	8×8×64	2×2	2	16×16×64
	14	Deconvolutional + LReLU	16×16×64	7×7×64×64	1	16×16×64
	15	Deconvolutional + LReLU	16×16×64	7×7×64×64	2	32×32×64
	16	Upsampling	32×32×64	2×2	2	64×64×64
	17	Deconvolutional + LReLU	64×64×64	7×7×64×64	1	64×64×64
	18	Deconvolutional	64×64×64	7×7×64×1	2	128×128×1

É importante mencionar que os pesos da Facenet não foram modificados durante o treinamento. Isso não foi feito, pois não era interesse da pesquisa melhorar a performance de reconhecimento da Facenet; o que precisamos é usar o gradiente da mesma como substituto para o gradiente de um sistema de reconhecimento qualquer, visando com isso regularizar o treinamento de modo a direcionar nossos parâmetros para regiões que mantenham ou melhorem o reconhecimento facial, mantendo a identidade da entrada da rede. Embora a Facenet seja treinada apenas com imagens 2D, já foi mostrado na literatura que modelos deste tipo podem generalizar relativamente bem para outras modalidades de reconhecimento facial (DAHIA et al., 2017).

Foi utilizada a distância entre os descritores da imagem de entrada e saída, respectivamente $h(\mathbf{x})$ e $h(\hat{\mathbf{y}})$, assim como entre a saída e a neutra correspondente, respectivamente $h(\hat{\mathbf{y}})$ e $h(\mathbf{y})$, para manter traços da identidade da entrada e da neutra correspondente. α é um hiperparâmetro que controla qual dos descritores que a saída da rede será mais parecida. Aumentar α prioriza o reconhecimento sem expressões faciais, e, empiricamente, $\alpha = 3$ foi encontrado como um bom valor neste trabalho.

Assim sendo, foram treinadas e avaliadas duas versões da rede proposta:

- **Rede A**: otimizada somente com a função L_2;
- **Rede B**: alternando os passos de otimização entre a função L_2 e o *recognition loss*.

Nas duas versões, a fase de treinamento foi dividida em duas partes: pré-treino baseado em *autoencoder* e treinamento *encoder-decoder* para remoção de expressões.

2.3.1 Pré-treino baseado em autoencoder

Essa fase usa a mesma imagem como entrada e saída esperada, para aprender como executar a redução de dimensionalidade da imagem de entrada e como reconstruir a mesma com o menor erro possível. É um problema mais fácil se comparado com o problema de remoção. Porém, o domínio é semelhante o suficiente para ser utilizado como uma boa inicialização para a próxima etapa do treinamento (SHRIVASTAVA et al., 2016). Essa fase foi otimizada com a função L_2 e o otimizador Adam (KINGMA; BA, 2014), taxa de aprendizado com decaimento polinomial de 10^{-2} até 10^{-5} e mini-batches de 256 exemplos por 50 épocas. Esses parâmetros foram definidos empiricamente, com base no desempenho no conjunto de validação. O mesmo processo de inicialização foi utilizado para as duas versões da rede e é ilustrado pela Figura 2.5.

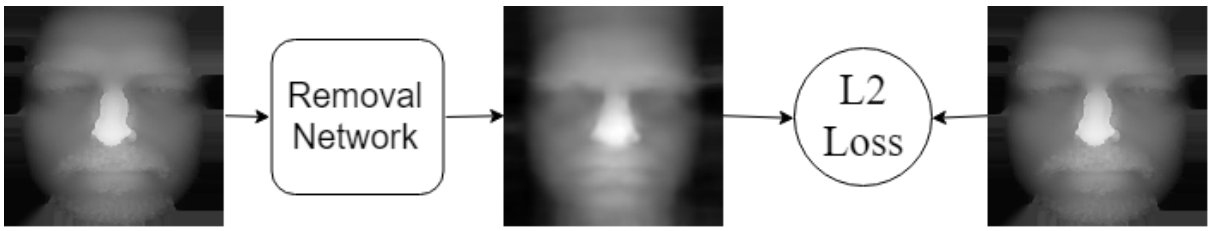

Figura 2.5 Ilustração do processo do pré-treino, baseado em autoencoder. Todas as imagens neutras do conjunto de treinamento são utilizadas para inicializar os pesos da rede.

2.3.2 Treinamento encoder-decoder

Foram utilizados todos os pares de faces, sendo uma não-neutra e a outra neutra, do conjunto de treinamento para a otimização da rede para remoção de expressões. Para a **Rede A** foi utilizada a mesma configuração de treinamento da etapa de inicialização, exceto pelo número de épocas, que foi alterado para 3000. Para a **Rede B**, foram intercalados batches de treinamento utilizando a função L_2 e o *recognition loss* na proporção de uma quatro para um. Além de também utilizar 3000 épocas, a taxa de aprendizado foi de 10^{-5} à 10^{-7} nos batches que utilizaram o *recognition loss*. O processo é ilustrado pela Figura 2.6.

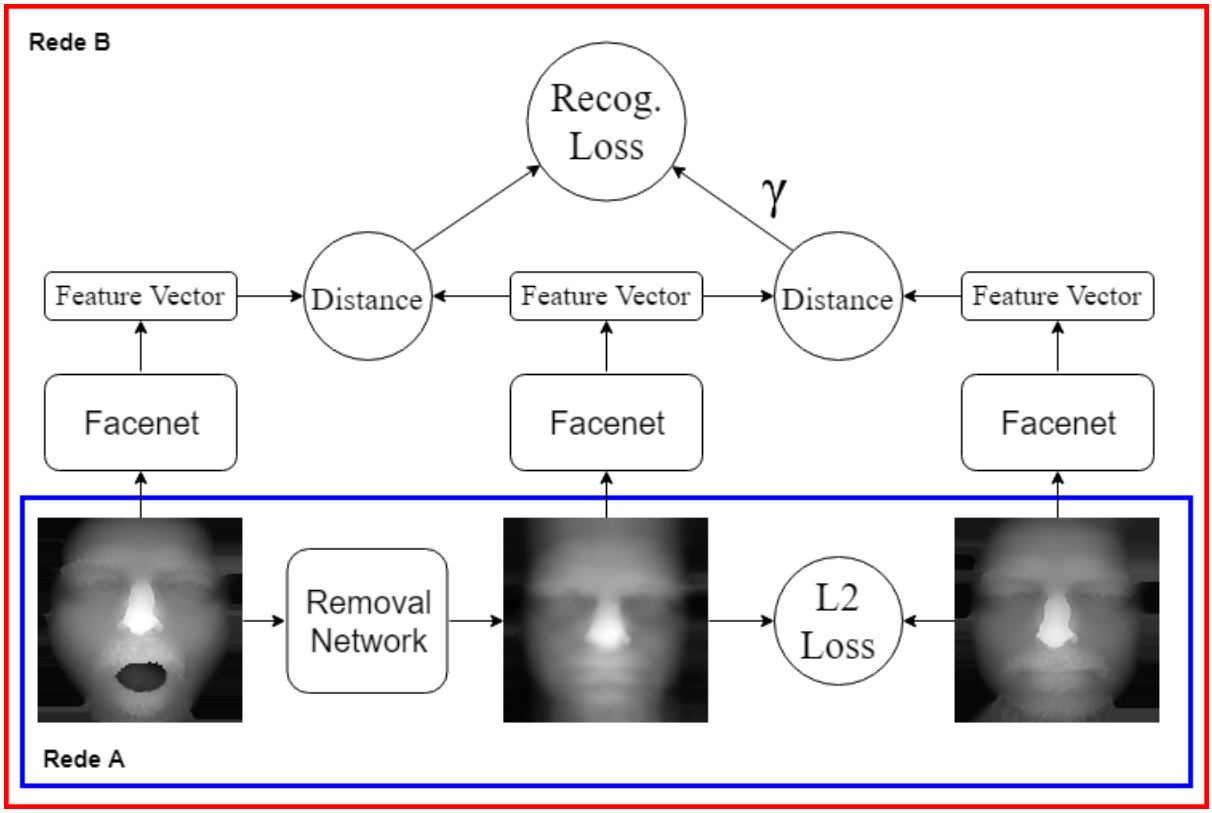

Figura 2.6 Ilustração do processo de treinamento das Redes A e B. O bloco azul é a Rede A, onde o apenas o L2 Loss é utilizado como função de custo. O bloco vermelho é a Rede B, apresentando também o Recognition Loss, que é intercalado com o L2 Loss.

Capítulo 3

RESULTADOS EXPERIMENTAIS

Os experimentos foram executados em um computador com processador Intel i7-6700k 4GHz com 32GB de memória Random Access Memory (RAM) e uma Graphics Processor Unit (GPU) NVidia Titan X Pascal 12GB. A implementação utiliza a biblioteca PointCloud Library (PCL) (RUSU; COUSINS, 2011) para manipular as nuvens de ponto, Open Source Computer Vision Library (OpenCV) para manipular as imagens, e Tensorflow (ABADI et al., 2015) para construir, treinar e testar as redes neurais.

Foram conduzidos três experimentos para avaliação do processo proposto. Primeiro, foi apresentada uma análise visual das imagens geradas. Então, foi executada uma análise quantitativa da estabilidade da técnica, que foi comparada ao estado-da-arte. Finalmente, foi comparada a performance de quatro métodos de reconhecimento facial, antes e depois de usar o trabalho proposto para remoção de expressões faciais, para avaliar o impacto da nossa abordagem na tarefa de reconhecimento.

Com o objetivo de corroborar o poder de generalização do método proposto, a base de dados foi divida de quatro formas diferentes, sempre respeitando a regra de que um indivíduo só pertence a um único subconjunto.

3.1 ANÁLISE VISUAL

Esta seção apresenta alguns resultados visuais da rede *encoder-decoder* usando imagens do conjunto de teste. Como pode ser visto na Figura 3.1, o principal problema do resultado é a suavização em regiões que são importantes para o reconhecimento, como as áreas ao redor dos olhos (KEIL, 2009). Somente alguns exemplos apresentam detalhes discerníveis na região dos olhos e, em alguns casos, uma perda substancial de informação é observada quando comparada com a imagem de entrada ou com o *ground truth*. Estes resultados, porém, são esperados, uma vez que foi utilizado uma correção baseada na distância L_2 durante o treinamento, e sabe-se que isso produz resultados suavizados (DOSOVITSKIY; BROX, 2016).

Ainda assim, mesmo que os resultados não sejam muito realísticos neste ponto, eles possuem uma propriedade interessante: quando duas imagens comparadas são processadas pela rede, se elas pertencem a mesma pessoa, elas continuarão muito parecidas entre

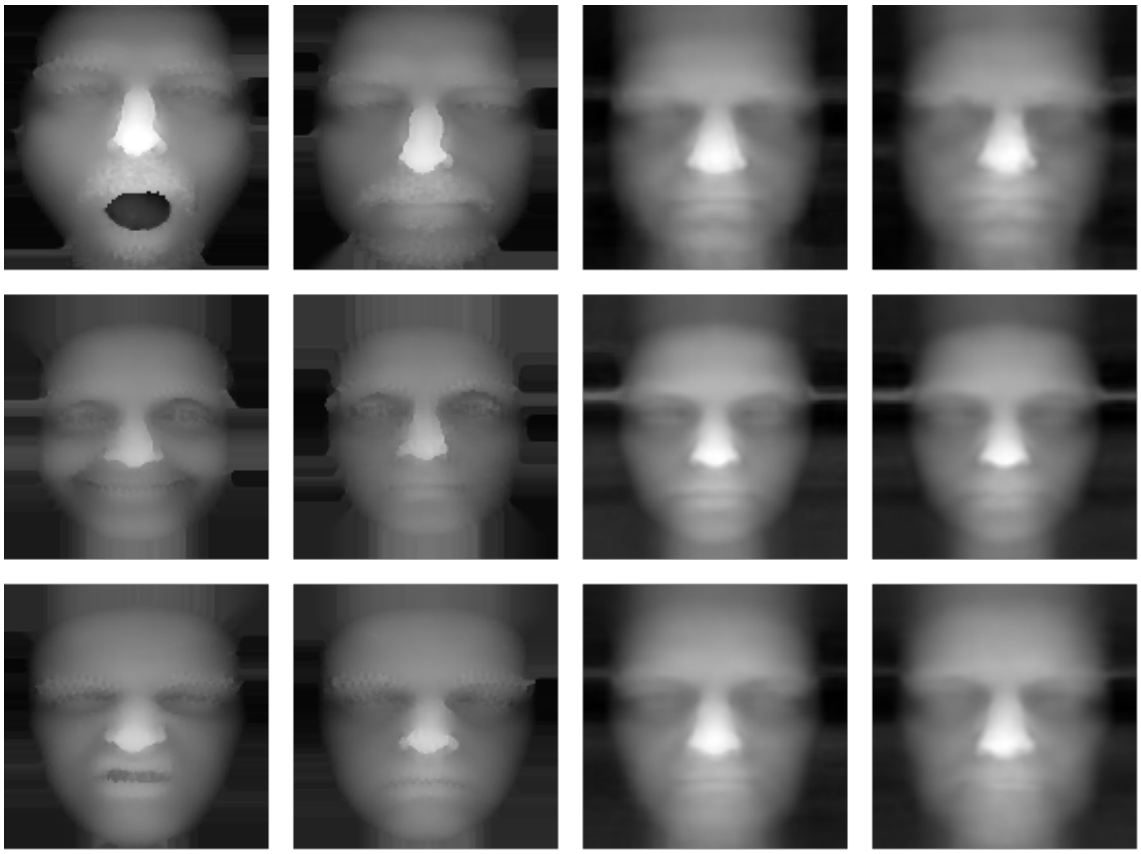

Figura 3.1 Resultados da rede *encoder-decoder* para remoção de expressões faciais em imagens que não estavam presentes no conjunto de treino. Cada linha apresenta a saída para uma imagem diferente, e é composta pela imagem de entrada com expressões em (a), a imagem neutral correspondente em (b), e os resultados após ser processadas pela **Rede A** em (c) e **Rede B** em (d).

si mesmo quando afetadas por diferentes expressões. Tal resultado pode ser observado na Figura 3.2, onde a diferença entre quatro pares de imagens é ilustrado na forma de mapas de calor. Fica claro que a diferença entre duas imagens neutras da mesma pessoa ou duas imagens de pessoas diferentes é preservada, enquanto a diferença entre imagens da mesma pessoa é reduzida quando ao menos uma é afetada por deformações causadas por expressões. Também é visível que a **Rede B**, ao menos visualmente, é melhor na execução da tarefa do que a **Rede A**, uma observação difícil de ser verificada sem a assistência dos mapas de calor. Note que mesmo imagens neutras devem ser processadas pela rede, uma vez que as imagens originais estão em um domínio diferente (devido aos seus detalhes nítidos e ausência de artefatos da rede).

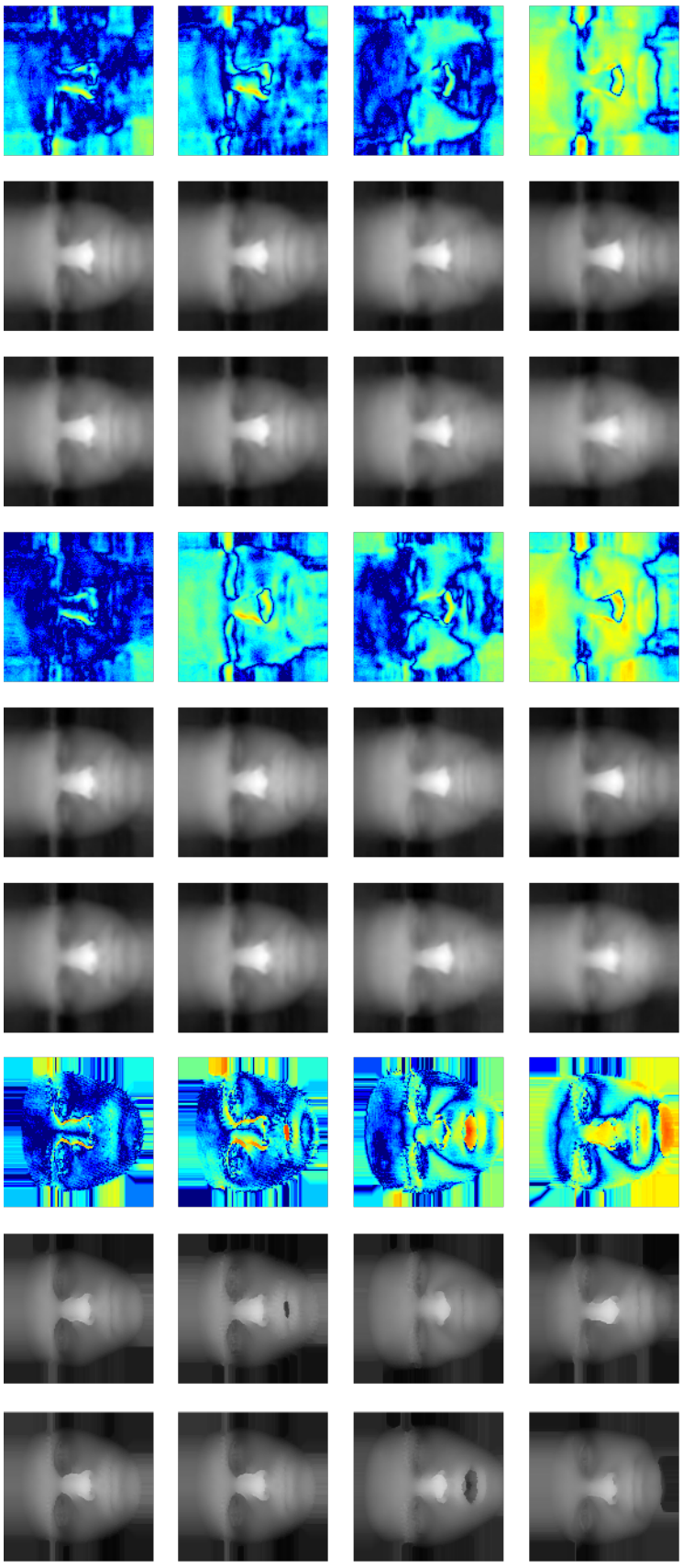

Figura 3.2 Resultados para ilustrar a aplicabilidade da remoção de expressões faciais no reconhecimento. As três primeiras linhas apresentam a comparação de pares de imagens da mesma pessoa em diferentes cenários: neutra vs. neutra, neutra vs. não-neutral e não-neutra vs. não-neutra, respectivamente. A última linha mostra a comparação de imagens neutras de pessoas diferentes. As imagens originais do conjunto de teste estão em (a) e (b), o mapa de calor da diferença entre elas está em (c). As colunas (d) e (e) apresentam as saídas da **Rede A** para (a) e (b), e a diferença entre elas é apresentada em (f). Finalmente, as saídas da **Rede B** para (a) e (b) são mostradas em (g) e (h), e a diferença entre elas em (i). Em todos os mapas de calor, áreas com grande diferença são apresentadas em vermelho, enquanto diferenças pequenas estão em azul. Todos os mapas usam escala logarítmica para realçar as diferenças.

Mesmo com a análise visual indicando um potencial para melhorar o reconhecimento, não é suficiente para validar os resultados. Então, foram executadas análises quantitativas do desempenho na Seção 3.2 e na Seção 3.3.

3.2 AVALIAÇÃO DA ESTABILIDADE DA REMOÇÃO DE EXPRESSÕES

Uma abordagem estável de remoção de expressões faciais deve aumentar consideravelmente a similaridade intraclasse e aumentar a separação entre as distribuições de similaridade intraclasse e interclasse. Para quantificar essa estabilidade, foi utilizado o Root Mean Square Error (RMSE) entre pares de imagens. Quanto mais próximo de zero for este valor, mais parecidas as imagens são. Foi computado o RMSE para cada par de imagens neutras e não-neutras no conjunto de testes e calculada a média dos valores para combinações intraclasse e interclasse. Esse procedimento foi repetido para as imagens originais e suas versões processadas usando a **Rede A** e a **Rede B**, e os valores obtidos são apresentados na Tabela 3.1. Também são apresentadas as Probability Density Function (PDF) para as classes, na Figura 3.3, mostrando a redução do RMSE nas duas classes e a ampliação da distância entre as mesmas. Como esperado, a média do RMSE para combinações intraclasse foi menor que o valor para combinações interclasse em todos os casos. Também foi visto que, após a remoção de expressões, os valores para ambas as classes foram reduzidos. Isto significa que imagens da mesma pessoa estão mais parecidas, porém, imagens de pessoas diferentes também se aproximaram. Porém, a separação entre os valores das classes foi ampliado, mostrando que remover as expressões torna mais fácil discriminá-las quando utilizada a métrica RMSE.

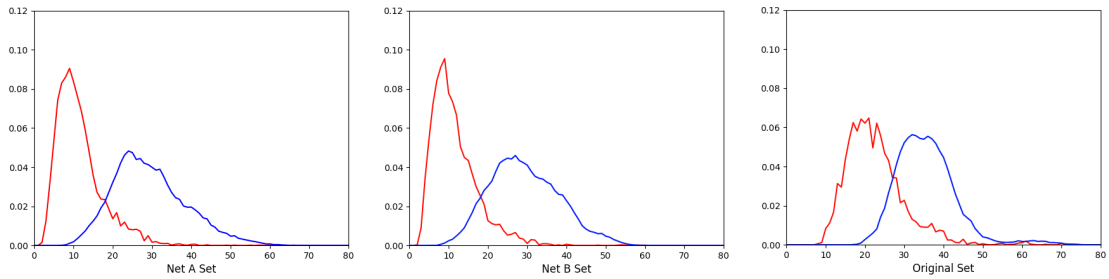

Figura 3.3 Curva da PDF da interclasse (azul) e intraclasse (vermelho). Observa-se uma maior concentração de valores dos erros da intraclasse e uma ampliação da distância entre as duas.

Quando comparada com o estado-da-arte, a abordagem proposta foi capaz de reduzir a média intraclasse em 47.8%, semelhante à redução de 43.6% alcançada por Pan et al. (2010). Embora eles tenham utilizado uma base de dados diferente, BU-3DFE (YIN et al., 2006), ela é bem semelhante ao subconjunto da Bosphorus utilizado neste trabalho (BU-3DFE vs. subconjunto da Bosphorus: 100 vs. 105 indivíduos; 1 vs. 2.84 faces neutras por pessoa; 24 vs. 20.8 faces não neutras por pessoa; intensidades diferentes de expressões em ambas). A principal diferença é que nesta pesquisa foram utilizados conjuntos fixos de treino e teste, enquanto Pan *et al.* segue o protocolo *leave-one-out*. Eles repetem o experimento deixando uma pessoa fora do treino por vez e calculam a média de todos

Tabela 3.1 Média do RMSE para todos os pares de imagens neutras e não-neutras no conjunto de teste usando suas versões originais e versões processadas pelas **Rede A** e **Rede B**. Cada média está acompanhada pelo seu desvio padrão. Após a remoção de expressões, o erro da intraclasse diminui e a distribuição se encontra mais concentrada ao redor da média, enquanto a separação entre intraclasse e interclasse aumenta.

	Intraclasse	**Interclasse**
Original	0.046176 ± 0.011230	0.072338 ± 0.011539
Net A	0.024199 ± 0.008934	0.059927 ± 0.014833
Net B	0.024082 ± 0.008785	0.059696 ± 0.014278

os resultados parciais. Isso significa que alcançamos um resultado comparável utilizando uma razão 2:1 de treino para teste enquanto Pan et al. (2010) usa uma razão 99:1.

Estes resultados sugerem que o sistema de reconhecimento facial pode se beneficiar desta técnica, então continuamos os experimentos com a avaliação dos impactos da nossa abordagem em diferentes sistemas de reconhecimento.

3.3 EFEITO DA REMOÇÃO DE EXPRESSÕES NO RECONHECIMENTO FACIAL

Para avaliar o efeito da nossa abordagem para remoção de expressões no reconhecimento, foi investigado o resultado de quatro métodos diferentes: Eigenfaces (TURK; PENTLAND, 1991), processo baseado no Principal Components Analysis (PCA); Fisherfaces (BELHUMEUR et al., 1997), processo baseado no Linear Discriminant Analysis (LDA); Local Binary Pattern Histogram (LBPH) (AHONEN et al., 2004), processo baseado na construção de histrogramas de Local Binary Pattern (LBP); e Facenet (SCHROFF et al., 2015), uma CNN considerada o estado-da-arte para imagens 2D e que pode ser aplicada em imagens de profundidade (DAHIA et al., 2017). Eigenfaces e Fisherfaces foram treinados utilizando apenas faces neutras no conjunto de treino. Fazendo isso, os modelos conseguem representar um sistema de reconhecimento que não sabe lidar com as variações causadas por expressões. O conjunto de validação foi usado para determinar o número de autovetores que maximiza a acurácia do reconhecimento, encontrando o número 30. Consequentemente, cada face no conjunto de teste foi descrita por um vetor de 30 dimensões. LBPH e Facenet não necessitam de nenhum treinamento, e foram usados para extrair vetores 4096 e 128 dimensões, respectivamente, de cada face de teste. LBPH é conhecido por ser robusto a variações de expressões (KHORSHEED; YURTKAN, 2016). Facenet, mesmo treinado em imagens 2D, possui um desempenho razoável em imagens 3D e ajuda a verificar se o método melhora a precisão de uma CNN.

Cada experimento de reconhecimento foi repetido três vezes, assim como na seção anterior, uma utilizando as imagens originais e as outras duas utilizando as imagens após serem processadas pelas **Rede A** e **Rede B**. Vale a pena dizer que mesmo as faces neutras devem ser processadas pelas redes de remoção de expressões para conseguir os resultados apresentados na Figura 3.2. Para cada combinação de técnica de reconhe-

cimento e conjunto de dados, foram executados dois experimentos: uma comparação "Todos contra Todos", com resultados reportados por curvas do tipo Receiver Operating Characteristic (ROC) na Figura 3.5 e valores de Equal Error Rate (EER) na Tabela 3.2; e um ranking de identificação (Rank-N) usando uma ou duas imagens neutras de cada indivíduo para formar a galeria e todas as imagens não-neutras como amostras a serem identificadas, com resultados reportados como curvas do tipo Cumulative Match Characteristic (CMC) na Figura 3.4. O primeiro experimento avalia o quão bem comparações genuínas e impostoras podem ser diferenciadas, enquanto o último diz o quão fácil é a identidade da amostra a ser reconhecida pode ser recuperada em um cenário controlado em que expressões faciais não são permitidas durante o cadastro biométrico.

Este experimento foi executado em quatro configurações de base de dados diferentes, com o objetivo de avaliar o poder de generalização do modelo proposto. Na média, o benefício da remoção, seja com a **Rede A** ou **Rede B**, é claro quando comparado com os resultados das imagens não processadas. Rank-1 e EER foram melhorados em todos os casos, com uma razoável vantagem da **Rede B** sobre a **Rede A** na maioria dos casos. Esse resultado mostra que a remoção de expressões melhora a precisão dos sistemas de reconhecimento; isso é observado mesmo quando o método de reconhecimento usado para otimizar a rede é diferente do que foi testado, mostrando que a abordagem usando o gradiente da Facenet como fator de regularização é promissor. Mesmo assim,

Tabela 3.2 EER considerando diferentes grupos de comparações genuínas: Todos contra Todos, Neutras contra Não-Neutras e Não-Neutras contra Não-Neutras. O conjunto de comparações impostoras é sempre o mesmo (Todos contra Todos).

Método	Todos vs Todos			Neutras vs Não-Neutras			Não-Neutras vs Não-Neutras		
	Rede A	Rede B	Orig.	Rede A	Rede B	Orig.	Rede A	Rede B	Orig.
PCA Média	0.092	**0.089**	0.110	0.093	**0.089**	0.107	0.139	**0.134**	0.194
PCA #1	0.108	**0.106**	0.139	0.087	**0.086**	0.120	0.131	**0.122**	0.185
PCA #2	0.076	**0.072**	0.109	0.086	**0.081**	0.123	0.133	**0.123**	0.200
PCA #3	0.084	**0.077**	0.084	0.087	**0.081**	**0.081**	0.144	**0.143**	0.190
PCA #4	**0.101**	**0.101**	0.109	0.111	**0.106**	**0.106**	0.150	**0.149**	0.201
LDA Média	0.096	**0.089**	0.116	0.100	**0.097**	0.122	0.189	**0.188**	0.249
LDA #1	0.137	**0.120**	0.183	0.096	**0.089**	0.128	0.170	**0.164**	0.256
LDA #2	0.104	**0.099**	0.148	0.152	**0.148**	0.214	0.178	**0.177**	0.240
LDA #3	0.055	**0.052**	0.057	0.066	0.063	**0.061**	0.194	**0.193**	0.238
LDA #4	0.086	**0.083**	0.078	0.086	0.087	**0.084**	**0.213**	0.218	0.264
LBPH Média	0.069	**0.067**	0.084	0.075	**0.072**	0.086	0.228	**0.227**	0.263
LBPH #1	0.071	**0.069**	0.097	0.062	**0.060**	0.085	0.229	**0.230**	0.278
LBPH #2	0.065	**0.064**	0.093	0.076	**0.075**	0.104	0.217	**0.213**	0.273
LBPH #3	0.065	**0.061**	0.069	0.072	0.065	**0.062**	0.261	**0.257**	0.267
LBPH #4	**0.074**	**0.074**	0.078	0.091	**0.090**	0.093	**0.204**	0.207	0.235
Facenet Média	0.146	**0.143**	0.158	0.155	**0.153**	0.161	0.194	**0.191**	0.256
Facenet #1	0.163	**0.160**	0.190	0.170	**0.168**	0.193	0.177	**0.174**	0.243
Facenet #2	0.132	**0.128**	0.161	0.128	**0.127**	0.168	0.189	**0.185**	0.266
Facenet #3	0.146	0.143	**0.131**	0.145	0.136	**0.123**	0.190	**0.185**	0.248
Facenet #4	0.145	**0.143**	0.150	0.177	0.184	**0.161**	**0.221**	0.222	0.268

várias outras observações podem ser feitas baseadas nos resultados apresentados:

1. Mesmo com pequenas divergências entre os resultados de cada uma das separações, os mesmos são similares. Isso corrobora o poder de generalização da rede que, independente de quem esteja no conjunto de treino/validação e teste, consegue atingir resultados similares.

2. Pode ser visto que o Eigenfaces é melhor que o Fisherfaces para o conjunto de imagens originais, mas estes trocam de posições quando as expressões são removidas pela **Rede B**. Sabendo que os dois métodos foram treinados apenas com imagens neutras, Fisherfaces provavelmente ficou muito especializado nesse domínio específico. Já o Eigenfaces procura representar os principais componentes do conjunto de treino, o que o ajuda a generalizar melhor para domínios similares (faces com expressões). O poder discriminativo do Fisherfaces é recuperado quando as imagens são processadas pela nossa rede, indicando que o domínio de treino (apenas imagens neutras) foi restabelecido com sucesso.

3. Na Figura 3.4 pode ser visto que o Rank-N aumenta nas imagens originais quando uma segunda imagem por indivíduo é adicionada a galeria, mas isso não acontece nos conjuntos de imagens processadas. Como mostrado por Bowyer *et al.* (BOWYER et al., 2006), adicionar múltiplas imagens da mesma pessoa na galeria só é útil se houver variações entre elas. Isso sugere que mesmo as faces neutras ficam muito semelhantes umas das outras depois de serem processadas pelas redes e que o método proposto pode reduzir o esforço durante o cadastro em um sistema de reconhecimento facial.

4. Existe uma pequena diferença entre as comparações não-neutra contra neutra e não-neutra contra não-neutra na Tabela 3.2 para ambas as redes avaliadas, indicando que não é necessário controlar a expressão do usuário durante o cadastro do sistema de reconhecimento se a nossa abordagem para remoção de expressões estiver em uso.

5. Existe uma melhoria na precisão mesmo nos métodos que são robustos a expressões, mostrando que os benefícios da remoção não são limitados aos métodos que supostamente não sabem lidar com tais deformações.

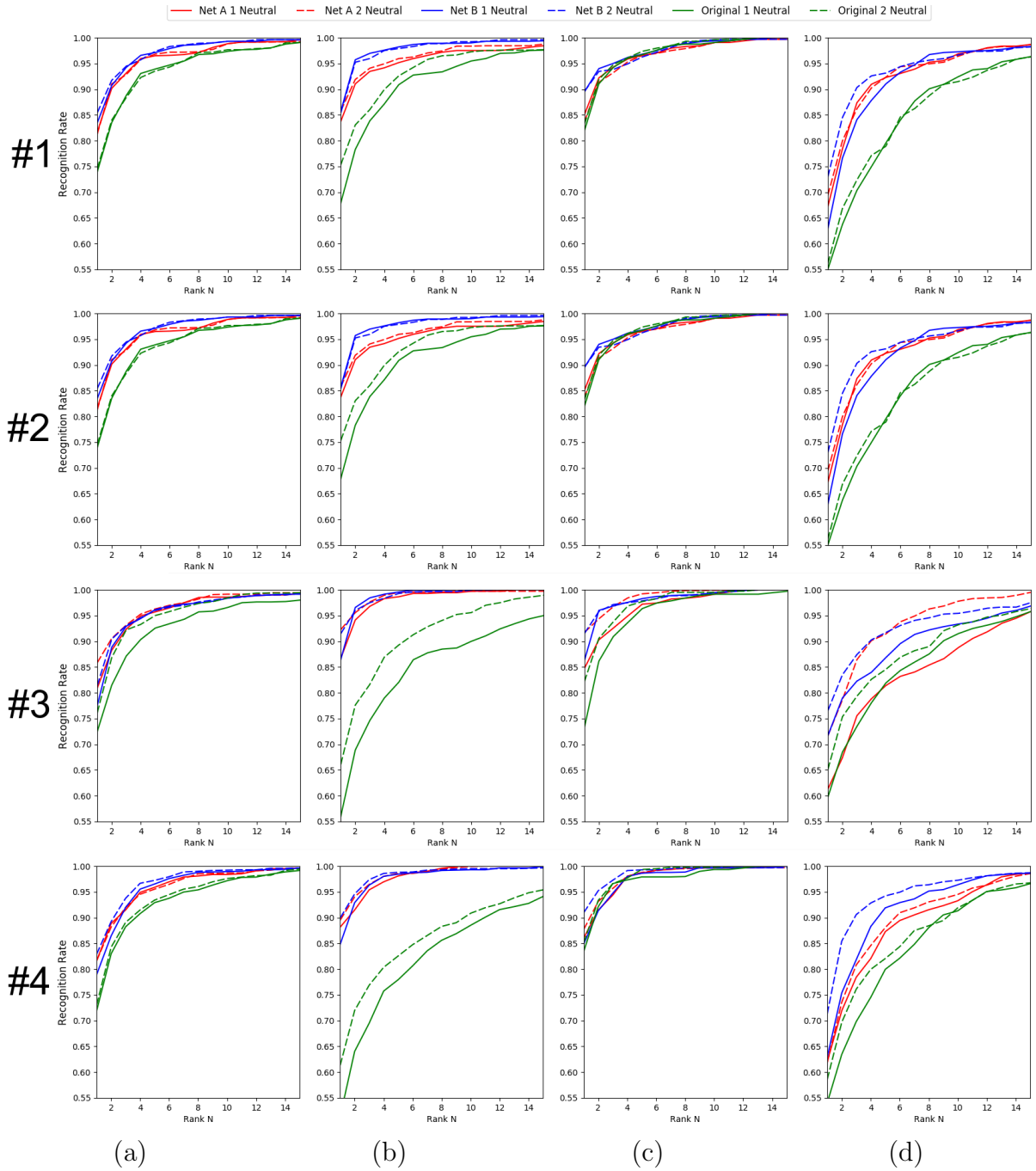

Figura 3.4 Curva CMC para resultados de identificação. Cada linha representa uma separação diferente do *dataset*. Os métodos utilizados foram (a) Eigenfaces, (b) Fisherfaces, (c) LBPH, e (d) Facenet. Linhas sólidas foram obtidas quando existia apenas uma imagem por pessoa na galeria, as tracejadas com duas imagens por pessoa.

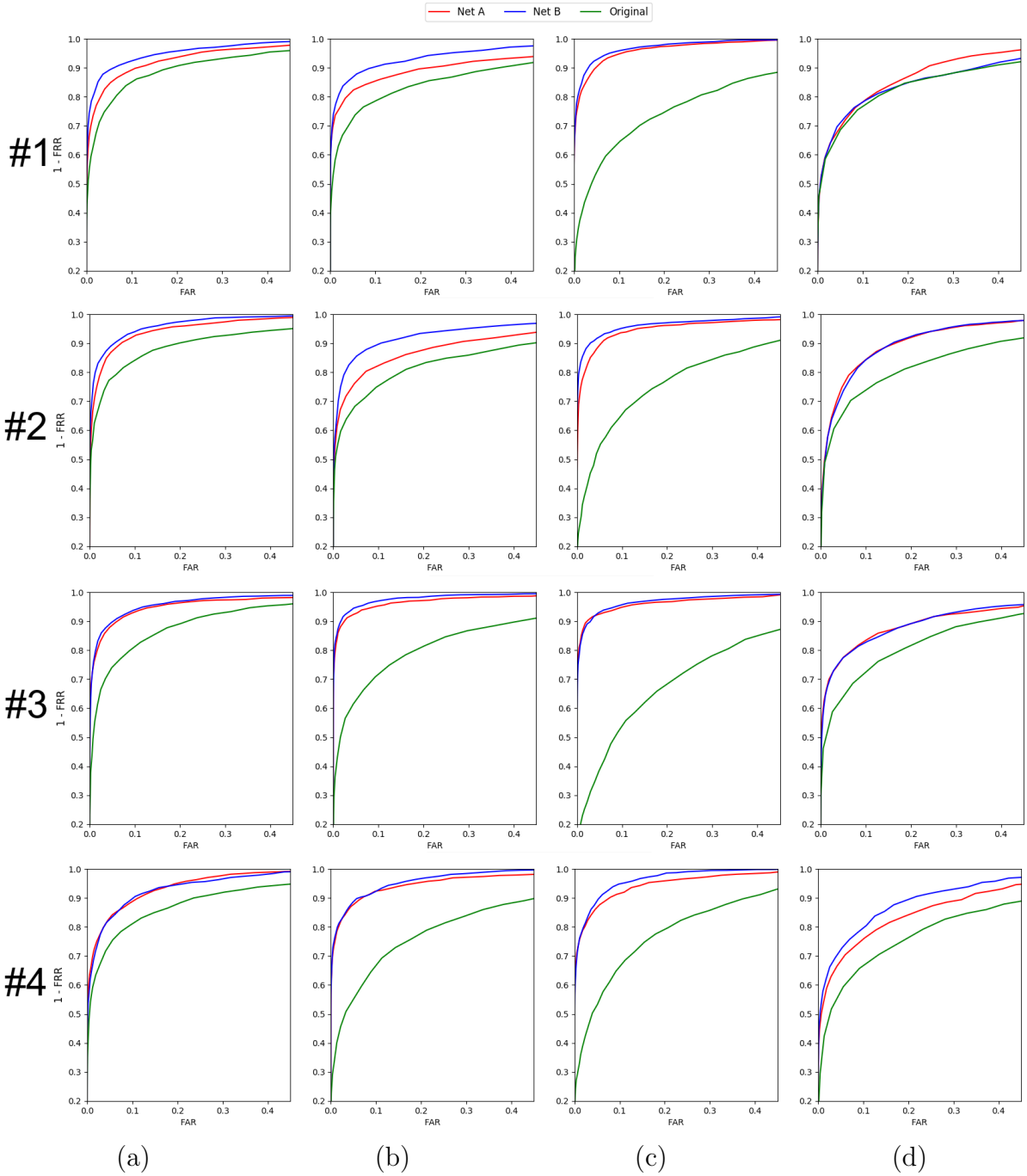

Figura 3.5 Curvas ROC para resultados de verificação. Cada linha representa uma separação diferente do *dataset*. Os métodos utilizados foram (a) Eigenfaces, (b) Fisherfaces, (c) LBPH, e (d) Facenet. FRR significa *False Rejection Rate* (Taxa de Falsos Negativos) e FAR significa *False Acceptance Rate* (Taxa de Falsos Positivos).

Capítulo 4

CONCLUSÃO

Foi utilizada uma rede neural *encoder-decoder* para remoção de expressões faciais em imagens 3D buscando melhorar a precisão em sistemas de reconhecimento. A principal contribuição foi o uso de um sistema de reconhecimento para guiar o processo de treinamento, assim mantendo os traços de identidade na imagem neutra de saída, e foram apresentadas as vantagens da abordagem proposta através de avaliações qualitativas e quantitativas.

A abordagem foi capaz de reduzir o RMSE entre a imagem neutra e a imagem não-neutra aproximadamente pela metade, o que é comparável ao estado-da-arte, mesmo que necessitando muito menos dados para treinamento. Ela também aumentou a separação entre os valores intraclasse e interclasse do RMSE. Quando utilizadas para fins de reconhecimento, as imagens que tiveram suas expressões removidas pela técnica proposta melhoraram a precisão de quatro métodos de reconhecimento facial diferentes. Finalmente, foi observado que a técnica reduz o esforço de cadastro em um sistema de reconhecimento, pois requer menos imagens na galeria e não necessita impor que o usuário esteja com uma expressão neutra.

Como trabalho futuro, pretendemos utilizar a combinação de múltiplas bases de dados de imagens faciais 3D para treinar uma CNN de reconhecimento 3D para ser utilizada como função de correção. Também pretendemos investigar se imagens mais realísticas, como as obtidas por redes do tipo Generative Adversarial Network (GAN) (GOODFELLOW et al., 2014) podem melhorar ainda mais os resultados de reconhecimento. Uma versão da implementação deste trabalho se encontra disponível para acesso no link https://github.com/lucasamparo/faceExpressionRemoval/ .

REFERÊNCIAS BIBLIOGRÁFICAS

ABADI, M. et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. 2015. Software available from tensorflow.org. Disponível em: ⟨https://www.tensorflow.org/⟩.

AGIANPUYE, A. S.; MINOI, J. L. Synthesizing neutral facial expression on 3d faces using active shape models. In: *IEEE REGION 10 SYMPOSIUM*. [S.l.: s.n.], 2014. p. 600–605.

AHONEN, T. et al. Face recognition with local binary patterns. In: *European Conference on Computer Vision*. [S.l.: s.n.], 2004. p. 469–481.

BADRINARAYANAN, V.; KENDALL, A.; CIPOLLA, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *CoRR*, abs/1511.00561, 2015. Disponível em: ⟨http://arxiv.org/abs/1511.00561⟩.

BELHUMEUR, P. N. et al. Eigenfaces vs. fisherfaces: recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 19, n. 7, p. 711–720, 1997.

BERRETTI, S. et al. 3d face recognition using isogeodesic stripes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 32, n. 12, p. 2162–2177, 2010.

BESL, P. J.; MCKAY, N. D. A method for registration of 3-d shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 14, n. 2, p. 239–256, 1992.

BOWYER, K. W. et al. Face recognition using 2-d, 3-d, and infrared: Is multimodal better than multisample? *Proceedings of the IEEE*, v. 94, n. 11, p. 2000–2012, 2006.

CAMPOS, T. E. et al. Eigenfaces versus eigeneyes: First steps toward performance assessment of representations for face recognition. In: *Mexican International Conference on Artificial Intelligence*. [S.l.: s.n.], 2000. p. 193–201.

CHANG, K. I. et al. Multiple nose region matching for 3d face recognition under varying facial expression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 28, n. 10, p. 1695–1700, 2006.

DAHIA, G. et al. A study of cnn outside of training conditions. In: *IEEE International Conference on Image Processing*. [S.l.: s.n.], 2017. p. 3820–3824.

DING, H. et al. Exprgan: Facial expression editing with controllable expression intensity. In: *AAAI*. [S.l.: s.n.], 2018.

DOSOVITSKIY, A.; BROX, T. Generating images with perceptual similarity metrics based on deep networks. *CoRR*, 2016.

ELAIWAT, S. et al. 3-d face recognition using curvelet local features. *IEEE Signal Processing Letters*, v. 21, n. 2, p. 172–175, 2014.

EMAMBAKHSH, M.; EVANS, A. Nasal patches and curves for expression-robust 3d face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 39, n. 5, p. 995–1007, 2017.

FRITH, C. Role of facial expressions in social interactions. *Philos. Trans. Royal Soc. B*, v. 364, n. 1535, p. 3453–3458, 2009.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. [S.l.]: MIT Press, 2016. ⟨http://www.deeplearningbook.org⟩.

GOODFELLOW, I. et al. Generative adversarial nets. In: *Neural Information Processing Systems*. [S.l.: s.n.], 2014. p. 2672–2680.

JAN, A. et al. Accurate facial parts localization and deep learning for 3d facial expression recognition. In: *IEEE FG*. [S.l.: s.n.], 2018.

KAKADIARIS, I. A. et al. Three-dimensional face recognition in the presence of facial expressions: An annotated deformable model approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 29, n. 4, p. 640–649, 2007.

KEIL, M. S. "i look in your eyes, honey": Internal face features induce spatial frequency preference for human face processing. *PLoS Computational Biology*, v. 5, n. 3, mar 2009.

KHORSHEED, J. A.; YURTKAN, K. Analysis of local binary patterns for face recognition under varying facial expressions. In: *Signal Processing and Communication Application Conference*. [S.l.: s.n.], 2016. p. 2085–2088.

KINGMA, D. P.; BA, J. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

KRIZHEVSKY, A. et al. Imagenet classification with deep convolutional neural networks. In: *Conference on Neural Information Processing Systems*. [S.l.: s.n.], 2012. p. 1097–1105.

LI, H. et al. Eigen-pep for video face recognition. In: *Asian Conference on Computer Vision*. [S.l.: s.n.], 2015. p. 17–33.

LIU, G. et al. Image Inpainting for Irregular Holes Using Partial Convolutions. *ArXiv e-prints*, abr. 2018.

LU, X.; JAIN, A. Deformation modeling for robust 3d face matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 30, n. 8, p. 1346–1357, 2008.

PAN, G. et al. Removal of 3d facial expressions: A learning-based approach. In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2010. p. 2614–2621. ISSN 1063-6919.

PATHIRAGE, C. S. N. et al. Stacked face de-noising auto encoders for expression-robust face recognition. In: *2015 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*. [S.l.: s.n.], 2015. p. 1–8.

RADFORD, A. et al. Unsupervised representation learning with deep convolutional generative adversarial networks. *CoRR*, 2015.

RUSU, R. B.; COUSINS, S. 3d is here: Point cloud library (pcl). In: *ICRA*. [S.l.: s.n.], 2011. p. 1–4.

SAVRAN, A. et al. Bosphorus database for 3d face analysis. In: SCHOUTEN, B. et al. (Ed.). *Biometrics and Identity Management*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008. p. 47–56. ISBN 978-3-540-89991-4.

SCHROFF, F. et al. Facenet: A unified embedding for face recognition and clustering. *CoRR*, 2015.

SHRIVASTAVA, A. et al. Learning from simulated and unsupervised images through adversarial training. *CoRR*, 2016.

SONG, L. et al. Geometry guided adversarial facial expression synthesis. *CoRR*, 2017.

TURK, M.; PENTLAND, A. Eigenfaces for recognition. *J. Cognitive Neuroscience*, v. 3, n. 1, p. 71–86, 1991.

YANG, H.; YIN, L. Cnn based 3d facial expression recognition using masking and landmark features. In: *ACII*. [S.l.: s.n.], 2017. p. 556–560.

YIN, L. et al. A 3d facial expression database for facial behavior research. In: *IEEE Face Gesture International Conference*. [S.l.: s.n.], 2006. p. 211–216.

ZHAO, W.-Y. et al. Face recognition: A literature survey. *ACM Computing Surveys*, v. 35, p. 399–458, 12 2003.