UNIVERSIDADE FEDERAL DO MARANHÃO

UNIVERSIDADE FEDERAL DO PIAUÍ
Doutorado em Ciência da Computação Associação UFMA/UFPI

Ricardo Teles Freitas

Uma Abordagem Multi-Nível Baseada em Redes Neurais Convolucionais para Redução do Viés Algorítmico na Localização de Pontos Faciais

Orientador: Prof. Dr. Kelson Rômulo Teixeira Aires
Coorientador: Prof. Dr. Anselmo Cardoso de Paiva

Teresina - PI
Setembro, 2023

Ricardo Teles Freitas

Uma Abordagem Multi-Nível Baseada em Redes Neurais Convolucionais para Redução do Viés Algorítmico na Localização de Pontos Faciais

TESE DE DOUTORADO

Tese apresentada como requisito parcial para obtenção do título de Doutor em Ciência da Computação, ao Doutorado em Ciência da Computação, Associação UFMA/UFPI.

Orientador: Prof. Dr. Kelson Rômulo Teixeira Aires
Coorientador: Prof. Dr. Anselmo Cardoso de Paiva

Teresina - PI
Setembro, 2023

Ficha gerada por meio do SIGAA/Biblioteca com dados fornecidos pelo(a) autor(a).
Diretoria Integrada de Bibliotecas/UFMA

Freitas, Ricardo Teles.
Uma abordagem multi-nível baseada em redes neurais convolucionais para redução do viés algorítmico na localização de pontos faciais / Ricardo Teles Freitas. - 2023.
98 f.

Coorientador(a): Anselmo Cardoso de Paiva.
Orientador(a): Kelson Rômulo Teixeira Aires.
Tese (Doutorado) - Programa de Pós-graduação Doutorado em Ciência da Computação - Associação UFMA/UFPI, Universidade Federal do Maranhão, Teresina, 2023.

1. Localização de Pontos Faciais. 2. Redes Neurais Convolucionais. 3. Viés Algorítmico. I. Aires, Kelson Rômulo Teixeira. II. Paiva, Anselmo Cardoso de. III. Título.

Ricardo Teles Freitas

Uma Abordagem Multi-Nível Baseada em Redes Neurais Convolucionais para Redução do Viés Algorítmico na Localização de Pontos Faciais

A presente Tese de Doutorado foi avaliada e aprovada por banca examinadora composta pelos seguintes membros:

Prof. Dr. Kelson Rômulo Teixeira Aires
Orientador
Universidade Federal do Piauí

Prof. Dr. Anselmo Cardoso de Paiva
Coorientador
Universidade Federal do Maranhão

Prof. Dr. Herman Martins Gomes
Universidade Federal de Campina Grande

Prof. Dr. Esteban Walter Gonzalez Clua
Universidade Federal Fluminense

Prof. Dr. Geraldo Braz Junior
Universidade Federal do Maranhão

Prof. Dr. André Macedo Santana
Universidade Federal do Piauí

gov.br
Documento assinado digitalmente
KELSON ROMULO TEIXEIRA AIRES
Data: 14/12/2023 16:28:44-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Kelson Rômulo Teixeira Aires
Orientador

gov.br
Documento assinado digitalmente
RODRIGO DE MELO SOUZA VERAS
Data: 19/12/2023 09:52:12-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Rodrigo de Melo Souza Veras
Coordenador

Teresina - PI, 29 de Setembro de 2023

Ao meu pai

Agradecimentos

Agradeço aos meus pais Marco e Cleonice, ao meu irmão Marquinho e à minha companheira Marcelle.

Ao meu orientador Professor Kelson pela orientação segura e paciência infinita no desenvolvimento deste trabalho.

Ao meu coorientador Professor Anselmo pela fundamental colaboração acerca da temática do trabalho.

Aos meus amigos Maylson, Humberto, Victor e Peri por todo apoio, paciência e compreensão.

Aos meus colegas da UFMA e UFPI.

Aos professores membros da banca examinadora.

A todos os professores e funcionários do DCCMAPI pela colaboração.

A todos que de uma forma ou de outra me estimularam ou me ajudaram.

"Reality denied comes back to haunt."

(Philip Kindred Dick)

Resumo

A localização de pontos de referência facial é uma desafiadora tarefa no contexto da Visão Computacional cujo resultado é aproveitado em diversas aplicações faciais de alto-nível. Para a solução desse problema, modelos baseados em redes neurais convolucionais já atingem níveis de desempenho próximos à anotação humana. O principal critério adotado para a avaliação de desempenho desses modelos é a média da distância ponto-a-ponto, considerando a totalidade de um *dataset*, entre a anotação de um especialista e o valor estimado. Contudo, estudos recentes lançaram luz sobre um problema ainda pouco explorado na avaliação desses modelos relativos a aplicações faciais. Ele consiste na existência de significativas diferenças de desempenho dos modelos de análise facial quando avaliados os resultados entre diferentes grupos demográficos. Isso caracteriza um viés algorítmico que pode levar a uma discriminação ou favorecimento na análise de um grupo em relação aos outros. Este trabalho propõe uma abordagem de localização de pontos faciais que visa a redução das diferenças de desempenho entre grupos demograficamente distintos. A abordagem concentra-se em uma estratégia multi-nível baseada em redes neurais convolucionais para modelagem dos atributos faciais. O nível de cima é composto por um modelo de regressão de coordenadas para detecção de subunidades da face. O nível de baixo utiliza as respostas de detecção para modelar a localização de cada ponto que compõe a subunidade correspondente. Os modelos foram treinados a partir dos *datasets Helen*, *LFPW*, *AFW*, e *300W* desbalanceados com respeito a indivíduos com problemas neurológicos, e aplicados ao *Toronto Neuroface*, balanceado com portadores de ELA (Esclerose Lateral Amiotrófica), sobreviventes de derrame, e um grupo de controle. A comparação com o estado da arte de localização de pontos faciais revelou dois avanços significativos: a aplicação dos modelos do nível de baixo isolados e em condições de detecção de subunidades ideal foi capaz de reduzir as diferenças de desempenho entre todos os grupos, além de reduzir significativamente o erro geral; e a aplicação da abordagem multi-nível reduziu as diferenças de desempenho entre os grupos de controle e portadores de ELA a níveis insignificantes, além de manter os resultados gerais comparáveis ao estado da arte. A abordagem demonstrou ser capaz de atenuar o viés algorítmico presente em modelos preditivos gerados a partir de um *dataset* desbalanceado.

Palavras-chave: Localização de Pontos Faciais, Viés Algorítmico, Redes Neurais Convolucionais.

Abstract

Facial landmarks localization is a challenging task in the context of Vision Computing whose results are explored in many high-level facial applications. To solve this problem, convolutional neural networks models are capable of achieving performance levels close to human annotation. The main criteria adopted to evaluate these models is the average point-to-point distance, regarding the whole dataset, from an expert annotation and the predicted value. However, recent studies shed light on a problem still obscure in those models' evaluations. The issue consists on the existence of significant performance differences of facial analysis models when applied to different demographic groups. This characterizes an algorithmic bias that may lead to discrimination or favoritism in the analysis of a group over another. This work proposes a face alignment approach aiming to reduce the performance gaps among demographically distinct groups with respect to facial landmarks location. The approach focuses on a multi-level strategy based on convolutional neural networks for face modeling. The top level comprises a coordinate regression model for facial subunit detection. The bottom level uses the responses to model the subunits landmarks coordinates. The models were trained with the unbalanced datasets *Helen*, *LFPW*, *AFW*, and *300W* and applied in the *Toronto Neuroface* balanced with *ALS* (amyotrophic lateral sclerosis) patients, post-stroke patients, and a control group. The comparison with the state-of-the-art models for face alignment revealed two significant advances: the application of bottom-level models in ideal facial subunits detection conditions was capable of significantly reducing the performance gap among all groups, besides the overall error was significantly reduced as well; and the application of the multi-level approach reduced the performance gap between *ALS* and the control group to insignificant, besides the overall performance is comparable to the state-of-the-art. The approach showed to be capable of mitigating the algorithmic bias present in predictive models generated after an unbalanced dataset.

Keywords: Face Alignment, Algorithmic Bias, Convolutional Neural Networks.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

ADNet	*Anisotropic Direction Network*
ALS	*Amyotrophic Lateral Sclerosis*
API	*Application Programming Interface*
ASM	*Active Shape Models*
AUC	*Area Under the Curve*
CED	*Cumulative Error Distribution*
CNN	*Convolutional Neural Networks*
ELA	*Esclerose Lateral Amitrófica*
FAN	*Face Alignment Network*
GAT	*Graph Attention Networks*
HC	*Healthy Control*
M_θ	*Modelo de Regressão de Rotação*
M_0	*Modelo de Localização de Pontos Multi-nível*
M_{68}	*Modelo de Localização de Pontos sem Divisão Facial*
M_S	*Modelo de Detecção de Subunidades*
NME	*Normalized Mean Error*
PDI	*Processamento Digital de Imagens*
PPB	*Pilot Parliaments Benchmark*
PS	*Post-stroke*
ReLu	*Rectifier Linear Unit*
RNA	*Redes Neurais Artificiais*
SPIGA	*Shape Preserving with GATs*
TF	*Taxa de Falha*

Sumário

1 Introdução

A localização de pontos faciais (ou alinhamento facial) é um problema que consiste em determinar uma função para estimar um vetor de coordenadas de um conjunto de pontos de referência a partir de uma imagem facial de entrada (KHABARLAK; KORIASHKINA, 2022). A composição desse conjunto varia de acordo com as necessidades do problema envolvido mas, em termos gerais, deve incluir pontos que fazem parte do contorno dos elementos mais representativos da face, como os olhos, o nariz, e a boca. Esses pontos representam características distinguíveis que podem ser aproveitadas na solução de problemas de mais alto nível (YANG; LIU; ZHANG, 2017), (JOHNSTON; CHAZAL, 2018) tais como reconhecimento de expressão facial (PANTIC; ROTHKRANTZ, 2000), estimativa de idade, e classificação de gênero (MOGHADDAM; YANG, 2000), (CAO et al., 2011). Nas aplicações de biometria, por exemplo, a localização desses pontos tem um importante papel no processo de normalização da imagem facial, sendo capaz de aumentar a robustez de métodos de reconhecimento facial em cenários adversos com variações de pose. Em relação a expressões faciais, a precisa localização dos pontos de referência pode fornecer um indicativo de deformidades na face causadas por expressões (JIN; TAN, 2017).

O principal objetivo almejado em abordagens de alinhamento facial consiste em estimar um vetor de coordenadas que se aproxime da localização real dos pontos modelados em uma imagem facial. Em geral, para avaliação de uma abordagem de alinhamento facial em um conjunto de imagens faciais, deve-se calcular, para cada face, a média da distância euclidiana entre a estimativa do modelo e uma observação real, determinada pela marcação humana (KHABARLAK; KORIASHKINA, 2022). O desempenho geral do modelo é medido após normalização do valor obtido no cálculo e agrupamento de todos os resultados alcançados naquele conjunto. Esses resultados subsidiam métricas de avaliação como a média do erro normalizado e a taxa de falha na localização.

Essas abordagens podem ser classificadas em dois tipos: generativas e discriminativas. Métodos generativos constroem modelos de forma e aparência em que o alinhamento facial é encarado como um problema de otimização, ou seja, busca-se uma configuração de parâmetros de forma e textura que resulte na menor distância entre a instância gerada pelo modelo e a aparência observada. Os métodos discriminativos tipicamente buscam inferir a posição dos pontos de referência a partir da aparência local, adotando uma estratégia de regularização global para correção da forma, ou a partir de uma função de regressão vetorial para toda a face, onde as restrições de forma já são implícitas ao modelo (JIN; TAN, 2017). A maioria dos modelos atuais de alinhamento

facial são discriminativos, baseados em redes neurais convolucionais, sendo capazes de atingir níveis de desempenho semelhantes à anotação de especialistas em diversos cenários (BULAT; TZIMIROPOULOS, 2017b), (XU et al., 2021).

Além da relevância do alinhamento facial para aplicações de mais alto nível dentro do domínio das aplicações faciais, há bastante apelo ao tema em virtude de lacunas ainda observadas nas soluções atuais. Fatores como oclusão, variações de pose, iluminação, e expressão facial, tipicamente observados em ambientes não-controlados, são alguns dos exemplos de desafios que atraem pesquisadores nessa área. Nas últimas três décadas, muitos trabalhos foram propostos na busca pelo aperfeiçoamento de soluções existentes com base em algum critério relevante ou sob condições adversas. Dentre esses critérios e cenários, é possível citar também, além do objetivo principal de redução da média geral do erro de localização ponto-a-ponto, a redução de custo computacional, e robustez a diversos ambientes (MILBORROW, 2007), (SESHADRI; SAVVIDES, 2009), (YANG; LIU; ZHANG, 2017).

Apesar do avanço conquistado na solução de problemas relacionados à análise facial, novas questões continuam a surgir e a oferecer desafios para os pesquisadores. Estudos recentes levantaram preocupações com o crescente número de sistemas computacionais de tomada de decisão que estabelecem classificações de indivíduos com o propósito de conceder ou negar um determinado benefício. Muitos desses sistemas são sedimentados em algoritmos que de alguma forma são mais susceptíveis a estigmatizar grupos de indivíduos distinguíveis com base em alguma característica demográfica, como gênero, tipo de pele, e idade (CITRON; PASQUALE, 2014). Buolamwini e Gebru (2018) demonstraram que algumas soluções comerciais de análise facial são inclinadas a privilegiar certos grupos demográficos, o que pode induzir sistemas computacionais a produzir resultados injustos. Em seu trabalho, as autoras avaliaram 3 aplicações comerciais de classificação de gênero a partir de um *benchmark* balanceado por gênero e tipo de pele. Elas concluíram que os classificadores apresentaram um melhor desempenho para indivíduos de pele mais clara e do sexo masculino. Esse tipo de favorecimento ou discriminação tem sido descrito como viés algorítmico ou viés demográfico. Nesse cenário, um algoritmo é considerado enviesado caso diferenças significativas de desempenho sejam observadas em diferentes grupos demográficos (DROZDOWSKI et al., 2020; DEALCALA et al., 2023). Alguns trabalhos propuseram metodologias para investigar, medir, ou reduzir os problemas oriundos desse viés em métodos ou *datasets* voltados para análise facial (HUPONT; FERNÁNDEZ, 2019), (GARCIA et al., 2019), (WU et al., 2020), (GEORGOPOULOS et al., 2021), (MENEZES et al., 2021). Os maiores focos de trabalhos que almejam maior justiça algorítmica abordam primariamente dois aspectos: caracterizar e medir o viés demográfico em métodos e *datasets*; e reduzir os efeitos desse viés nos modelos preditivos (DROZDOWSKI et al., 2020).

1.1 Justificativa

A maioria das abordagens atuais de alinhamento facial concentra-se em propor modelos que visam à redução do erro de localização de pontos avaliando a totalidade de um *dataset* ou *benchmark*, sejam estes definidos em condições mais controladas ou em cenários mais desafiadores. Isso significa que não há distinção nas análises de desempenho no que diz respeito à existência de possíveis subgrupos destacados por suas características demográficas. É possível perceber na literatura uma escassez de trabalhos que busquem tanto caracterizar o viés demográfico em modelos preditivos quanto medir e mitigar os efeitos dos problemas relacionados a esse viés no contexto do alinhamento facial (TAATI et al., 2019). Nesse panorama, Bandini et al. (2021) apresentaram o *Toronto Neuroface*, um *dataset* balanceado de indivíduos com problemas neurológicos (portadores de ELA, sobreviventes de derrame, e um grupo de controle com adultos saudáveis) com o objetivo de investigar e mitigar o viés demográfico observado no estado da arte do alinhamento facial. Além de contar com a presença de indivíduos de 3 grupos clínicos distintos, esse *dataset* foi construído com imagens capturadas em ambiente controlado e livre de condições adversas para aplicações de análise facial como oclusão e variações de iluminação. Após aplicação dos modelos, foi possível observar que eles produziram resultados consideravelmente distintos para cada um dos grupos. Nesse caso, os modelos avaliados apresentaram desempenhos significativamente melhores quando aplicados ao grupo de controle, caracterizando um favorecimento dos métodos aos indivíduos sem problemas neurológicos. Eles creditaram esse favorecimento a dois fatores principais: 1) baixa variabilidade demográfica de indivíduos com problemas neurológicos nos *datasets* de treinamento dos modelos; 2) assimetria orofacial mais acentuada nos grupos demográficos com pior desempenho. Além disso, uma estratégia de ajuste fino com amostras do *Toronto Neuroface* no modelo que obteve o melhor resultado geral foi empregada visando à redução do viés. Contudo, o que se observou foi somente uma melhora no desempenho geral, ao passo que as diferenças dos resultados daquele modelo, quando avaliados os 3 grupos separadamente, permaneceram quase inalteradas.

O problema deste trabalho consiste em reduzir as diferenças de desempenho na localização de pontos faciais resultantes da aplicação de modelos de alinhamento facial em diferentes grupos demográficos, no sentido de produzir resultados que sejam mais igualitários para os grupos envolvidos.

Diante do problema observado e da avaliação dos métodos e *datasets* explorados, algumas questões apontaram os caminhos trilhados no desenvolvimento de uma abordagem baseada em redes neurais convolucionais para alinhamento facial que apresente resultados mais justos entre diferentes grupos demográficos:

- Que técnicas de aprendizagem de máquina devem ser exploradas para a concepção da abordagem de alinhamento facial?
- Que técnicas de pré-processamento podem ser aplicadas nas imagens para a diminuição das diferenças de desempenho entre grupos?
- Que estratégia deve ser adotada na modelagem facial para superar o problema da baixa variabilidade demográfica dos *datasets* de alinhamento facial?

Para contornar os problemas apontados, neste trabalho foi desenvolvida uma abordagem de alinhamento facial multi-nível baseada em redes neurais convolucionais. A abordagem divide a análise facial em dois níveis: 1) detecção de subunidades faciais; e 2) localização de pontos de referência das subunidades. Algumas hipóteses foram formuladas para verificação da abordagem desenvolvida:

- A uniformização de rotação das imagens contribui na redução das diferenças de desempenho entre grupos;
- A divisão da modelagem facial fundamentada nos elementos que compõem a face reduz as diferenças de desempenho entre grupos;
- A adição de amostras dos grupos desfavorecidos na modelagem facial reduz as diferenças de desempenho entre grupos.

1.2 Objetivos

O objetivo geral desta pesquisa é desenvolver uma abordagem multi-nível de alinhamento facial baseada em redes neurais convolucionais que reduza as diferenças de desempenho entre grupos clínicos observadas nos modelos atuais sem que haja prejuízo no desempenho geral.

1.2.1 Objetivos Específicos

Para alcançar o objetivo geral, os seguintes objetivos específicos foram perseguidos:

- Caracterizar e medir as diferenças de desempenho de modelos do estado da arte do alinhamento facial aplicados em populações clínicas;
- Desenvolver um modelo de normalização facial para pré-processamento;
- Desenvolver um modelo de detecção de subunidades faciais;

- Desenvolver modelos de localização de pontos de referência facial para os elementos da face;
- Validar e integrar os modelos preditivos desenvolvidos;
- Comparar a abordagem desenvolvida com o estado da arte de alinhamento facial.

1.3 Estrutura do Trabalho

Os demais capítulos desta tese estão organizados da seguinte forma:

- O Capítulo 2 elenca e discute brevemente os trabalhos relacionados aos dois aspectos fundamentais desta pesquisa: alinhamento facial e viés algorítmico na análise facial;
- O Capítulo 3 apresenta os conceitos teóricos necessários para concretização da pesquisa, incluindo métodos e métricas de avaliação;
- O Capítulo 4 descreve todo o fluxo da abordagem de alinhamento facial desenvolvida para a redução do viés e os *datasets* explorados;
- O Capítulo 5 mostra os resultados atingidos com a abordagem desenvolvida;
- O Capítulo 6 apresenta as conclusões sobre a abordagem, aponta problemas observados que devem ser resolvidos, e possíveis caminhos para trabalhos futuros sobre o tema.

2 Trabalhos Relacionados

As primeiras pesquisas desenvolvidas para a solução do problema de localização de pontos faciais fundamentavam-se em modelos faciais deformáveis que estimavam os pontos de referência com base na textura local, e recorriam a algoritmos estatísticos para correção global da forma (WANG et al., 2018). Durante os anos seguintes esses modelos receberam diversos aprimoramentos até serem superados por abordagens baseadas em redes neurais convolucionais (*CNNs*, do inglês *convolutional neural networks*). Os modelos mais atuais utilizam redes neurais convolucionais para o desenvolvimento de métodos majoritariamente assentados em duas estratégias: regressão direta do vetor de coordenadas; ou regressão de *heatmaps* (KHABARLAK; KORIASHKINA, 2022). Muito embora o desempenho geral desses modelos tenha atingido níveis próximos à anotação do especialista, ainda existem lacunas a serem investigadas no tema. Trabalhos recentes lançaram luz sobre o problema do viés algorítmico na análise facial e dos efeitos danosos que esse tipo de distorção pode provocar em um modelo de aprendizagem de máquina. O viés algorítmico, oriundo tanto do processo de geração dos conjuntos de dados quanto do projeto e construção dos modelos preditivos, tem sido indicado como um importante caminho para pesquisas sobre o tema (DROZDOWSKI et al., 2020).

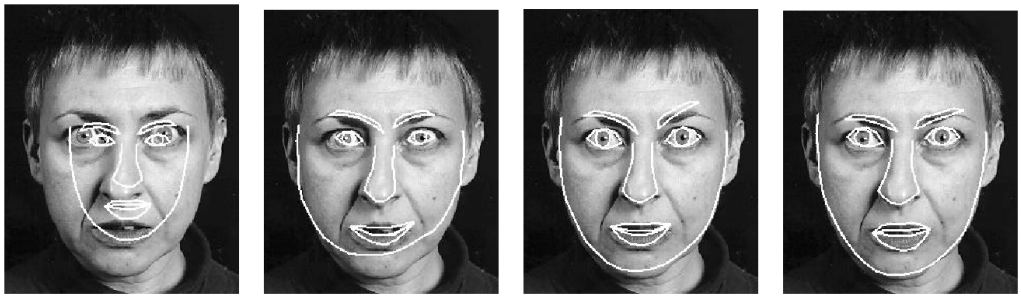
Neste capítulo são enumerados e discutidos os avanços alcançados no problema do alinhamento facial, com ênfase nas metodologias mais recentes, e os principais trabalhos relacionados ao viés algorítmico no contexto da análise facial.

2.1 Alinhamento Facial

A concepção de modelos preditivos para determinação das coordenadas de pontos de referência da face é uma tarefa que vem sendo perseguida por pesquisadores em Visão Computacional ao longo de quase três décadas. Cootes e Taylor (1992) propuseram o *Active Shape Models* (*ASM*), um método de construção de modelos deformáveis que aprende os padrões de variabilidade de formas a partir de um conjunto de imagens marcadas com pontos de referência. A formulação do *ASM* define como os modelos devem ser aplicados em novas imagens de maneira iterativa para busca pela forma alvo com base no gradiente local, seguido da aplicação de restrições que garantem que a instância deformável se mantenha coerente ao conjunto de treinamento. A aplicação do *ASM* foi demonstrada inicialmente para a localização do contorno das mãos e de resistores. O método foi refinado para aplicação em múltiplas resoluções e demonstrado na busca facial (COOTES; TAYLOR; LANITIS, 1994). Posteriormente, a

função de busca local foi substituída pela distância de *Mahalanobis* (COOTES; TAYLOR, 2001). A Figura 1 ilustra o encaixe de um modelo facial com aplicação do *ASM*.

Figura 1 – Busca facial com *Active Shape Models*.



Fonte: (COOTES; TAYLOR, 2001).

A partir de então, muitas abordagens de alinhamento facial propuseram melhorias na localização dos pontos a partir da formulação original do *ASM*. Du et al. (2008) propuseram uma alteração na maneira como novas estimativas são geradas durante a aplicação dos modelos. A ideia mais estabelecida do *ASM* consiste em calcular novas coordenadas para os pontos com base na distância de *Mahalanobis* entre a amostra de textura dos pontos e o modelo de textura. Contudo, existem casos em que esse modelo não segue uma distribuição gaussiana, o que restringe a eficácia dessa abordagem. Para superar esse problema, eles propuseram a utilização de máquinas de vetores de suporte na classificação das texturas correspondentes a cada ponto. A aplicação dessa proposta foi capaz de melhorar tanto os resultados gerais na localização dos pontos, quanto a frequência de convergência do método quando comparado ao *ASM* clássico.

Seshadri e Savvides (2009) apresentaram um método mais robusto para a localização de pontos em imagens faciais frontais a partir de uma versão modificada do *ASM*. Além de aperfeiçoar o modelo com o aumento do número de pontos de referência, eles também atacaram o problema do cálculo de novas coordenadas para os pontos faciais a cada iteração. Ao invés de calcular a distância de *Mahalanobis* com base na média do perfil de textura dos pontos, eles utilizaram um vetor reconstruído a partir da projeção do perfil de textura (amostrado em fase de aplicação do método) no subespaço de textura modelado para aquele ponto. Essas alterações também impactaram positivamente a localização de pontos faciais em relação ao *ASM* clássico.

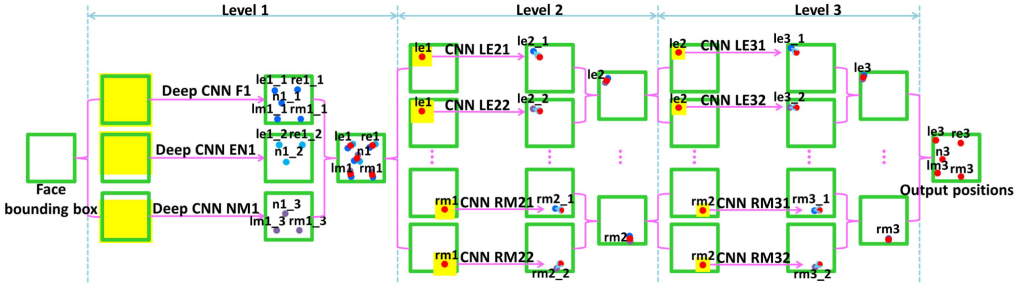
Milborrow (2007) também investigou e propôs aprimoramentos ao método *ASM*. Ele propôs as seguintes extensões ao *ASM* clássico: selecionar sub-conjuntos de pontos de referência para modelagem de textura em 2D; aplicar instâncias em séries de diferentes resoluções; aumentar o número de pontos que compõem o modelo; ignorar valores das matrizes de covariância de alguns perfis de textura. O melhor modelo gerado foi testado no *dataset BioId* (JESORSKY; KIRCHBERG; FRISCHHOLZ, 2001),

no qual atingiu resultados gerais melhores que o *ASM* clássico.

Nos últimos anos, as abordagens de alinhamento facial baseadas em aprendizagem profunda têm superado os métodos tradicionais de aprendizagem de máquina. Considerando que o desempenho geral de um modelo é o resultado que ele atinge quando aplicado na totalidade de um *dataset*, é possível observar que as abordagens mais destacadas do alinhamento facial estão divididas em duas categorias de *CNNs*: métodos de regressão de coordenadas; e métodos baseados em mapas de ativação de classe (*heatmaps*) (HSU et al., 2020).

Sun, Wang e Tang (2013) propuseram um método de regressão de coordenadas em cascata com 3 níveis baseado em redes neurais convolucionais. O método dos autores inicia a localização de 5 pontos de referência facial no primeiro nível e refina os resultados nos níveis posteriores. A Figura 2 ilustra como a abordagem opera em cascata para localização e refinamento dos pontos modelados.

Figura 2 – Abordagem em cascata para localização de pontos faciais.



Fonte: (SUN; WANG; TANG, 2013).

Bulat e Tzimiropoulos (2017b) apresentaram a *Face Alignment Network* (*FAN*). Trata-se de uma arquitetura de rede neural convolucional para detecção de pontos faciais baseada em regressão de *heatmaps* capazes de localizar pontos em 2D ou 3D. A *FAN-2D* foi testada em alguns dos principais *benchmarks* de alinhamento facial (*300W*, *Menpo*, etc) e comparada com modelos do estado da arte, produzindo resultados gerais superiores.

Wang, Bo e Fuxin (2019) também elaboraram um trabalho de alinhamento facial baseado em *heatmaps*. Os autores propuseram uma nova função de perda para a regressão dos mapas de ativação de classe, denominada *Adaptive Wing*. O método adapta a forma da função de perda de acordo com diferentes tipos de *pixels* existentes no *heatmap*. A definição do *pixel* é feita por uma outra função de acordo com o plano ao qual ele pertence dentro do *heatmap*. Tal função, denominada *Weighted Loss Map*, foi também proposta na pesquisa de Wang, Bo e Fuxin (2019). O modelo alcançou o estado da arte nos *bechmarks COFW* (BURGOS-ARTIZZU; PERONA; DOLLAR, 2022), *300W* (SAGONAS et al., 2016) e *WFLW* (WU et al., 2018).

Kumar et al. (2020) apresentaram um novo *framework* para execução de múltiplas tarefas associadas ao alinhamento facial: localização de pontos; cálculo de incerteza na localização; e estimativa de visibilidade de cada ponto. Para tanto, os autores introduziram três novos componentes em uma arquitetura *DU-Net* (TANG et al., 2020): um estimador médio de coordenadas dos pontos faciais que computa uma média ponderada a partir dos valores positivos do *heatmap* correspondente; uma rede para estimar incerteza na localização baseada na decomposição de Cholesky; e outra rede para estimar a visibilidade do ponto. Os três novos elementos compõem a função de perda proposta pelos autores e chamada de *LUVLi*. O trabalho também produziu uma anotação adicional ao *dataset AFLW* (KÖSTINGER et al., 2011) com rótulos de visibilidade para todos os pontos das imagens faciais, denominada *MERL-RAV*.

Huang et al. (2020) fundamentaram seu modelo de alinhamento facial na estrutura de *hourglasses* empilhados de Bulat e Tzimiropoulos (2017a) visando aumentar a robustez com respeito a imagens faciais em ambientes não-controlados. Na arquitetura *PropagationNet* proposta, cada *hourglass* produz um mapa de atributos para o *hourglass* seguinte juntamente com os *heatmaps* dos pontos modelados treinados com as marcações. Na sequência, um módulo de propagação gera *heatmaps* de bordas e outro mapa de atributos para o *hourglass* posterior. A cadeia é composta por quatro *hourglasses* acompanhados dos módulos de propagação em sequência. O modelo alcançou o estado da arte nos *bechmarks COFW*, *300W* e *WFLW*.

Xu et al. (2021) focaram seu trabalho em ambientes não-controlados, especialmente com grande variação de posições faciais. Eles utilizaram uma estratégia de divisão-e-conquista na qual, primeiramente, o espaço da imagem facial é dividido com auxílio de *templates* que funcionam como âncoras de referência para a regressão. A segunda etapa consiste na agregação dos resultados de estimativa de cada *template*. O método *AnchorFace* utiliza como *backbone* a *ShuffleNet-V2* (MA et al., 2018) para regressão dos deslocamentos dos *templates*, e para uma estimativa de confiança que é utilizada como peso na etapa de agregação. O método alcançou o estado da arte nos *datasets AFLW*, *300W*, *Menpo*, e *WFLW*.

Bulat, Sanchez e Tzimiropoulos (2021) apontaram problemas induzidos pelo processo de discretização tanto na codificação quanto na decodificação do *heatmaps* em métodos que exploram esta abordagem. A geração de *heatmaps* com imagens em resolução original é geralmente um processo proibitivo por questões de escala. A alternativa mais usual consiste na aplicação de escala e corte na imagem, tipicamente para resoluções de 64×64 *pixels*. Assim, o vetor de coordenadas dos pontos deve ser transformado utilizando os mesmos parâmetros. Após aplicação das transformações, as novas coordenadas precisam ser discretizadas para refletir a localização dos pontos no novo espaço. Durante a codificação dos *heatmaps*, evita-se a quantização gerando

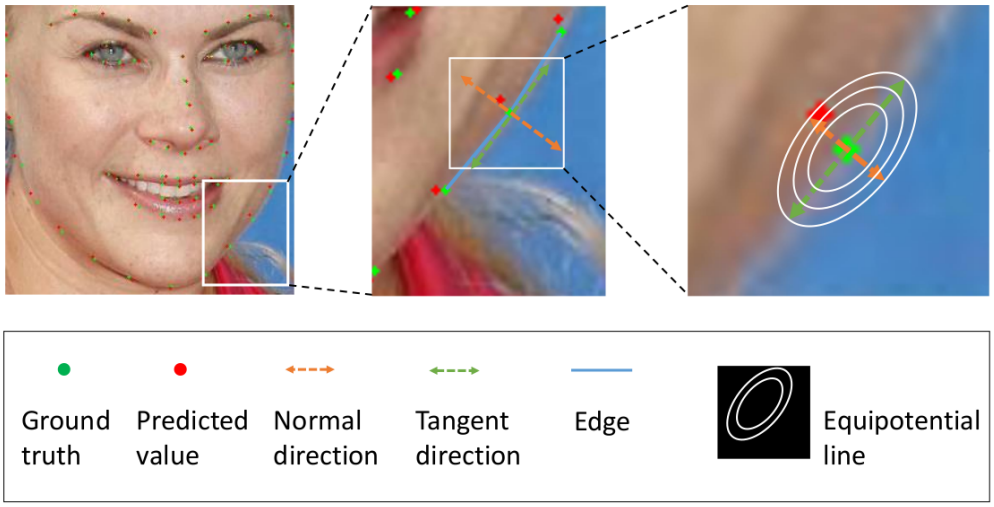
uma Gaussiana nas coordenadas em escala, e uma amostra é retirada em forma de *grid*. Para decodificar o *heatmap*, deve ser feita uma busca pelo máximo local, uma extração de amostra em torno dele, e a aplicação do *soft-argmax* local. O desempenho geral do método supera o estado da arte em *datasets* como *300W*, *WFLW*, e *AFLW-19*, considerando as métricas do erro médio normalizado (*NME*, do inglês *normalized mean error*), área sob a curva do erro acumulado (*AUC*, do inglês *area under the curve*), e a taxa de falha (TF).

Jin, Liao e Shao (2021) apontaram três problemas com modelos baseados em regressão de *heatmaps*: alto custo computacional; ausência, em geral, de restrições de forma explícitas; e diferenças de desempenho sensíveis ao domínio. A proposta dos autores baseia-se na regressão de *heatmaps*, contudo, sem necessidade de *upsampling*, o que reduz o custo computacional. Os mapas de atributos em baixa resolução apontam os *grids* mais prováveis para as coordenadas dos pontos. Para cada *grid*, é feita uma estimativa de deslocamento, com origem no canto esquerdo superior, visando refinamento da localização. O método é realizado em uma única fase, pois as estimativas de *grid* e deslocamento podem ser realizadas em paralelo. O *PIPNet* se mostrou preciso e robusto quando comparado ao estado da arte, com destaque para seu baixo custo computacional.

Huang et al. (2021) investigaram um tipo de erro intimamente vinculado à rotulação ambígua de alguns pontos faciais. A determinação do vetor de coordenadas dos pontos que é realizada pelo especialista em uma imagem facial pode render divergências a depender de fatores como a qualidade da imagem, a descrição do ponto, a textura da face, etc. Pontos faciais que determinam o contorno da mandíbula, por exemplo, podem facilmente apresentar desafios para a marcação manual. Nesses casos, eles observaram que a introdução de um viés de rotulação torna-se mais provável e tende a dificultar a convergência dos modelos, ao passo que prejudica o desempenho das metodologias. Visando reduzir esse problema, os autores propuseram para seu modelo – *ADNet* (*Anisotropic Direction Network*) – uma nova função de perda baseada em direção anisotrópica e um módulo de atenção anisotrópica para a regressão dos *heatmaps*. O desempenho atingiu o estado da arte nos *datasets* 300W, *WFLW*, e *COFW*. A Figura 3 exemplifica como a estimativa do ponto tende a ser corrigida ao longo da direção normal em detrimento da tangente.

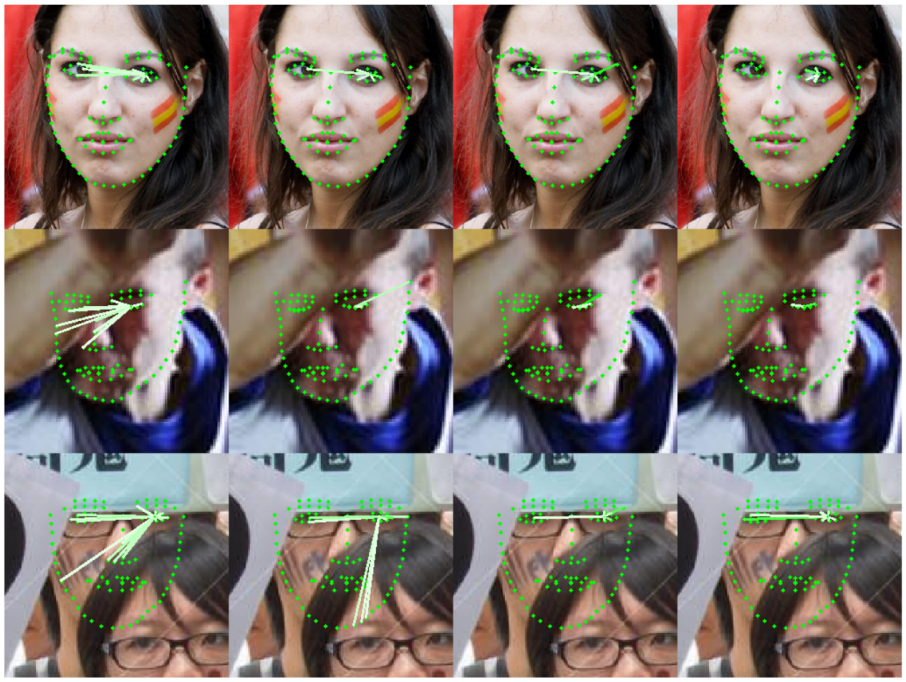
Prados-Torreblanca, Buenaposada e Baumela (2022) apresentaram o *SPIGA* (*Shape Preserving with Graph Attention Networks*), um regressor de pontos faciais que combina uma *CNN* com *Graph Attention Networks* (*GATs*) em cascata. A CNN fornece a representação local da aparência e cada regressor *GAT* lida com a codificação posicional e com o mecanismo de atenção que aprende as relações geométricas entre os pontos para forçar o modelo a produzir formas faciais coerentes. O desempenho do

Figura 3 – Diagrama da função de perda da *ADNet*.



Fonte: (HUANG et al., 2021).

Figura 4 – Exemplos das matrizes de adjacências por módulo da *GAT*.



Fonte: (PRADOS-TORREBLANCA; BUENAPOSADA; BAUMELA, 2022).

SPIGA atingiu o estado da arte nos *datasets WFLW*, *COFW-68* e *MERL-RAV* (*ALFW*). A Figura 4 mostra alguns exemplos de como as matrizes de atenção são computadas em cada módulo para composição das redes de atenção.

A Tabela 1 sumariza os resultados obtidos pelos modelos de alinhamento facial mais recentes nos conjuntos de teste dos *datasets 300W* e *WFLW*. Os modelos apresentados foram construídos a partir dos subconjuntos de treino de cada *dataset*,

com exceção dos modelos testados no *300W* dos trabalhos de Bulat e Tzimiropoulos (2017b), Kumar et al. (2020), e de Bulat, Sanchez e Tzimiropoulos (2021), que foram treinados com base no *300W-LP-2D* (BULAT; TZIMIROPOULOS, 2017b). Os melhores resultados por *dataset* e métrica estão em negrito.

Tabela 1 – Desempenhos dos modelos de alinhamento facial mais recentes nos *datasets 300W* e *WFLW*. Em vermelho, os erros de localização foram normalizados com base na média geométrica de altura e largura da caixa delimitadora da face. Os demais resultados (em preto) consideram a distância inter-ocular para normalização

Teste	**Referência**	**Métrica**		
		NME %	**AUC 7%**	
300W	Bulat e Tzimiropoulos (2017b) *	2,32	0,665	
	Kumar et al. (2020) *	2,10	0,702	
	Bulat, Sanchez e Tzimiropoulos (2021) *	**2,04**	**0,711**	
	Wang, Bo e Fuxin (2019)	3,07	-	
	Huang et al. (2020)	**2,93**	-	
	Xu et al. (2021)	3,72	-	
	Jin, Liao e Shao (2021)	3,19	-	
	Huang et al. (2021)	**2,93**	-	
	Prados-Torreblanca, Buenaposada e Baumela (2022)	2,99	-	
			AUC 10%	**TF 10%**
WFLW	Wang, Bo e Fuxin (2019)	4,36	0,5719	2,84
	Kumar et al. (2020)	4,37	0,5770	3,12
	Huang et al. (2020)	4,05	0,6158	2,96
	Bulat, Sanchez e Tzimiropoulos (2021)	**3,72**	**0,6310**	**1,55**
	Xu et al. (2021)	4,62	0,5516	4,20
	Jin, Liao e Shao (2021)	4,31	-	-
	Huang et al. (2021)	4,14	0,6022	2,72
	Prados-Torreblanca, Buenaposada e Baumela (2022)	4,06	0,6056	2,08

* Modelos treinados com a extensão *300W-LP-2D*

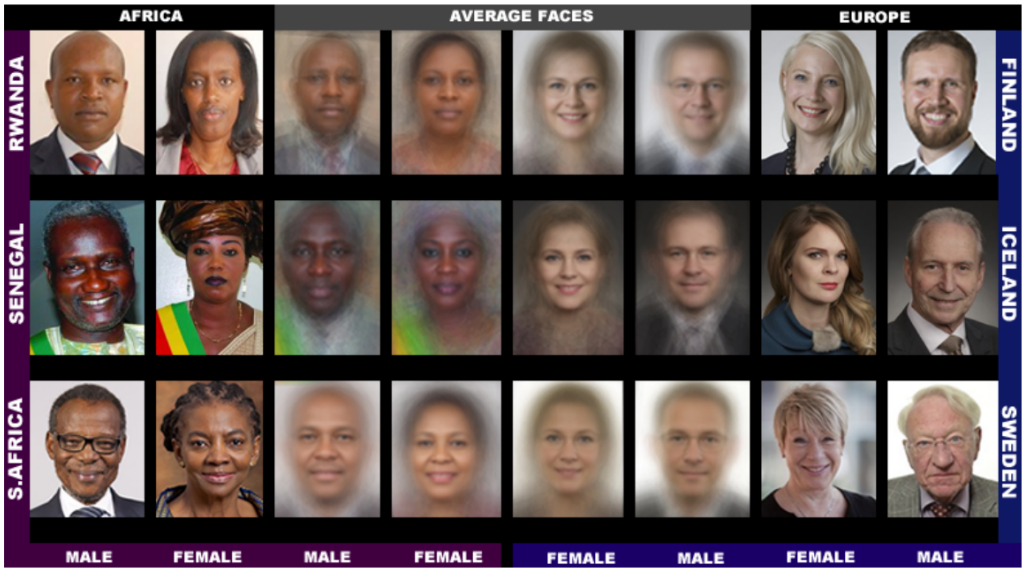
2.2 Viés Algorítmico e Análise Facial

Métodos baseados em aprendizagem de máquina têm sido amplamente explorados na construção de sistemas computacionais para tomada ou auxílio em decisões que afetam diretamente seus usuários. Eles estão presentes na identificação e marcação automática de pessoas em redes sociais, na decisão sobre a aprovação ou rejeição de exames clínicos, no auxílio à concessão de empréstimos bancários, em aplicações que ajudam no controle de fronteiras entre países, etc. A utilização desses modelos tem trazido rapidez e facilidades para os usuários desses sistemas em questões que variam do entretenimento à segurança. Entretanto, a concepção de tais modelos pode embutir comportamentos indesejados com sérias implicações sociais nessas soluções. Um problema que tem sido observado recentemente em modelos de aprendizagem de máquina, e que tem atraído bastante atenção dos pesquisadores é a introdução de viés algorítmico, ou viés demográfico, nesses modelos.

Buolamwini e Gebru (2018) avaliaram a presença de viés algorítmico em 3

importantes soluções comerciais para classificação de gênero em imagens faciais. Para verificação e caracterização do viés nessas soluções, foi elaborado o *Pilot Parliaments Benchmark* (*PPB*), um *dataset* composto por 1.270 indivíduos de três países africanos (Ruanda, Senegal, e África do Sul) e três países europeus (Islândia, Finlândia, e Suécia). O *dataset* é balanceado com base no gênero (masculino e feminino) e tipo de pele dos indivíduos (pele mais clara e pele mais escura). As soluções computacionais de classificação de gênero avaliadas pertencem a *API*s (*application programming interface*) proprietárias da *Microsoft*, *IBM*, e *Face++*. Os resultados das avaliações indicaram que os classificadores tiveram melhor desempenho no grupo de indivíduos do sexo masculino de pele mais clara, ao passo que desempenharam pior em indivíduos do sexo feminino de pele mais escura. A Figura 5 exemplifica os diferentes grupos do *benchmark PPB*.

Figura 5 – Exemplos do *dataset Pilot Parliaments Benchmark*.



Fonte: (BUOLAMWINI; GEBRU, 2018).

Hupont e Fernández (2019) quantificaram o desbalanceamento demográfico em *datasets* públicos de análise facial em relação a gênero e etnia. Eles demonstraram que o viés demográfico impacta negativamente os desempenhos dos modelos do estado da arte de reconhecimento facial (*FaceNet*, *SphereFace*, e *VGGFace2*). Como contribuição ao estudo e desenvolvimento das abordagens de reconhecimento facial, eles elaboraram o *DemogPairs*, um *dataset* composto por 10,8 mil imagens de indivíduos balanceado com base em gênero, etnia, e identidade.

Georgopoulos, Panagakis e Pantic (2020) afirmaram que o viés demográfico observado em métodos de análise facial baseados em aprendizagem profunda é oriundo tanto de limitações de diversidade dos *datasets* explorados na concepção dos modelos, quanto da própria elaboração dos algoritmos. Eles investigaram a presença

de viés demográfico no estado da arte de reconhecimento facial, estimativa de idade, e classificação de gênero. Para tanto, eles desenvolveram um *dataset* de imagens e vídeos direcionado à análise facial em ambientes não-controlados denominado *KANFace*. Seus experimentos demonstraram que os modelos de reconhecimento facial e estimativa de idade apresentaram viés mais acentuado para indivíduos com idade abaixo de 18 anos e acima de 60. No caso de classificação de gênero, o viés se manifestou majoritariamente em indivíduos do sexo masculino mais jovens.

Drozdowski et al. (2020) apresentaram uma vasta compilação de estudos sobre viés demográfico em biometria. Eles apontaram várias questões ainda sem solução dentro da temática tais como: a quantidade limitada de *datasets* e *benchmarks* balanceados de acordo com atributos demográficos; a deficiência de trabalhos teóricos que abordam as noções de viés e justiça de uma forma mais rigorosa; e a ausência de abordagens estritamente teóricas para comprovar a redução de viés em métodos e *datasets*.

Suresh e Guttag (2021) buscaram enriquecer a discussão no campo teórico sobre o viés algorítmico no contexto dos modelos de aprendizagem de máquina. Nesse trabalho, eles conferiram mais profundidade ao entendimento de como o viés é afetado pelo conjunto de dados usado no treinamento desses modelos, e dos prejuízos que esse viés pode provocar em várias etapas do ciclo de vida dos mesmos. Eles propuseram a discriminação da ideia de viés algorítmico em 7 conceitos mais granulares que dizem respeito à origem do viés:

- Viés histórico - representa um viés do mundo real introduzido nos *datasets* de treino e validação;
- Viés de representação - ocorre quando há sub-representação de parte da população nos *datasets*;
- Viés de medição - surge no momento da definição de atributos e rótulos que são abstratos mas que derivam de valores observáveis. Índices para concessão de crédito, por exemplo, são conceitos abstratos que precisam ser construídos a partir de elementos observáveis. A definição desses elementos pode introduzir o viés durante a construção dos *datasets*;
- Viés de agregação - é oriundo de modelos muito genéricos que são aplicados a grupos que apresentam caraterísticas distintas;
- Viés de aprendizado - ocorre quando as escolhas da modelagem aumentam as diferenças de desempenho em grupos distintos. Determinar a função de custo de um modelo é um exemplo de escolha que pode introduzir esse viés;

- Viés de avaliação - surge quando um *dataset* de teste usado para avaliar o modelo não corresponde à população de uso. Esse viés pode introduzir problemas ainda maiores que outras fontes de viés, pois um *benchmark* com problemas de representação pode encorajar o desenvolvimento de metodologias que tenham bom desempenho apenas no subconjunto representado por ele;
- Viés de implantação - ocorre quando há divergência entre o problema para o qual um modelo foi elaborado e a sua real utilização em um sistema.

O objetivo desse trabalho foi a formalização de um *framework* para entendimento de como cada tipo de viés pode introduzir aspectos danosos no *pipeline* de um modelo de aprendizagem de máquina genérico, fornecendo subsídios para o projeto de abordagens que satisfaçam a noção de justiça.

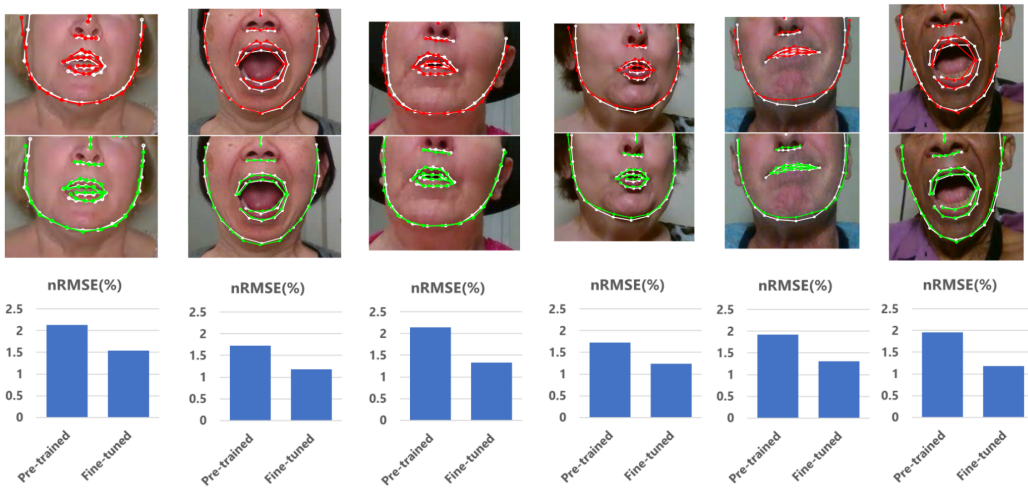
DeAlcala et al. (2023) propuseram uma nova forma para medir o viés presente em modelos de aprendizagem de máquina a partir de uma abordagem estatística baseada no método N-Sigma. Para verificar a qualidade do instrumento de medição proposto, eles fizeram ajuste fino em um modelo da rede *ResNet-100* (HE et al., 2016) treinado com a base *MS1Mv3* (GUO et al., 2016). O ajuste teve como objetivo introduzir viés de representação no modelo de tal forma que fosse possível quantificá-lo após aplicação no *dataset* alvo. Os experimentos foram conduzidos no *dataset Racial Faces in the Wild* (WANG et al., 2019) e demonstraram a capacidade do método N-Sigma em produzir valores que representassem, de maneira mais acessível, as diferenças de resultados geradas por modelos com viés algorítmico.

Em relação ao viés demográfico na localização de pontos de referência facial, o número de pesquisas ainda é bastante modesto. Taati et al. (2019) investigaram o viés que afeta o desempenho de algoritmos de detecção de pontos faciais em adultos com demência, e buscaram formas de mitigar as diferenças observadas. Eles descobriram que os métodos de alinhamento facial têm desempenho consideravelmente pior quando aplicados a imagens faciais de adultos com demência, em comparação com o outro grupo avaliado composto por idosos cognitivamente saudáveis. Os autores tentaram aprimorar os modelos avaliados com amostras de imagens do grupo de idosos sem problemas neurológicos, contudo, a estratégia não reduziu as diferenças de desempenho dos modelos entre os grupos, apesar de ter melhorado o desempenho geral.

Bandini et al. (2021) investigaram o viés inerente a métodos clássicos de aprendizagem de máquina e a modelos do estado da arte do alinhamento facial. Eles desenvolveram e publicaram o *Toronto Neuroface*, um *dataset* com imagens obtidas em vídeo de indivíduos com assimetria orofacial causada por transtornos neurológicos. O *dataset* é composto por 3.303 *frames* de vídeo com imagens faciais capturadas

em ambiente controlado de dois grupos de indivíduos com problemas neurológicos: portadores de ELA (*ALS*, do inglês *amyotrophic lateral sclerosis*), e sobrevivente de derrame (*PS*, do inglês *post-stroke*); e um grupo de controle de adultos saudáveis (*HC*, do inglês *healthy control*). Durante a captura das imagens, os participantes realizaram expressões faciais típicas de exames clínicos. Cada face do *dataset* foi marcada seguindo o padrão de marcação de pontos faciais *Multi-PIE* (GROSS et al., 2010). Foram aplicados ao *dataset*: um modelo de alinhamento facial generativo (*AAM*); três modelos discriminativos (*CLM*, *ERT*, e *SDM*); e um modelo baseado em redes neurais convolucionais (*FAN-2D*). A análise estatística (KRUSKAL; WALLIS, 1952) dos resultados obtidos revelou que todos os modelos apresentaram significativas diferenças de desempenho (média do erro normalizado) entre os grupos avaliados, caracterizando assim o viés. Os autores implementaram uma estratégia de ajuste fino no modelo baseado em aprendizagem profunda (*FAN-2D*) com exemplares do *Toronto Neuroface*. Contudo, o que se observou foi apenas uma melhora no desempenho geral, confirmando que a inclusão de exemplares do próprio *benchmark* no conjunto de treino não surtiu efeito na diminuição do viés, tal como observado por Taati et al. (2019). A Figura 6 mostra alguns resultados obtidos pela *FAN-2D* no *dataset Toronto Neuroface*.

Figura 6 – Casos de teste da *FAN-2D* no *Toronto Neuroface* antes (marcações em vermelho) e depois (marcações em verde) do ajuste fino.



Fonte: (BANDINI et al., 2021).

2.3 Considerações Finais

A utilização de *CNNs* em métodos de localização de pontos faciais representou grande avanço no desempenho geral dos modelos ao longo da última década. A literatura recente estabelece claramente duas categorias de modelos de regressão: coordenadas; e *heatmaps*. De maneira geral, os modelos de regressão de coordenadas são mais simples e menos custosos computacionalmente. Contudo, os modelos basea-

dos em *heatmaps* se apresentam mais dominantes no estado da arte pois conseguem atingir desempenho superior geralmente em arquiteturas com um único estágio, ao passo que modelos de regressão de coordenadas precisam de dois ou mais estágios para atingir desempenho semelhante (JIN; LIAO; SHAO, 2021).

Entretanto, os modelos de avaliação adotados quase pela totalidade dos trabalhos relacionados de alinhamento facial leva em consideração apenas métricas como o erro médio normalizado ou a taxa de falha sobre todo o conjunto de teste ou *benchmark*. O grande avanço no desempenho geral dos modelos foi movido pela tentativa de solucionar imprecisões herdadas de ambientes não-controlados como oclusão, variações de iluminação, variações de pose facial, etc. Pouco se observou sobre os efeitos que os modelos apresentam em grupos de indivíduos distintos por alguma característica demográfica ou clínica, por exemplo.

Recentemente, questões relacionadas ao viés algorítmico introduzido nos modelos preditivos vieram à tona e revelaram os prejuízos que podem surgir das suas aplicações. É importante notar que o viés, mesmo que seja originário de uma modelagem correta, pode introduzir aspectos danosos aos modelos de aprendizagem de máquina ao reproduzir alguma caraterística desigual e injusta do mundo real. Os trabalhos mais recentes de alinhamento facial que focaram em investigar o impacto do viés no desempenho dos métodos não obtiveram grande sucesso na redução desse problema, ou seja, os desempenhos apresentaram significativas divergências a depender do grupo analisado (TAATI et al., 2019; BANDINI et al., 2021). Outro aspecto importante de se observar é que os modelos avaliados foram concebidos a partir de *datasets* cuja composição estava sub-representada com respeito a indivíduos dos grupos minoritários, evidenciando um viés de representação. Ademais, mesmo com adição de amostras dos grupos sub-representados não foi possível observar grandes avanços na redução do viés.

A abordagem de alinhamento facial desta pesquisa introduz uma modelagem baseada na divisão dos componentes da face em subunidades. A metodologia utiliza modelos intermediários quando comparada tanto às abordagens de regressão de pontos com base na textura global da face, quanto àquelas que fazem uso de textura local. O método desenvolvido tem algumas similaridades com o trabalho de Sun, Wang e Tang (2013) porque opera em diferentes níveis de abstração dos atributos faciais e utiliza um *backbone* semelhante para a estrutura da rede. Contudo, esta abordagem se diferencia em dois pontos fundamentais: ao invés de realizar a detecção parcial dos pontos, o primeiro nível de regressão de coordenadas do método localiza os componentes faciais; e o segundo nível de regressão é baseado na textura de todo o componente facial detectado, em vez da textura local.

A ideia que fundamenta a abordagem escolhida para este trabalho leva em

consideração que os modelos dos componentes faciais são capazes de aumentar a capacidade de expressão quando combinados em um modelo de alto nível, o que pode ser benéfico para superar os problemas de representação de indivíduos com assimetria orofacial acentuada, uma vez que eles representam o grupo mais prejudicado. Em um panorama mais amplo, diferentemente dos métodos desenvolvidos até então, que visam somente a ganhos no desempenho geral dos modelos, a abordagem deste trabalho se preocupa, adicionalmente, com os efeitos danosos que o viés algorítmico pode provocar, seja ele originário dos *datasets* de treino ou do próprio projeto do método. Assim, o próprio *design* do método de alinhamento facial desenvolvido foi orientado no sentido de compensar o viés presente nos *datasets*, objetivando resultados mais igualitários entre os grupos envolvidos.

No próximo capítulo são apresentados os elementos teóricos necessários para formulação e avaliação de um novo modelo de alinhamento facial. Os elementos apresentados são necessários tanto para o projeto da metodologia em questão, que visa primariamente à redução dos efeitos negativos do viés, quanto para a avaliação dos resultados da abordagem.

3 Fundamentação Teórica

Este capítulo apresenta os conceitos teóricos usados no desenvolvimento da abordagem de alinhamento facial visando à redução das diferenças de desempenho entre grupos clínicos. Assim, faz-se necessária a formalização da modelagem facial destacando os pontos de referência usados e as subunidades da face. Ademais, uma vez que o trabalho envolve um problema de Visão Computacional, são explorados conceitos básicos sobre a representação da imagem digital, bem como as operações abordadas na construção do método, e um conceito base para o elemento central da pesquisa: as redes neurais convolucionais. Para finalizar, são apresentadas as métricas comumente utilizadas na avaliação de modelos de alinhamento facial, e os testes estatísticos explorados para caracterizar as diferenças de desempenho dos modelos.

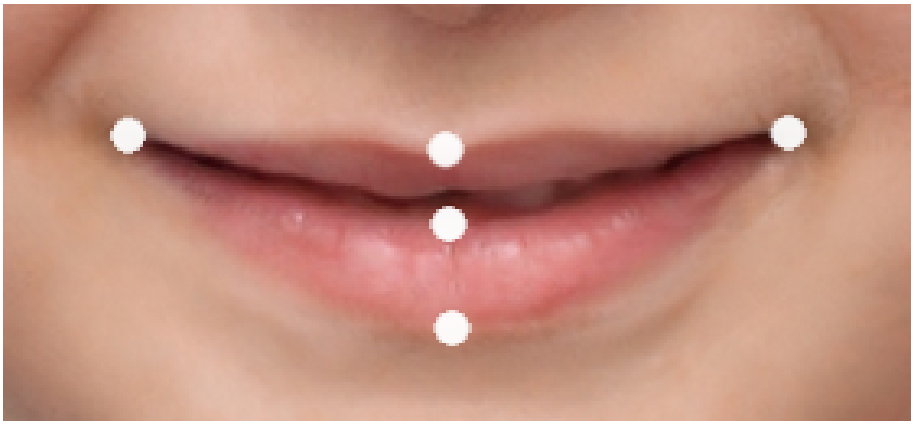
3.1 Pontos de Referência Facial

A análise facial é um tema que envolve importantes problemas tais como reconhecimento facial, rastreamento facial, reconhecimento de expressão facial, estimativa de pose. A solução dessas tarefas é capaz de fornecer informações valiosas sobre identidade e comportamento de indivíduos para muitas aplicações e sistemas computacionais. Tais informações podem ser extraídas diretamente a partir da textura da imagem facial, considerando toda a informação de uma face detectada. Além disso, a precisa localização dos elementos faciais é capaz de contribuir para a determinação de expressões e estimativa de pose facial (JOHNSTON; CHAZAL, 2018). Nesse contexto, os pontos de referência facial têm papel fundamental na correta localização dessas estruturas. Um ponto de referência facial representa uma região na face que tem uma característica distinguível. Ele é representado por coordenada mapeada do mundo real para o domínio da imagem.

Para dar suporte àquelas soluções de mais alto nível (identificação facial, identificação de expressões faciais, estimativa de gênero, etc), outras tarefas geralmente são empregadas dentro do fluxo de análise facial, seja como um passo inicial, seja como um caminho intermediário. A detecção facial, por exemplo, é considerada um ponto de partida para qualquer tarefa de análise facial (ZAFEIRIOU; ZHANG; ZHANG, 2015). Analogamente, a localização de pontos faciais pode ser considerada como um importante e essencial passo intermediário para muitas tarefas de análise facial (JIN; TAN, 2017). Nas aplicações de localização de pontos faciais (problema também conhecido como alinhamento facial), geralmente é pré-determinado um conjunto de pontos faciais alvo que são definidos com base no centro ou no contorno de compo-

nentes da face, como os olhos, a boca, ou o nariz. A Figura 7 ilustra um conjunto de 5 pontos de referência para a definição da boca. É importante notar que, no contexto do elemento facial, todos os pontos devem ser distinguíveis entre si. Cootes et al. (1995) definiram uma classificação de pontos de referência que devem ser modelados para compor um modelo de forma ativa (*ASM*). A definição tem três classes: 1) pontos de referência dependentes da aplicação; 2) pontos não dependentes da aplicação mas que representam um extremo local na forma; e 3) interpolações de outros pontos. Tal classificação é útil na definição de quais pontos de referência devem ser incluídos em um modelo facial.

Figura 7 – Exemplo de pontos de referência da boca.



Fonte: Elaborado pelo autor.

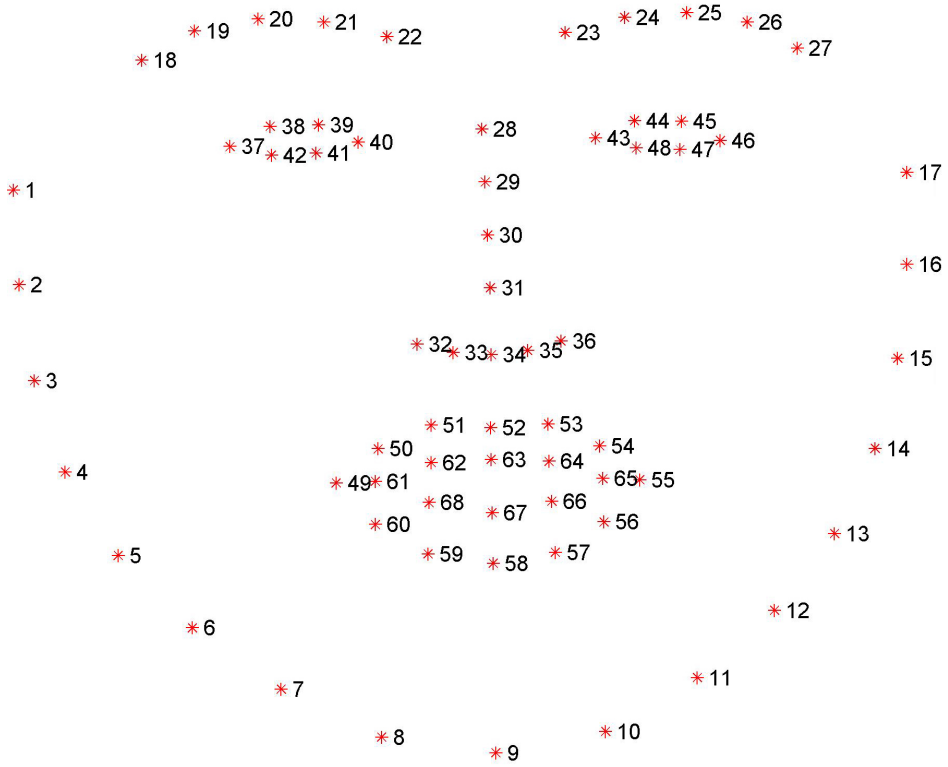
3.1.1 Padrão de Anotação de Pontos de Referência Facial

Os *datasets* de alinhamento facial são geralmente construídos a partir de conjuntos pré-determinados de pontos faciais que constituem padrões de marcação facial. As abordagens de localização de pontos faciais buscam adotar aqueles padrões que já são consolidados na temática. Um dos fatores que explicam a adoção de padrões e *datasets* já estabelecidos se deve à complexidade de rotulação dos conjuntos de imagens faciais. A rotulação desse tipo de material consiste em determinar um vetor $S = (x_1, y_1, x_2, y_2, \ldots, x_L, y_L)$ cujos pares (x_i, y_i) representam as coordenadas do i-ésimo ponto de referência de uma face na imagem, e L é a quantidade de pontos faciais (GOGIĆ; AHLBERG; PANDŽIĆ, 2021). Essa rotulação deve ser feita para cada exemplar dos conjuntos a serem utilizados no problema. Assim, a adoção de padrões e *datasets* já estabelecidos significa não somente a possibilidade de estudos comparativos mais precisos, mas também uma redução na carga de trabalho dos pesquisadores.

Os padrões mais utilizados atualmente são oriundos de *datasets* de alinhamento facial (*Helen*, *AFW*, *LFPW*, etc). Nesta pesquisa foi adotada a configuração de 68 pontos

estabelecida no *Multi-PIE* (GROSS et al., 2010) por ser um dos padrões de marcação mais utilizados na modelagem de soluções de alinhamento facial observados em trabalhos relacionados. A Figura 8 ilustra os pontos faciais usados nesse padrão de marcação. É possível notar que cada ponto possui um rótulo que o identifica dentro do conjunto de pontos adotado.

Figura 8 – 68 pontos de referência usados na marcação das imagens.



Fonte: (SAGONAS et al., 2013).

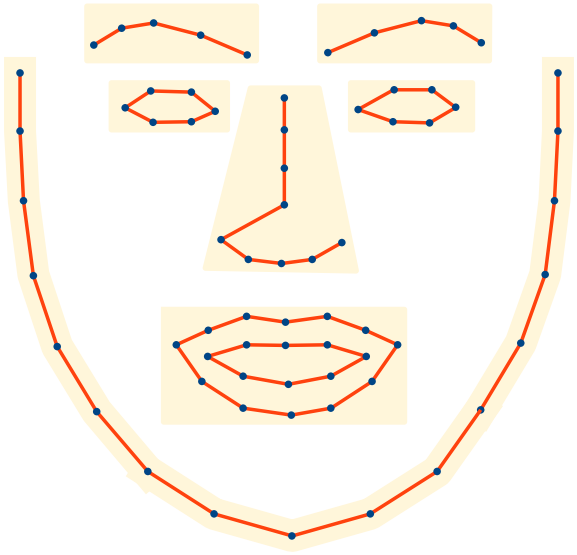
3.1.2 Subunidades Faciais

Uma subunidade facial pode ser definida como um subconjunto de pontos de referência da configuração de marcação adotada que representa um elemento facial, parte dele, ou um agregado de elementos. A Figura 9 ilustra um exemplo de divisão dos elementos faciais em subunidades a partir do padrão *Multi-PIE*.

3.2 Processamento Digital de Imagens

Os métodos de processamento digital de imagens (PDI) aplicam transformações em imagens digitais com dois objetivos principais: melhorar a representação da imagem original para o olhar humano; e processar a imagem de modo a facilitar seu armazenamento, transmissão, e sua representação para a percepção da máquina, a depender da aplicação (GONZALEZ; WOODS, 2018b).

Figura 9 – Exemplo de configuração de 7 subunidades faciais.



Fonte: Elaborado pelo autor.

Para tanto, tais métodos atuam no domínio das imagens digitais. Uma imagem pode ser definida como uma função bidimensional $f(x, y)$, onde x e y representam coordenadas espaciais em um plano, e a amplitude de f em qualquer ponto (x, y) representa a intensidade, ou nível de cinza, da imagem naquele ponto. Quando o domínio tanto das coordenadas quanto da amplitude da função é discreto, ou seja, estão definidos dentro de um espaço de valores finito, diz-se que a imagem é digital. Cada elemento da função dentro desse domínio discreto é conhecido como *pixel* (GONZALEZ; WOODS, 2018b).

Imagens coloridas naturais são geralmente oriundas de escâneres de cor ou de câmeras de video colorido. Tais dispositivos incorporam três sensores, cada um mais sensível a determinada faixa de comprimento de onda na porção de luz visível do espectro eletromagnético. Os sensores geram sinais de cor que são proporcionais às quantidades de vermelho, verde, e azul detectadas (PRATT, 2007). Após o processo de digitalização desses sinais, o resultado é uma imagem digital que possui três canais de cor, onde cada um é representado por uma função bidimensional no domínio discreto.

3.2.1 Transformações Geométricas

Transformações geométricas são operações usadas para transformar o arranjo dos *pixels* de uma imagem. As transformações geométricas em imagens digitais consistem basicamente de duas operações: transformação espacial das coordenadas; e interpolações que atribuem novos valores de intensidade aos *pixels* transformados

(GONZALEZ; WOODS, 2018b). A Equação 3.1 mostra como uma nova posição (x', y') pode ser mapeada a partir de uma matriz de transformação T:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = T \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}. \tag{3.1}$$

A transformação afim é feita a partir de uma matriz que inclui vários tipos de transformações. Em duas dimensões, ela tem por característica preservar pontos, linhas retas, e planos. Ela pode ser obtida a partir da Equação 3.1. Para incluir também a translação (transformação que move todos os pontos da imagem para uma mesma direção, a uma mesma distância) é necessário fazer uso de coordenadas homogêneas (GONZALEZ; WOODS, 2018b). A Equação 3.2 mostra a definição de uma matriz A de transformação afim com coordenadas homogêneas:

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = A \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}, \tag{3.2}$$

e a Tabela 2 resume as transformações geométricas exploradas nesta pesquisa.

Tabela 2 – Matrizes de transformações geométricas

Transformação	**Ação**	**Matriz**
Translação	Desloca a imagem em uma mesma direção	$\begin{bmatrix} 1 & 0 & t_x \\ 0 & 1 & t_y \\ 0 & 0 & 1 \end{bmatrix}$
Escala	Aumenta ou diminui a imagem	$\begin{bmatrix} c_x & 0 & 0 \\ 0 & c_y & 0 \\ 0 & 0 & 1 \end{bmatrix}$
Rotação	Rotaciona a imagem em torno da origem	$\begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix}$

3.2.2 Convolução

O conceito de convolução está intimamente relacionado ao de correlação, uma vez que tais operações exploram a mesma mecânica de ação. A correlação é definida matematicamente como mostrado na Equação 3.3:

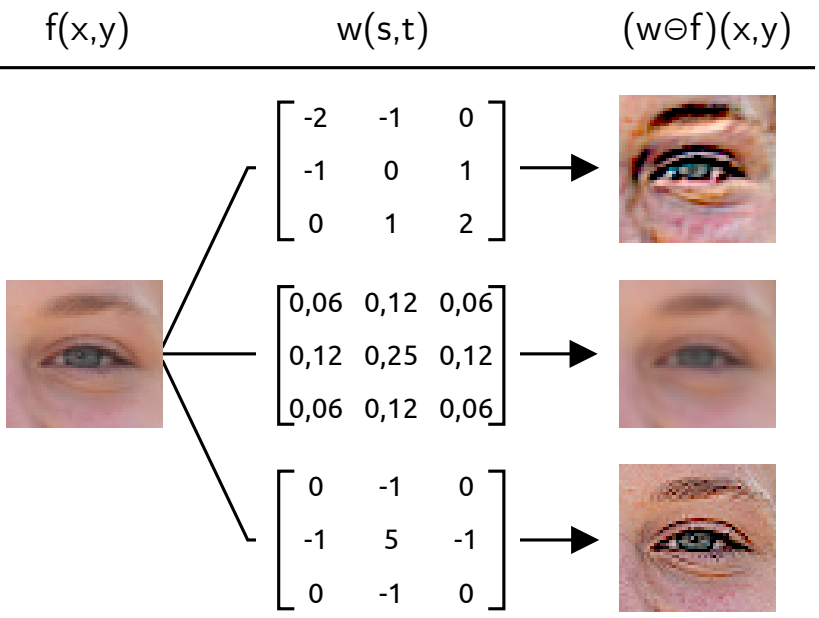
$$(w \oplus f)(x, y) = \sum_{s=-a}^{a} \sum_{t=-b}^{b} w(s, t) f(x + s, y + t), \tag{3.3}$$

e significa, no contexto do PDI, a aplicação de um filtro $w(s,t)$ (também chamado *kernel* ou máscara) sobre uma imagem $f(x,y)$, onde o centro de w move-se sobre a imagem e são computadas as somas dos produtos em cada coordenada. O processo de convolução é o mesmo com uma única diferença: o *kernel* w é rotacionado em 180° antes da operação. Matematicamente, a operação de convolução está expressa na Equação 3.4:

$$(w \ominus f)(x,y) = \sum_{s=-a}^{a} \sum_{t=-b}^{b} w(s,t) f(x-s, y-t). \tag{3.4}$$

É comum encontrar na literatura o uso do termo filtro de convolução, ou máscara de convolução, para denotar um *kernel* de filtragem espacial sem que necessariamente o *kernel* esteja sendo usado para convolução. O próprio termo "*convolução de um filtro com imagem*" é frequentemente utilizado para denotar o processo de soma de produtos feito na imagem sem haver diferenciação entre correlação e convolução (GONZALEZ; WOODS, 2018b). Um detalhe importante sobre o filtro é que ele deve ter lado ímpar, ou seja, se o tamanho de w é $M \times N$, então $M = 2a+1$ e $N = 2b+1$, onde a e b são inteiros. A Figura 10 ilustra alguns exemplos de *kernels* e os respectivos resultados das filtragens em uma imagem colorida.

Figura 10 – Exemplos de filtragem espacial.



Fonte: Elaborado pelo autor.

3.3 Redes Neurais Artificiais

As redes neurais artificiais (RNA) são modelos computacionais de aprendizado de máquina amplamente empregadas em problemas de classificação, regressão, agru-

pamento, etc. Assim como outros métodos de inteligência artificial, as RNA foram concebidas a partir de inspiração na natureza, especificamente no sistema nervoso do ser humano. O cérebro humano pode ser entendido como um computador altamente complexo e não-linear capaz de executar difíceis tarefas relacionadas à visão de maneira mais eficaz que os poderosos computadores atuais. As RNA possuem grande poder de generalização, o que significa que elas conseguem produzir saídas a partir de entradas não observadas durante o treino. Isso é possível por conta da maneira como a rede é estruturada (HAYKIN, 2009). As unidades elementares que compõem uma rede, chamadas de neurônios artificiais, são organizadas com interconexões similares à forma como os nerônios estão conectados no córtex visual de seres humanos (GONZALEZ; WOODS, 2018a).

3.3.1 Perceptron

O perceptron é a forma mais simples de rede neural usado para a classificação de padrões linearmente separáveis. Ele consiste de um único neurônio artificial com entradas que contêm pesos ajustáveis, um módulo de soma que computa a combinação linear das entradas ponderadas, um viés externo, e uma função de saída que aplica limites rígidos ao resultado do combinador linear (HAYKIN, 2009). O perceptron aprende uma fronteira linear entre duas classes cujos padrões são linearmente separáveis. Essa fronteira pode ser matematicamente interpretada, em um domínio bidimensional, como uma equação linear $y = ax + b$, onde a é a inclinação da reta e b indica o ponto de interseção da reta com o eixo vertical y. O parâmetro b não afeta a inclinação da reta e é comumente denominado viés.

Utilizando a equação geral da reta têm-se $x_2 + (w_1/w_2)x_1 + (w_3/w_2) = 0$, onde $y = x_2$, $x = x_1$, $a = w_1/w_2$, $b = w_3/w_2$. Seja $x_1, x_2, \ldots, x_n$ os componentes de um ponto no espaço n-dimensional, e $w_1, w_2, \ldots, w_n, w_{n+1}$ os coeficientes da fronteira que separa duas classes, onde o último representa o viés. A classificação do ponto é feita por avaliação da seguinte soma de produtos: $w_1x_1 + w_2x_2 + \cdots + w_nx_n + w_{n+1}$. Caso o resultado seja positivo, o ponto pertence à classe c_1, caso seja negativo, pertence a c_2.

O algoritmo que possibilita a aprendizagem do perceptron converge quando os pontos observados pertencem a classes linearmente separáveis. Nesse contexto, seja α uma taxa de correção, ou taxa de aprendizagem. O algoritmo de correção modifica os pesos iterativamente da seguinte forma:

- Se $x(k) \in c_1$ e $\mathbf{w}^T(k)\mathbf{x}(k) \leq 0$:

$$\mathbf{w}(k+1) = \mathbf{w}(k) + \alpha\mathbf{x}(k); \tag{3.5}$$

- Se $x(k) \in c_2$ e $\mathbf{w}^T(k)\mathbf{x}(k) \geq 0$:

$$\mathbf{w}(k+1) = \mathbf{w}(k) - \alpha\mathbf{x}(k); \tag{3.6}$$

- Caso contrário:

$$\mathbf{w}(k+1) = \mathbf{w}(k), \tag{3.7}$$

onde **w** é o vetor de coeficientes, **x** o vetor de coordenadas do ponto, e k é o índice de um passo iterativo do algoritmo de correção (GONZALEZ; WOODS, 2018a).

3.3.2 Rede Neural Multi-camadas

Um perceptron isolado tem uma limitada capacidade de classificação de dados, o que torna ineficaz seu uso para solução de problemas complexos observados no mundo real. Contudo, a concatenação de múltiplos perceptrons a partir de suas entradas e saídas aumenta a capacidade de generalização da estrutura. Uma rede neural multi-camadas consiste de neurônios artificiais dispostos em várias camadas: uma camada de entrada, uma de saída, e pelo menos uma camada oculta (HAYKIN, 2009). O neurônio artificial é bastante similar ao perceptron, com a diferença que a função de ativação é suavizada para evitar problemas de instabilidade nas interconexões. Assim, na estrutura da rede de neurônios artificiais interconectados, as entradas de um neurônio são valores ativados pelos resultados de uma função suavizada (sigmóide, tangente, ou *ReLu*) dos neurônios anteriores. Foi demonstrado experimentalmente que a função de retificação *ReLu* (*rectifier linear unit*) tende a obter melhores resultados para a convergência da rede (GONZALEZ; WOODS, 2018a).

As camadas que constituem uma rede neural artificial podem ter quantidades distintas de nós, mas cada neurônio tem apenas uma saída. Quando todos os neurônios de uma camada estão conectados a todos os neurônios da camada seguinte, diz-se que a rede é totalmente conectada. Um requisito importante dessas redes, conhecidas como redes *feedforward*, é a ausência de ciclos entre os nós, ou seja, os sinais das entradas avançam somente no sentido da camada de entrada para a camada de saída. Geralmente, denomina-se rede neural rasa toda rede neural artificial que possui apenas uma camada interna de neurônio. Caso ela possua duas ou mais camadas entre as camadas de entrada e saída, chama-se de rede neural profunda (GONZALEZ; WOODS, 2018a).

O treinamento de uma rede neural artificial para classificação ou regressão envolve conhecimento do conjunto de valores esperados para a camada de saída, dada uma entrada conhecida. Contudo, não se conhece, inicialmente, os valores esperados de saída dos neurônios das camadas ocultas. O treinamento por *backpropagation* das redes é feito em quatro passos: 1) entrada dos valores a serem processados; 2) passo

à frente dos valores do treinamento a serem processados para determinação do erro; 3) passo de *backpropagation* que retorna o erro da saída de volta para as camadas; 4) atualização de pesos e viés com base no erro retornado. O objetivo desse processo é reduzir o erro que a rede apresenta para as entradas de treinamento. Este erro é calculado com base em uma função de custo levando em conta a saída esperada e o resultado apresentado pela rede.

3.3.3 Redes Neurais Convolucionais

As redes neurais artificiais têm como entrada vetores de atributos previamente modelados e extraídos do conjunto de dados para seu treinamento. Uma característica importante das redes é a capacidade de aprender padrões diretamente do conjunto de treino. No domínio das imagens, é possível modelar a entrada de tal forma que os *pixels* constituam o vetor de atributos. Contudo, essa forma de representação não é capaz de aproveitar as relações espaciais entre os *pixels* que possibilitam a identificação de padrões na imagem como bordas, cantos, ou outra característica relevante.

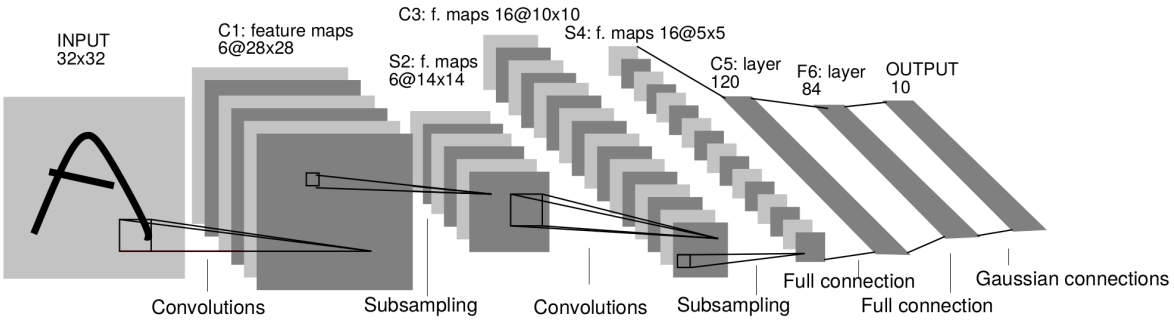
Uma rede neural convolucional (*CNN*) é uma rede na qual um sinal de entrada alimenta um conjunto empilhado de camadas de convolução cuja saída é repassada a um conjunto de camadas de neurônios artificiais densamente (totalmente) conectadas (VENKATESAN; LI, 2018). A convolução é uma operação espacial realizada na imagem que computa uma soma de produtos entre os *pixels* e um *kernel* de pesos (filtro de convolução). A operação é realizada para todo ponto (x, y) da imagem, resultando num valor escalar. Se ao valor for adicionado um viés e o resultado for passado a uma função de ativação, a operação da *CNN* se assemelha ao funcionamento de uma rede neural (GONZALEZ; WOODS, 2018a).

A imagem de entrada de uma *CNN* é inicialmente processada por uma camada de convolução com uma quantidade f_1 de *kernels* que tem como resultado f_1 mapas de atributos. Cada mapa tem dimensionalidade reduzida em uma camada de subamostragem. Cada um dos f_1 mapas resultantes passa por outra camada de convolução com f_2 *kernels*. Esse processo gera f_1 respostas para cada elemento (x, y) dos f_2 mapas subamostrados, que são combinados por superimposição. Assim, a segunda camada de convolução tem como resultado f_2 mapas de atributos. Os processos são repetidos com diferentes *kernels* até que os mapas resultantes são transformados em um vetor unidimensional que serve de entrada para as camadas de neurônios totalmente conectadas. Nesse momento, o processamento ocorre de forma análoga ao de uma rede neural totalmente conectada. Desse modo, as etapas de convolução e subamostragem das *CNN*s representam a extração de atributos do conjunto de dados modelado. As *CNN*s são métodos muito apropriados para problemas de classificação que envolvem imagens, uma vez que elas têm a capacidade de abstrair

as relações espaciais presentes entre seus atributos durante as fases de convolução e subamostragem.

LeCun et al. (1989) usaram a estratégia de *back-propagation* para estimar os coeficientes dos filtros de convolução diretamente de imagens de dígitos escritos à mão. Esse trabalho foi pioneiro na utilização de redes neurais convolucionais para classificação de imagens e fez com que essa abordagem se tornasse base para os métodos de Visão Computacional. A Figura 11 ilustra a arquitetura de uma rede com o mesmo propósito. É possível observar as operações descritas anteriormente em uma imagem de 32x32 *pixels*. Na primeira camada de convolução são utilizados 6 *kernels* 5x5 que resultam em 6 mapas de atributos com dimensão 28x28 cada. Depois a subamostragem com tamanho 2x2 reduz a dimensão dos 6 mapas de atributos para 14x14. Na sequência, são aplicados e combinados os resultados de 16 *kernels* nos 6 mapas, gerando 16 novos mapas de atributos com dimensão 10x10, que são novamente reduzidos pela metade. Ao final, os mapas resultantes são vetorizados e introduzidos em uma camada densa.

Figura 11 – Rede neural convolucional para classificação de dígitos.



Fonte: (LECUN et al., 1998)

Atualmente, as redes neurais convolucionais são fundamentais para os métodos de análise facial, especialmente para o alinhamento facial (HSU et al., 2020).

3.4 Métricas de Avaliação

A comparação de trabalhos relacionados ao problema de alinhamento facial deve ser realizada de forma padronizada para reduzir imprecisões na análise dos resultados. A aferição do erro de predição em um método de localização de pontos faciais é feita, comumente, a partir da distância euclidiana entre os pontos reais e o resultado do modelo preditivo (KHABARLAK; KORIASHKINA, 2022). Além disso, quando os resultados de dois modelos são próximos, é necessária a adoção de testes estatísticos para confirmar as diferenças.

3.4.1 Média do Erro Normalizado

A distância entre o vetor de pontos faciais estimado e a marcação de um especialista para um determinado caso de teste é normalizada de acordo com o número de pontos do padrão de marcação e uma distância entre dois pontos pré-definidos. Esse valor calculado é conhecido como média do erro normalizado (*NME*) (WANG, 2019). A Equação 3.8 mostra como é calculado o *NME* de um caso de teste:

$$NME = \frac{1}{dL} \sum_{i=1}^{L} \|x_i^g - x_i^e\|_2 \times 100, \tag{3.8}$$

onde d é a distância entre os pontos pré-definidos, L é a quantidade de pontos do padrão de marcação, x_i^g são as coordenadas da localização correta do i-ésimo ponto, e x_i^e são as coordenadas da localização estimada.

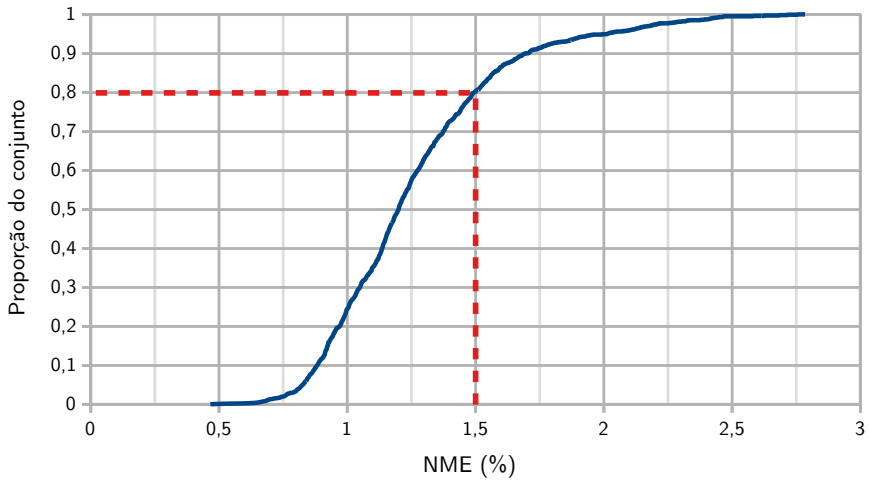
Em trabalhos relacionados, três escolhas geralmente são feitas para cálculo da distância explorada na normalização: distância entre os pontos mais afastados dos olhos; distância entre as pupilas; e tamanho da diagonal do caixa delimitadora da face.

3.4.2 Gráfico de Erro Cumulativo, Área Abaixo da Curva, e Taxa de Falha

O valor geral do *NME* em um conjunto de teste, ou seja, a média de todos os *NME* calculados pode não refletir idealmente a resposta de um modelo de alinhamento facial naquele conjunto. Nos trabalhos relacionados, explora-se com frequência gráficos de erro cumulativo para avaliação dos métodos. O gráfico de erro cumulativo (*CED*, do inglês *cumulative error distribution*) relaciona o *NME* máximo observado até uma determinada proporção do conjunto de teste. Assim, é possível observar com mais clareza as diferenças de desempenho que os modelos apresentam entre os melhores e os piores casos avaliados. A Figura 12 exemplifica um gráfico de erro cumulativo. Nele, a série de dados apresenta todos os *NME*s do conjunto avaliado (curva azul). A linha tracejada em vermelho indica que 80% do conjunto apresentou *NME* inferior ou igual a 1,5%.

Aliado à curva de erro cumulativo, trabalhos relacionados utilizam o cálculo da área abaixo da curva até um determinado limite para avaliação de resultados. A métrica é conhecida como área abaixo da curva (*AUC*) e seu valor calculado é relativizado pelo limite de erro α pré-estabelecido. A taxa de falha é outra métrica bastante utilizada e indica a porcentagem do teste que apresenta erro superior a um limite também pré-estabelecido (BULAT; SANCHEZ; TZIMIROPOULOS, 2021).

Figura 12 – Exemplo de gráfico de erro cumulativo.



Fonte: Elaborado pelo autor.

3.4.3 Avaliação Estatística

Quando os resultados de desempenhos de diferentes modelos de alinhamento facial são muito próximos, faz-se necessária, visando a uma comparação mais precisa, a utilização de testes estatísticos para determinar se as diferenças entre eles são de fato significativas ou se são apenas frutos de variações casuais. O mesmo pode ocorrer quando um único modelo é aplicado em diferentes grupos dentro de um contexto específico e apresenta resultados distintos para cara um deles. Nesse caso, deve-se conduzir a avaliação sob a ótica de uma métrica de distribuição.

Nesta pesquisa, é necessária a avaliação de desempenho dos métodos quando aplicados a 3 grupos distintos (*HC*, *ALS*, e *PS*). Nesse caso, tendo em vista a distribuição de resultados comumente observada em modelos de alinhamento facial, o teste de Kruskal-Wallis (KRUSKAL; WALLIS, 1952) se mostra uma alternativa adequada para determinar se há diferenças significativas de desempenho. Caso a hipótese nula seja rejeitada (i.e. existe diferença significativa entre pelo menos dois grupos) deve-se aplicar o teste de Dunn (DUNN, 1961) visando confirmar quais grupos apresentam as diferenças. Esses testes permitem, adicionalmente, uma interpretação sobre o nível das diferenças de desempenho. Por exemplo, dado um limiar de significância de $\alpha = 0,05$, para valor $p \geq 0,05$ calculado, a hipótese nula não pode ser rejeitada. Caso contrário, é possível inferir que quanto menor o valor de p, mais diferentes são as distribuições avaliadas.

Além dos testes de hipótese, que indicam se as amostras são originárias da mesma distribuição, é possível expressar, de maneira mais inteligível, as diferenças entre os desempenhos obtidos nos grupos avaliados quando é caracterizado o viés

no modelo. O método N-Sigma produz uma distância N entre duas distribuições (DEALCALA et al., 2023). A Equação 3.9 mostra como N é expresso:

$$N = \frac{\mu_{G1} - \mu_{G2}}{\sigma_{G1}}, \tag{3.9}$$

em que μ_{G1} e μ_{G2} são as médias das duas populações em comparação, e σ_{G1} é o desvio padrão da população usada como referência.

A análise de dados resultantes dos experimentos pode sugerir algum nível de associação linear entre as variáveis observadas. Neste trabalho foi utilizado o coeficiente de correlação amostral de Pearson para medir o nível de correlação entre duas variáveis. A Equação 3.10 mostra como é calculado o coeficiente:

$$r_P = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}, \tag{3.10}$$

onde x_i e y_i são as observações das variáveis analisadas, $\bar{x}$ e $\bar{y}$ suas respectivas médias. O valor do coeficiente pode variar no intervalo de -1 (indicando correlação negativa perfeita entre as variáveis) a +1 (indicando correlação positiva), com valor igual a 0 indicando que não há dependência linear (MORETTIN; SINGER, 2022).

3.5 Considerações Finais

Neste capítulo foram descritos os principais elementos utilizados no desenvolvimento da pesquisa, desde a formalização do modelo facial até os métodos de avaliação dos resultados. No próximo capítulo são descritos o método de alinhamento facial e os *datasets* usados para treino e teste dos modelos.

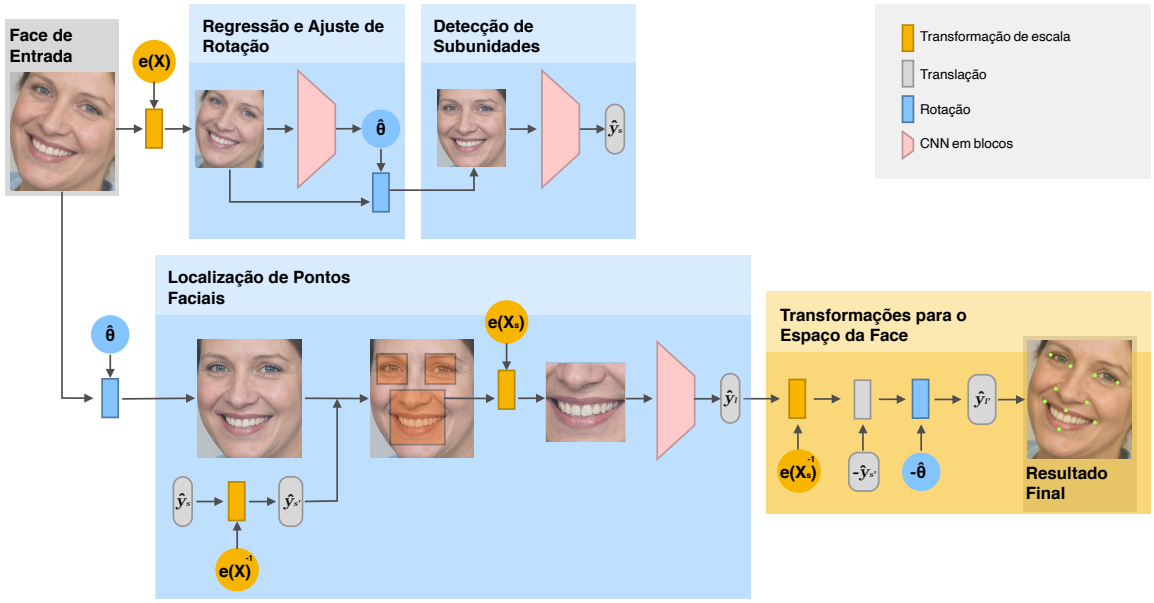
4 Metodologia e Materiais

O problema observado nos métodos mais atuais de alinhamento facial está relacionado às diferenças de desempenho que eles apresentam quando aplicados em subgrupos distintos com base em algum atributo demográfico. Estudos demonstraram que os modelos do estado da arte não foram capazes de apresentar desempenhos igualitários de localização de pontos faciais entre populações clínicas. Bandini et al. (2021) caracterizaram o viés da *FAN-2D* quando aplicada no alinhamento facial de indivíduos com problemas neurológicos (portadores de ELA e sobreviventes de derrame) comparados a adultos saudáveis (grupo de controle). A proposta de ajuste fino na *FAN-2D* com amostras do *Toronto Neuroface* no sentido de conferir maior representatividade ao modelo foi incapaz de reduzir as disparidades de desempenho. A maior dificuldade do modelo observada por Bandini et al. (2021) foi a presença de assimetria orofacial mais acentuada em indivíduos dos grupos *ALS* e *HC*, que é potencializada pelas expressões faciais exploradas em exames clínicos.

Nesse sentido, este trabalho busca dividir a análise criando modelos de localização de pontos faciais específicos para cada subunidade facial que são agregados em uma etapa de detecção de subunidades. A ideia é demonstrar que a detecção dos pontos dentro de cada subunidade é mais robusta às variações de expressão e à assimetria orofacial, o que pode contribuir para a redução das diferenças de desempenho da abordagem quando aplicada aos grupos do *Toronto Neuroface*. A Figura 13 ilustra o fluxo de etapas da metodologia. Pode ser observado na figura que é incorporado o conhecimento prévio da face detectada. A primeira etapa é uma uniformização de escala da entrada com fator $e(\mathbb{X})$. Depois, o regressor de rotação estima o ângulo $\hat{\theta}$ e a face em escala uniformizada tem rotação ajustada. A partir da face ajustada, o detector de subunidades estima os centros $\hat{y}_s$ de cada elemento facial. Para localização dos pontos faciais da subnidade s, é extraída uma amostra da face de entrada rotacionada em $\hat{\theta}$ graus com o centro estimado de s após aplicação de transformação de escala $e(\mathbb{X})^{-1}$. A amostra serve de entrada ao localizador de pontos que estima o vetor de coordenadas $\hat{y}_l$ com os pontos do elemento facial. Finalmente, o vetor $\hat{y}_l$ é transformado para o espaço da face de entrada. Essas etapas são detalhadas neste capítulo.

Outro aspecto importante abordado neste capítulo são os *datasets* utilizados na construção dos modelos oriundos dessa metodologia e na avaliação dos resultados. Para treinamento foram utilizados os subconjuntos de treino dos *datasets Helen* (LE et al., 2012), *AFW* (ZHU; RAMANAN, 2012), *LFPW* (BELHUMEUR et al., 2013), e o subconjunto de imagens faciais *indoor* do *benchmark 300W* (SAGONAS et al., 2013). Neste trabalho foi utilizada a anotação de pontos seguindo o padrão *Multi-PIE*

Figura 13 – Fluxograma da metodologia.



Fonte: Elaborado pelo autor.

fornecida para o desafio *300 Faces In-the-Wild* (SAGONAS et al., 2013). O *dataset* usado para avaliação dos modelos foi o *Toronto Neuroface* com dados balanceados de diferentes grupos clínicos (BANDINI et al., 2021). A partir deste capítulo as seguintes convenções são adotadas: $\mathbb{S} = \{(x^{(i)}, y^{(i)})\}_{i=1}^{N}$ refere-se a um conjunto de treinamento[1] com N exemplares; no qual $x^{(i)} \in \mathbb{X}$ representa a i-ésima entrada de treino, e $y^{(i)} \in \mathbb{Y}$ representa a saída esperada para $x^{(i)}$; notações sem serifa como x, y, l, w, e h representam coordenadas ou atributos no domínio da imagem.

4.1 Preparação das Imagens Faciais

Tendo em vista que a detecção da face é um conhecimento previamente fornecido (coordenadas da imagem onde se encontra a face, chamada de caixa delimitadora), é possível realizar uma preparação das imagens a partir da exclusão de atributos desnecessários, isto é, que não pertencem à face detectada. Esse processo – que deve ser realizado para todos os *datasets* – traz duas vantagens à pesquisa: simplificação da metodologia, pois filtra detalhes desimportantes para a descrição do modelo; e economia de recursos computacionais, o que contribui para a realização de experimentos.

Para utilização das imagens faciais como entrada dos modelos baseados em *CNNs* da metodologia, é necessário que a face detectada tenha lados de mesmo

[1] O conjunto de treinamento depende do modelo em questão, podendo ser: 1) faces de entrada e os respectivos ângulos de rotação, para o modelo de rotação; 2) faces com ajuste de rotação e o vetor de coordenadas das subunidades, para o modelo de detecção de subunidades; 3) amostras das subunidades e o vetor de coordenadas dos pontos, para os modelos de localização de pontos.

tamanho. Contudo, o aspecto altura×largura da caixa delimitadora pode variar em razão da pose, rotação, ou da própria morfologia facial do indivíduo. Portanto, caso a caixa tenha lados de tamanhos distintos, deve-se realizar uma expansão da mesma ao longo do eixo do menor lado. Inicialmente, determina-se qual deve ser o tamanho do lado da caixa expandida. A Equação 4.1 mostra como calcular o lado da nova caixa:

$$\mathsf{l}'_{bbox} = \max(\mathsf{w}^{(i)}_{bbox}, \mathsf{h}^{(i)}_{bbox}), \tag{4.1}$$

onde $\mathsf{w}^{(i)}_{bbox}$ e $\mathsf{h}^{(i)}_{bbox}$ são largura e altura, respectivamente, da caixa delimitadora da face $x^{(i)}$. A nova caixa delimitadora da i-ésima face do conjunto em questão tem origem calculada da seguinte forma:

$$(\mathsf{x_o}', \mathsf{y_o}')^{(i)} = \begin{cases} (\mathsf{x_o}, \mathsf{y_o} - \delta_{bbox})^{(i)}, & \text{se } \mathsf{w}^{(i)}_{bbox} > \mathsf{h}^{(i)}_{bbox} \\ (\mathsf{x_o} - \delta_{bbox}, \mathsf{y_o})^{(i)}, & \text{se } \mathsf{w}^{(i)}_{bbox} \leq \mathsf{h}^{(i)}_{bbox} \end{cases} \tag{4.2}$$

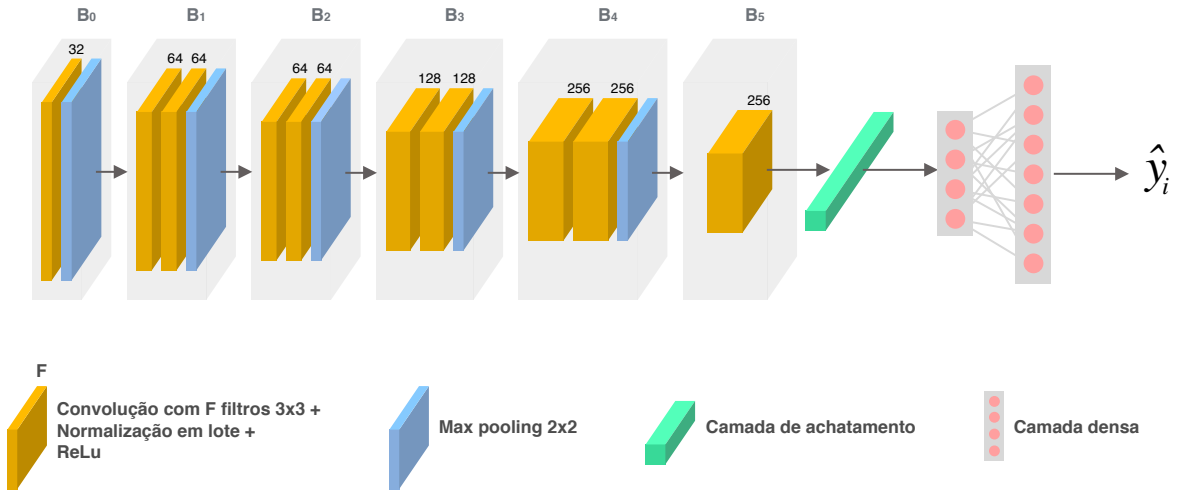
onde o par $(\mathsf{x_o}, \mathsf{y_o})$ indica a origem da caixa no espaço da imagem, e $\delta_{bbox} = |(\mathsf{w}^{(i)}_{bbox} - \mathsf{h}^{(i)}_{bbox}|/2$ é o deslocamento aplicado à origem ao longo do eixo do menor lado da caixa. Nesse processo não é realizado nenhum *padding* na imagem resultante. O mesmo deslocamento deve ser aplicado às anotações $y^{(i)}$ da imagem facial correspondente.

4.2 Arquitetura de *CNN* em Blocos

As *CNN* utilizadas nas regressões (rotação, detecção de unidades, e localização de pontos faciais) são construídas com base em uma arquitetura adaptável em função da escala, com blocos de convolução e subamostragem. Para cada tarefa, as estruturas modeladas diferem na quantidade de blocos e também no domínio da saída. A Figura 14 esquematiza a arquitetura da *CNN* adaptada à escala máxima, ou seja, o modelo mais complexo atingido nesta pesquisa. Esse modelo foi utilizado para as etapas de ajuste de rotação e detecção de subunidades. Após a detecção facial, a escala das entradas (para localização dos pontos de cada elemento facial) tende a diminuir, em comparação com a face. Nesses casos, a estrutura é adaptada com redução de blocos para comportar escalas menores.

O primeiro bloco tem uma camada de convolução com normalização em lote ativada por uma função para anulação de valores negativos (*ReLu*) e uma de subamostragem que escolhe o valor mais alto dentro do filtro (*max pooling* 2×2). Os blocos intermediários são compostos por duas camadas de convolução e uma de subamostragem. O último bloco contém apenas uma camada de convolução. Por fim, o vetor achatado resultante das convoluções e subamostragens alimenta as camadas densas que têm como saída uma estimativa $\hat{y}$ que deve ser: 1) um ângulo para uniformização da rotação; 2) um vetor de coordenadas dos centros dos elementos faciais detectados; 3) um vetor de coordenadas dos pontos de um elemento facial.

Figura 14 – Arquitetura da *CNN* para regressão de rotação facial e vetores de coordenadas.



Fonte: Elaborado pelo autor.

Uma vez que a regressão para o cálculo de rotação facial, assim como a detecção de subunidades, leva em consideração toda a textura da face detectada, é importante que a estrutura da rede seja capaz de modelar atributos de alto nível tal qual os componentes faciais agregados. Assim, a estrutura de *CNN* adotada nessas tarefas é análoga à *Deep CNN F1* usada para regressão de coordenadas em primeiro nível apresentada no trabalho de Sun, Wang e Tang (2013). A principal diferença entre a estrutura usada neste trabalho e a *Deep CNN F1* reside no número de camadas de convolução e subamostragem. Neste trabalho, a estrutura da rede é definida com 15 camadas, enquanto a rede de Sun, Wang e Tang (2013) é limitada a 7 camadas por conta da escala reduzida da entrada. Como apontado pelos autores, o acréscimo de camadas contribuiu para a melhora no desempenho do regressor. A estrutura ilustrada na Figura 14 com as camadas organizadas em 6 blocos saturou os recursos computacionais disponíveis para o treinamento, o que apontou um limite para a complexidade dos modelos experimentados nesta pesquisa.

A escala dos elementos faciais observados em treino pode não permitir a utilização de todos os blocos dessa arquitetura. Essa restrição ocorre porque, assim como na *Deep CNN F1*, não é realizado nenhum tipo de *padding* durante o processamento. Logo, as respostas de cada convolução (além das subamostragens) têm dimensão sempre menor que as entradas, assim, à medida que a imagem é processada pela rede seus atributos diminuem até atingirem o limite da representação. Os regressores de rotação e detecção de subunidades fazem uso da estrutura completa, contudo, caso o conjunto de treino seja composto por entradas em escala mais reduzida, é necessário adaptar a estrutura. Esse aspecto fica evidente para as subunidades faciais que têm menos atributos que as faces de entrada.

O agrupamento das camadas em blocos (B_0 a B_5) facilita a definição da arquitetura mais adequada à escala envolvida. A definição da estrutura para cada modelo de localização de pontos é realizada após ajuste de escala da entrada. A Tabela 3 resume como as estruturas podem ser montadas a partir de um valor mínimo de tamanho de entrada.

Tabela 3 – Definição das estruturas para treinamento dos modelos

Estrutura	**Blocos**	**Entrada mínima (px)**	**Intervalo da escala (px)**
E_0	$B_0 \rightarrow B_1 \rightarrow B_2 \rightarrow B_3 \rightarrow B_4 \rightarrow B_5$	218×218	$\mu_{\text{l}}(\mathbb{X}) \geq 218$
E_1	$B_0 \rightarrow B_1 \rightarrow B_2 \rightarrow B_3 \rightarrow B_4$	154×154	$218 > \mu_{\text{l}}(\mathbb{X}) \geq 154$
E_2	$B_0 \rightarrow B_1 \rightarrow B_2 \rightarrow B_3$	74×74	$154 > \mu_{\text{l}}(\mathbb{X}) \geq 74$
E_3	$B_0 \rightarrow B_1 \rightarrow B_2$	34×34	$74 > \mu_{\text{l}}(\mathbb{X}) \geq 34$
E_4	$B_0 \rightarrow B_1$	14×14	$34 > \mu_{\text{l}}(\mathbb{X}) \geq 14$
E_5	B_0	4×4	$14 > \mu_{\text{l}}(\mathbb{X}) \geq 4$

4.3 Uniformização de Escala

A etapa de uniformização de escala das entradas das *CNN* tem dois aspectos importantes neste trabalho. O primeiro diz respeito à grande diferença de escala das imagens observadas no conjunto de treino. O *dataset Helen*, por exemplo, é um conjunto bastante heterogêneo em relação a essa propriedade. Nele, o tamanho das faces – determinado pela caixa delimitadora – varia de 250x250 a 1500x1500 *pixels*. O outro aspecto relativo à escala diz respeito às imagens de teste. No caso do *Toronto Neuroface* não há grandes diferenças de escala entre as faces, contudo, é necessário que a face de teste do modelo seja ajustada para a escala das faces observadas em treinamento. Além disso, os elementos faciais, representados nesta metodologia por subunidades, herdam as variações de escala observadas nas faces e também passam por esse processo de uniformização.

Para tanto, foi definido um fator que deve ser calculado a partir do tamanho da entrada no conjunto de treino, e transferido para aplicação no conjunto de teste, uma vez que a caixa é admitida como conhecimento prévio. Para as subunidades faciais, o processo é análogo, contudo, no lugar do conhecimento prévio da caixa é utilizada a estimativa do detector de subunidades.

O fator de escala da entrada é um parâmetro de ajuste dependente do elemento de entrada da rede e definido no domínio do modelo com auxílio das caixas que delimitam as faces (conhecimento prévio) ou os elementos faciais (estimados pelo detector). Primeiramente é calculada a média das maiores dimensões das caixas. A Equação 4.3 descreve como calcular essa média para as entradas de um conjunto de treinamento $\mathbb{X}$:

$$\mu_{\mathsf{l}}(\mathbb{X}) = \frac{1}{N}\sum_{i=1}^{N} \mathsf{l}^{(i)}, \tag{4.3}$$

onde N é o número de observações em $\mathbb{X}$, $\mathsf{l}^{(i)}$ é o lado da caixa delimitadora da face – ou da subunidade – $x^{(i)}$ do conjunto, considerando que as imagens já passaram por fase de preparação. A depender do conjunto de imagens utilizado, a média calculada pode representar imagens com escala muito superior à dimensão mínima da estrutura E_0 (arquitetura mais complexa do trabalho). Essa média deve ser usada para determinar a estrutura de blocos mais adequada à escala em questão, seja para a face de entrada, seja para as subunidades detectadas. A Tabela 3 descreve as estruturas possíveis, tamanho mínimo de entrada da rede, e uma regra parametrizada pela média calculada que determina a estrutura mais adequada.

O fator de escala da entrada no contexto desses modelos é calculado de acordo com a seguinte equação:

$$e^{(i)}(\mathbb{X}) = \frac{\min(\mu_{\mathsf{l}}(\mathbb{X}), 218)}{\mathsf{l}^{(i)}}. \tag{4.4}$$

A normalização de escala pelo fator $e^{(i)}(\mathbb{X})$ estabelece o padrão de entrada do modelo para qualquer x independente do conjunto ao qual pertence (treino ou teste). De maneira análoga, as anotações y das saídas referentes a x são transformadas com base no mesmo fator. Seja $A_e^{(i)} = T_e(e^{(i)}(\mathbb{X}))$ a matriz de transformação de escala isotrópica com fator $e^{(i)}(\mathbb{X})$ calculado para uniformizar $x^{(i)}$ e $y^{(i)}$, a Equação 4.5 descreve o conjunto de entrada do treino ($\mathbb{S}$) com escala uniformizada $\mathbb{S}_e$:

$$\mathbb{S}_e = \{(A_e^{(i)}x^{(i)}, A_e^{(i)}y^{(i)})\}_{i=1}^{N}. \tag{4.5}$$

4.4 Regressão e Ajuste de Rotação

Muito embora as imagens faciais alvo desta pesquisa (*Toronto Neuroface*) tenham sido capturadas em ambiente controlado, algumas variações de pose facial persistiram durante a construção do *dataset*. Para corrigir as variações de pose em torno do eixo longitudinal, é aplicada uma correção de rotação. Essa é a etapa de ajuste da rotação que, a partir de uma face com escala uniformizada, estima a variação de rotação do eixo determinado pelos centros dos olhos em relação ao eixo horizontal, e aplica uma rotação inversa tanto na imagem quanto nos pontos faciais.

A rotação de uma face x qualquer é estimada por um modelo de *CNN* treinado com os exemplares de $\mathbb{X}_e$ como entrada para regressão do conjunto $\mathbb{Y}_\theta = \{\theta^{(i)}\}_{i=1}^{N}$. O ângulo $\theta^{(i)}$ correspondente à rotação horizontal da i-ésima face de treino é calculado de acordo com a Equação 4.6:

$$\theta^{(i)} = \arctan\frac{\mathsf{y}_l^{(i)} - \mathsf{y}_r^{(i)}}{\mathsf{x}_l^{(i)} - \mathsf{x}_r^{(i)}}, \tag{4.6}$$

onde $(\mathrm{x}_k^{(i)}, \mathrm{y}_k^{(i)})$ são as coordenadas do k-ésimo ponto da i-ésima face de entrada de $\mathbb{X}_e$, e l e r são os índices dos pontos que representam os cantos dos olhos esquerdo e direito, respectivamente.

Após treinamento do regressor de rotação com entrada $\mathbb{X}_e$ e saída $\mathbb{Y}_\theta$, deve-se utilizar o modelo resultante para auto-ajuste do conjunto de treino $\mathbb{S}_e$ com as rotações estimadas. Seja $A_\theta^{(i)} = T_\theta(-\hat{\theta}^{(i)})$ a matriz de rotação com o ângulo $\hat{\theta}^{(i)}$ estimado pelo regressor de rotação para a i-ésima face de $\mathbb{S}_e$, a Equação 4.7 descreve o conjunto de entrada do treino $\mathbb{S}_e$ com rotação ajustada $\mathbb{S}_{\hat{\theta}}$:

$$\mathbb{S}_{\hat{\theta}} = \{(A_\theta^{(i)} x_e^{(i)}, A_\theta^{(i)} y_e^{(i)})\}_{i=1}^{N}. \tag{4.7}$$

O conjunto de entrada $\mathbb{X}_e$ após auto-ajuste da rotação é transformado em $\mathbb{X}_{\hat{\theta}}$, que integra a detecção de subunidades. A mesma transformação de rotação é realizada na saída $\mathbb{Y}_e$, resultando no conjunto $\mathbb{Y}_{\hat{\theta}}$.

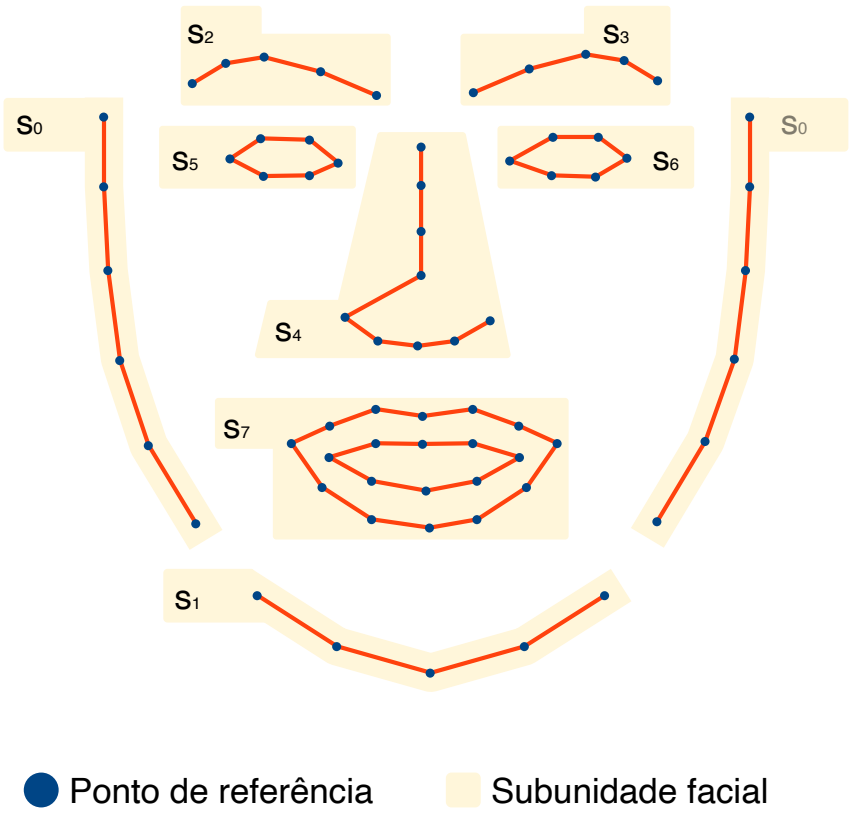
4.5 Detecção de Subunidades Faciais

Após o processo de uniformização facial, é realizada a etapa de detecção de subunidades faciais. Como explicado no início do capítulo, a divisão da análise facial em subunidades busca aumentar a variabilidade do modelo visando conferir mais robustez às variações de expressão facial. A modelagem da face em subunidades foi realizada com base na divisão semântica de acordo com os componentes faciais tais como os olhos, as sobrancelhas, o nariz, etc. A ideia que fundamenta essa escolha é que a textura do elemento facial é suficiente para a localização dos pontos que pertencem àquela estrutura. Existe ainda a possibilidade de se dividir um componente em duas ou mais subunidades, ou de concatenar dois ou mais componentes.

Foram idealizadas algumas configurações de divisão da face para avaliar diferentes modelos de detecção de subunidades. As primeiras tentativas consideravam modelagem semelhante à apresentada na Subseção 3.1.2, com a diferença que os pontos da mandíbula e do queixo estavam agrupados como uma única subunidade. O problema dessa estratégia é que ela limita a variabilidade da forma facial na região da boca. Assim, o modelo apresenta maior dificuldade na localização dos pontos da boca em faces que apresentam expressões nas quais há movimentação vertical dos lábios. Portanto, buscou-se dividir esta subunidade em três: parte esquerda da mandíbula; parte direita da mandíbula; e queixo. Essa configuração permite maior variação para o modelo, contudo, os *patches* das partes laterais da mandíbula acabam incluindo atributos de fundo da imagem, desnecessários para localização dos pontos e ignorados pela detecção facial. A estratégia mais bem sucedida considerando não somente a detecção das subunidades, mas também o resultado geral do método, levou em consideração o agrupamento das duas partes laterais da mandíbula em uma única subunidade. A

Figura 15 descreve a configuração de melhor desempenho observada nesta pesquisa.

Figura 15 – Divisão dos pontos de referência em subunidades.

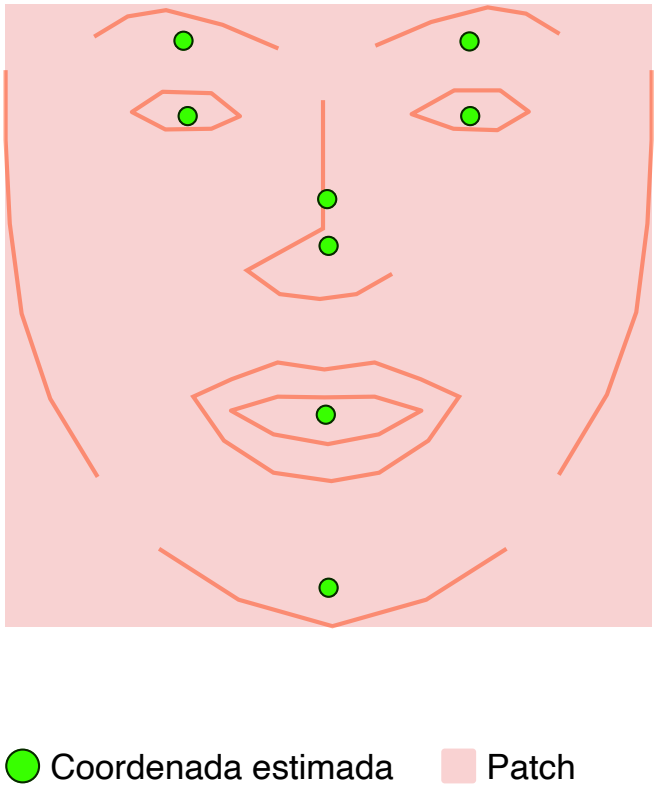


Fonte: Elaborado pelo autor.

A caixa delimitadora de cada subunidade é definida com base nos pontos extremos da mesma, e indica a amostra da imagem facial usada na localização dos pontos em etapa posterior. A utilização somente do conteúdo determinado pela caixa para realizar a localização dos pontos daquele componente é uma das ideias que fundamentam a escolha da estratégia principal deste trabalho, que é a divisão da modelagem facial. Essa é uma noção intermediária entre os métodos que consideram apenas textura global, e aqueles que fazem uso de textura local para localização dos pontos faciais. Por exemplo, a localização dos pontos que compõem o nariz depende apenas da textura do próprio nariz. Nesse caso, um método de alinhamento facial baseado em textura global utiliza mais informação do que de fato é necessário para determinação individual desses pontos. Já um método baseado em textura local pode enfrentar dificuldade em distinguir padrões de textura semelhantes em diferentes componentes faciais. A busca por uma solução intermediária, nesse sentido, evita o uso de informações desnecessárias para a localização ao mesmo tempo que amplifica a variabilidade do modelo geral ao combinar diferentes modelos de localização de

pontos nas subunidades. A detecção das subunidades, contudo, é feita com base na textura global, assim como o ajuste de rotação. A Figura 16 esquematiza a textura facial usada na detecção de subunidades, e os centros das caixas delimitadoras de cada subunidade.

Figura 16 – Esquema de textura global e centros de subunidades detectadas.



Fonte: Elaborado pelo autor.

Um vez definida a configuração de subunidades da face, deve-se treinar a estrutura que realiza a detecção. A primeira abordagem escolhida para modelagem do vetor de saída do processo de detecção de subunidades faciais foi baseada nos centros de massa. Entretanto, dependendo da divisão facial, pode ocorrer um desequilíbrio na distribuição dos pontos. No caso do nariz, a distribuição de pontos é mais densa na base do que no dorso nasal, que resulta em um centro de massa mais próximo da base do que da raiz nasal. A utilização dos centros das caixas delimitadoras elimina os problemas relativos a distribuições desequilibradas. O cálculo dos centros das subunidades para construção do vetor de saída da *CNN* de detecção de subunidades é realizado a partir do espaço transformado pelo ajuste de rotação.

Seja $s_j^{(i)}$ a j-ésima subunidade da configuração definida na entrada $x_{\hat{\theta}}^{(i)}$, e o par $(\mathrm{x}_{k,\hat{\theta}}^{(i)}, \mathrm{y}_{k,\hat{\theta}}^{(i)})$ o k-ésimo ponto de $y_{\hat{\theta}}^{(i)} \in \mathbb{Y}_{\hat{\theta}}$. A coordenada horizontal $\mathrm{cx}_j^{(i)}$ do centro

de $s_j^{(i)}$ em $x_{\hat{\theta}}^{(i)}$ é igual à média entre as coordenadas horizontais máxima e mínima dos pontos que compõem $s_j^{(i)}$. Para a coordenada vertical, a definição é análoga. O vetor de centros de todas as subunidades $y_s^{(i)} = [(\mathrm{cx}_0^{(i)}, \mathrm{cy}_0^{(i)}), (\mathrm{cx}_1^{(i)}, \mathrm{cy}_1^{(i)}), \ldots, (\mathrm{cx}_m^{(i)}, \mathrm{cy}_m^{(i)})]$ compõe a saída esperada da *CNN* para detecção das m subunidades da configuração. Esse conjunto, calculado no espaço rotacionado, é denominado $\mathbb{Y}_s = \{y_s^{(i)}\}_{i=1}^{N}$

Após treinamento do regressor de subunidades com a entrada $\mathbb{X}_{\hat{\theta}}$ e saída $\mathbb{Y}_s$, utiliza-se o modelo resultante para detecção das subunidades no conjunto de entrada $\mathbb{X}_{\hat{\theta}}$. O conjunto de saídas estimadas $\mathbb{Y}_{\hat{s}}$ é propagado para a etapa posterior de localização de pontos faciais. Nesta pesquisa, a configuração de melhor desempenho tem $m = 8$ subunidades.

4.6 Localização de Pontos Faciais

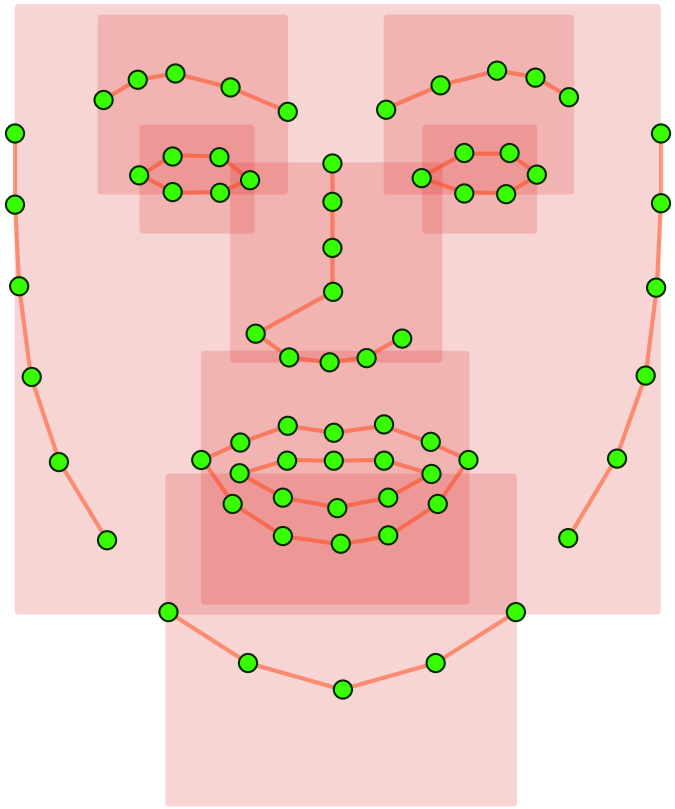
O último processo do método consiste na localização dos pontos de referência de cada subunidade facial e nas transformações geométricas inversas que devem ser aplicadas aos pontos localizados. A localização dos pontos de referência é feita a partir de m *CNN*s, que possuem, cada uma, estrutura própria em bloco (Seção 4.2) definida com base no processo de uniformização de escala, tal como descrito na Seção 4.3.

A construção dos regressores de localização de pontos faciais é conduzida com auxílio de 3 elementos: estimativa gerada pela localização dos centros das subunidades ($\mathbb{Y}_{\hat{s}}$); do conjunto de treinamento inicial ($\mathbb{S}$); e do conjunto de rotações estimadas no treino ($\mathbb{Y}_{\hat{\theta}}$). A ideia para o treinamento desses regressores é aproveitar os atributos em escala original. Para extração desses atributos, deve-se: transformar $\mathbb{S}$ com base nos ângulos estimados $\mathbb{Y}_{\hat{\theta}}$; transformar $\mathbb{Y}_{\hat{s}}$ para a escala de $\mathbb{S}$ através da aplicação da inversa de $T_e(e(\mathbb{X}))$. Assim, as estimativas de detecção de subunidades e a face original rotacionada passam a fazer parte do mesmo espaço, facilitando a extração das amostras que devem compor as entradas das *CNN* de localização de pontos.

A extração da amostra leva em consideração o centro estimado e o tamanho da caixa que delimita a subunidade. Para isso, inicialmente calcula-se a caixa delimitadora dos pontos que compõem a subunidade facial e aplica-se o método de extensão da caixa como explicado na Equação 4.1. A Figura 17 ilustra como as caixas delimitadoras ficam expandidas em forma de quadrado após o processo. A caixa expandida da subunidade $s_j^{(i)}$ tem lado $\mathrm{l}_j^{(i)}$. Então, extrai-se a amostra da subunidade $s_j^{(i)}$ em escala original, com lado $\mathrm{l}_j^{(i)}$, da face original $x^{(i)}$ após o ajuste de rotação. Esse processo visa aproveitar os atributos da imagem original ao máximo.

Por fim, as amostras extraídas passam pelo mesmo processo de uniformização de escala realizada para o conjunto $\mathbb{S}$, como descrito nas Equações 4.3, 4.4, e 4.5. A escala média calculada (Equação 4.3) para as amostras das subunidades incorpora

Figura 17 – Esquema de amostragem das subunidades faciais para regressão dos pontos de referência.



Fonte: Elaborado pelo autor.

o modelo na forma de parâmetro a ser transferido para a aplicação em teste. Além disso, ela determina qual estrutura de *CNN* deve ser adotada para cada regressor de localização de pontos. O conjunto de amostras extraídas e com escala uniformizada compõe a entrada da *CNN* que modela o regressor de pontos da subunidade correspondente. O vetor de coordenadas dos pontos da subunidade passa pelas mesmas transformações e compõe a saída da rede para treino. É importante notar que esse vetor além de estar rotacionado e em diferente escala da face original, por conta da uniformização de escala da amostra, está deslocado do espaço da imagem em virtude da localização do centro da subunidade. Por último, a estimativa de localização dos pontos da subunidade é levada de volta ao espaço da imagem com base nas inversas dessas transformações que foram introduzidas ao longo do processo.

4.7 Definição dos Modelos

As estruturas de *CNN* foram definidas com base nas regras expostas na Tabela 3. A configuração de subunidades adotada nesta pesquisa que atingiu os melhores

resultados está ilustrada na Figura 15. Assim, 10 modelos foram treinados para a concretização da metodologia. A Tabela 4 descreve as estruturas adotadas e as tarefas de cada modelo. É possível notar que os modelos de localização de pontos de subunidades que representam elementos faciais de menor escala têm estrutura de blocos mais simples.

Apesar da utilização de diferentes estruturas para os modelos, o conjunto de hiperparâmetros é o mesmo para todas as redes. A função de custo empregada foi a média do erro ao quadrado (*MSE*, do inglês *mean squared error*). Por questões de simplicidade na implementação, os vetores de coordenadas esperadas para cada modelo foram achatados.

Os modelos foram treinados com um máximo de 300 épocas utilizando uma estratégia de *early stopping* limitada a 30 épocas posteriores a partir da época de erro de validação mínimo. Ao longo do treino, é selecionado o modelo responsável pelo melhor desempenho no conjunto de validação. A Figura 18 ilustra as curvas de aprendizagem dos modelos que atingiram os melhores resultados. Assim como no trabalho de Sun, Wang e Tang (2013), foi experimentada uma estratégia de *drop out* na rede em 2 momentos: antes da camada de achatamento a uma probabilidade de 0,25; e logo após a primeira camada densa, com probabilidade de 0,5. Contudo, foram observadas: uma queda no desempenho do modelo geral; e maior dificuldade de convergência em treino. Em relação aos conjuntos de treino e validação, a proporção é de 3:1. A Tabela 5 descreve os quantitativos de faces de cada conjunto utilizado no treinamento.

Tabela 4 – Definição dos modelos

Modelo	**Tarefa**	**Estrutura de CNN**	**Blocos**
M$_\theta$	Regressão de rotação	E_0	$B_0 \rightarrow B_1 \rightarrow B_2 \rightarrow B_3 \rightarrow B_4 \rightarrow B_5$
Ms	Detecção de subunidades	E_0	$B_0 \rightarrow B_1 \rightarrow B_2 \rightarrow B_3 \rightarrow B_4 \rightarrow B_5$
M$_{S_0}$	Localização de pontos (S_0)	E_0	$B_0 \rightarrow B_1 \rightarrow B_2 \rightarrow B_3 \rightarrow B_4 \rightarrow B_5$
M$_{S_1}$	Localização de pontos (S_1)	E_1	$B_0 \rightarrow B_1 \rightarrow B_2 \rightarrow B_3 \rightarrow B_4$
M$_{S_2}$	Localização de pontos (S_2)	E_2	$B_0 \rightarrow B_1 \rightarrow B_2 \rightarrow B_3$
M$_{S_3}$	Localização de pontos (S_3)	E_2	$B_0 \rightarrow B_1 \rightarrow B_2 \rightarrow B_3$
M$_{S_4}$	Localização de pontos (S_4)	E_1	$B_0 \rightarrow B_1 \rightarrow B_2 \rightarrow B_3 \rightarrow B_4$
M$_{S_5}$	Localização de pontos (S_5)	E_3	$B_0 \rightarrow B_1 \rightarrow B_2$
M$_{S_6}$	Localização de pontos (S_6)	E_3	$B_0 \rightarrow B_1 \rightarrow B_2$
M$_{S_7}$	Localização de pontos (S_7)	E_1	$B_0 \rightarrow B_1 \rightarrow B_2 \rightarrow B_3 \rightarrow B_4$

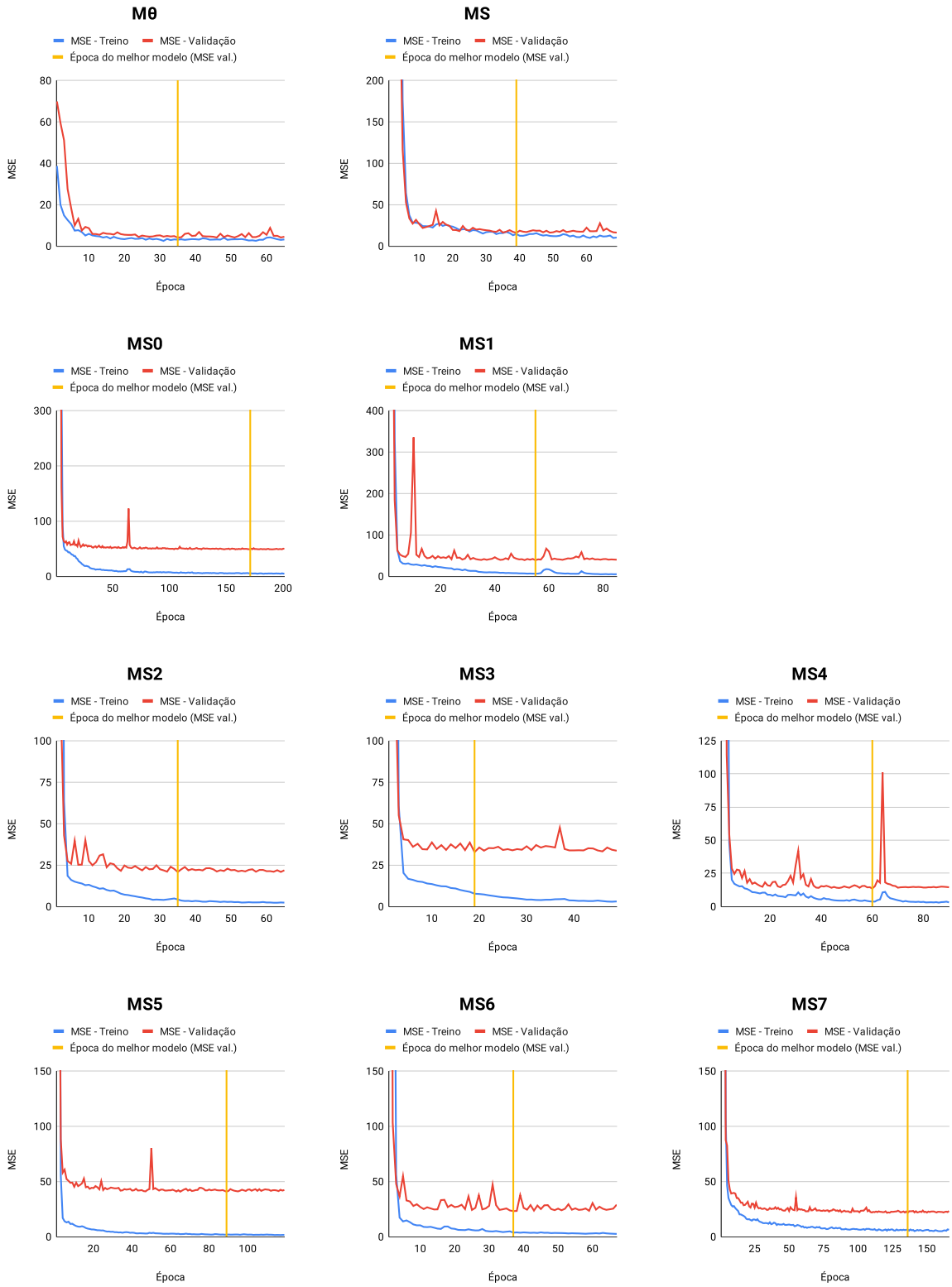


Figura 18 – Curvas de aprendizagem de cada modelo. O estágio final do modelo é determinado pela época cujo erro no conjunto de validação é o mínimo.

4.8 *Toronto Neuroface*

Este *dataset* foi construído com o objetivo de investigar o viés algorítmico dos métodos de alinhamento facial em populações com problemas neurológicos. Ele contou

Tabela 5 – Quantitativo de imagens faciais do conjunto de treinamento

Dataset	**Quantidade**
Helen	1853
LFPW	757
AFW	333
300W (*Indoor*)	300
Total	**3243**

com 3 grupos de participantes: portadores de ELA; sobreviventes de derrame; e um grupos de controle de adultos saudáveis. As imagens do *dataset* foram capturadas em ambiente controlado onde os participantes realizaram expressões faciais que são comumente utilizadas em exames clínicos para aferição do grau de severidade dos problemas neurológicos. As expressões faciais foram codificadas nos arquivos das imagens e estão descritas na Tabela 6. A Figura 19 ilustra algumas amostras de imagens faciais do *Toronto Neuroface*: nas colunas, as imagens estão separadas por expressão facial; nas linhas, estão separadas por grupo clínico do indivíduo.

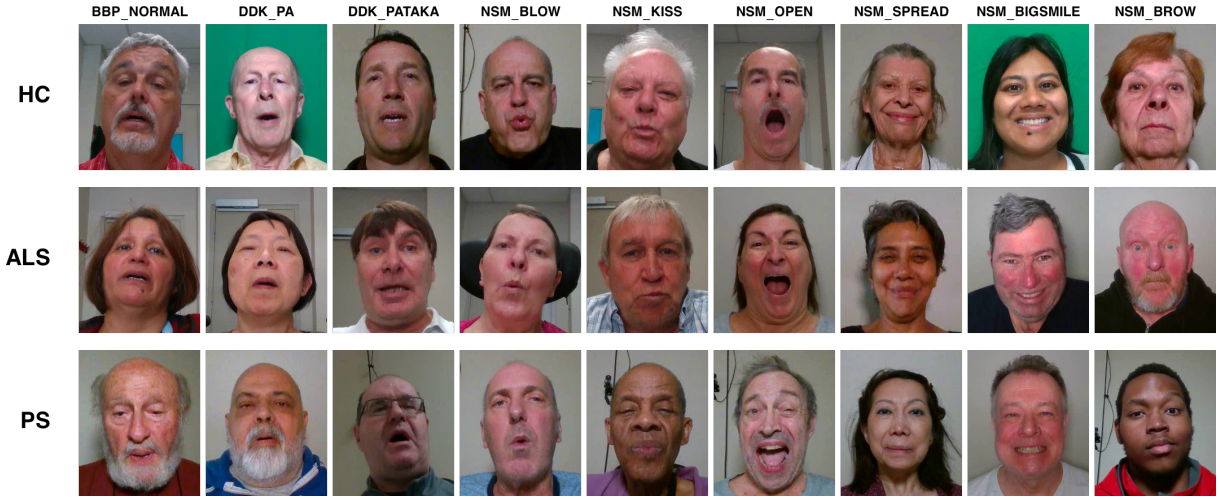
Tabela 6 – Descrição das expressões faciais

Código	**Descrição**
BBP_NORMAL	Repetição da sentença "*Buy Bobby a Puppy*"
DDK_PA	Repetição da sílaba /*pa*/ o mais rápido com uma única respiração
DDK_PATAKA	Repetição das sílabas /*pataka*/ o mais rápido com uma única respiração
NSM_BLOW	Imitar um assoprar de velas
NSM_KISS	Imitar um beijo
NSM_OPEN	Abrir a mandíbula ao máximo
NSM_SPREAD	Fingir um sorriso de lábios fechados
NSM_BIGSMILE	Mostrar um grande sorriso
NSM_BROW	Levantar as sobrancelhas

É importante salientar que a distribuição das imagens dentro do conjunto não é uniforme em relação às expressões faciais. A Tabela 7 e a Figura 20 enumeram os quantitativos de imagens faciais por expressão e grupo clínico.

A composição desse *dataset* serve ao propósito de analisar, exclusivamente, questões relativas ao viés algorítmico. Diferentemente dos *datasets* usados para treinamento dos modelos, o *Toronto Neuroface* não possui imagens faciais com oclusão, grande variação de pose facial, variação de iluminação, etc. A ausência de desafios, comumente observados em *benchmarks* de alinhamento facial, é coerente com esta pesquisa, uma vez que auxilia na exclusão de características que poderiam atrapalhar a análise sobre o viés observado.

Figura 19 – Amostras do *Toronto Neuroface*.



Fonte: Elaborado pelo autor.

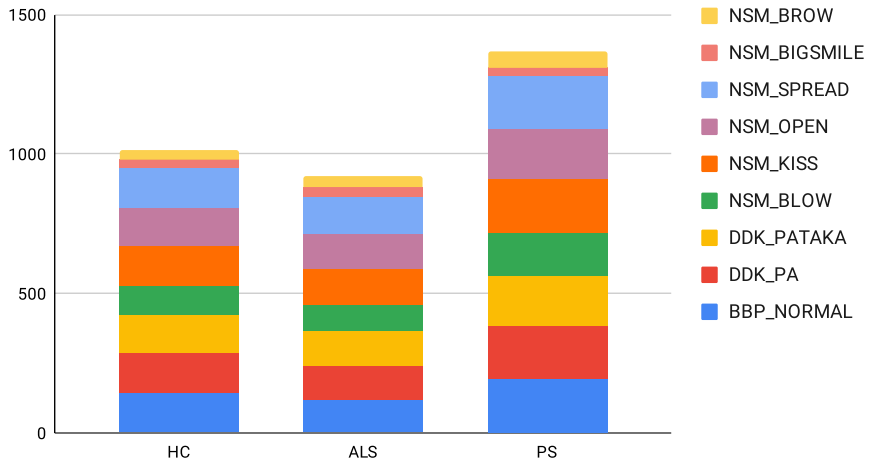
Tabela 7 – Quantitativo de imagens faciais do *Toronto Neuroface*

Expressão	**Grupos**		
	HC	*ALS*	*PS*
BBP_NORMAL	145	119	194
DDK_PA	141	126	191
DDK_PATAKA	138	122	178
NSM_BLOW	105	90	157
NSM_KISS	141	129	192
NSM_OPEN	140	130	183
NSM_SPREAD	141	131	185
NSM_BIGSMILE	27	36	27
NSM_BROW	36	36	63
Total	**1014**	**919**	**1370**

4.9 Considerações Finais

O método descrito neste capítulo foi testado no *Toronto Neuroface* em alguns cenários para investigação das hipóteses levantadas no Capítulo 1. O próximo capítulo descreve os cenários, os resultados obtidos, e faz uma discussão sobre os desempenhos observados.

Figura 20 – Quantitativo de imagens do *Toronto Neuroface*.



Fonte: Elaborado pelo autor.

5 Resultados

Diversos modelos foram gerados a partir da metodologia desenvolvida com o objetivo de testar as hipóteses levantadas no Capítulo 1, utilizando o *Toronto Neuroface* como *dataset* de referência, e de aprimorar os resultados do alinhamento facial. Os resultados são expostos de acordo com as métricas de avaliação descritas na Seção 3.4. Também foram realizados testes com os modelos do estado da arte do alinhamento facial: *FAN-2D* (BULAT; TZIMIROPOULOS, 2017b); *ADNet* (HUANG et al., 2021); e *SPIGA* (PRADOS-TORREBLANCA; BUENAPOSADA; BAUMELA, 2022), visando caracterizar o viés algorítmico em relação a indivíduos com problemas neurológicos e estabelecer um comparativo com os modelos deste trabalho. Os modelos da *ADNet* e *SPIGA* foram gerados com o *dataset* de treinamento do desafio 300W (SAGONAS et al., 2016). A *FAN-2D* foi treinada com o conjunto *300W-LP-2D* (BULAT; TZIMIROPOULOS, 2017b). Os modelos gerados pela metodologia desenvolvida neste trabalho utilizaram o subconjunto do *300W* com *Helen*, *AFW*, *LFPW*, e imagens *indoor* do desafio como treino. Os seguintes experimentos e análises foram realizados:

- Cenário ideal: teste dos modelos de localização de pontos das subunidades considerando um cenário no qual a detecção de subunidades é conhecimento prévio. Nesse cenário, as caixas delimitadoras das subunidades que são usadas pelos modelos de localização de pontos são definidas com base na localização real dos centros das subunidades. Este cenário visa evidenciar o conceito de que a divisão da análise facial em subunidades é capaz de reduzir as diferenças de desempenho entre grupos em comparação ao estado da arte do alinhamento facial;

- Modelo de rotação: teste do modelo de ajuste de rotação. São apresentados os resultados da etapa de uniformização de rotação bem como o impacto desse ajuste sobre o modelo de localização de pontos;

- Modelo de detecção de subunidade: teste do modelo de localização dos elementos faciais. Diversos experimentos foram realizados com o modelo do nível superior do método apresentado. Foram realizadas análises isoladas do seu comportamento sob aspectos como expressão facial, erro de localização por elemento facial, e aspecto da caixa delimitadora;

- Modelos da abordagem: teste do método como um todo. Nesse cenário, os melhores modelos gerados a partir da metodologia desenvolvida são testados considerando conhecimento prévio somente da detecção facial. Este cenário visa

demonstrar que a metodologia também mostra eficácia na redução das diferenças de desempenho. Além disso, verifica se a estratégia de divisão do alinhamento facial em subunidades contribui efetivamente para essa redução em comparação a um modelo gerado sem a divisão facial. Adicionalmente, é feito o teste de um modelo gerado com adição de exemplares do *Toronto Neuroface*;

- Análise por expressões faciais: investigação sobre o efeito das expressões faciais nos resultados da metodologia desenvolvida. É feita uma análise dos resultados da metodologia e dos modelos do estado da arte dividida pelas expressões faciais presentes e rotuladas no conjunto de teste como apresentado na Seção 4.8;
- Desempenho geral: Análise dos resultados gerais (i.e. sem distinção de grupo) dos modelos gerados comparados ao estado da arte.

Após compilação dos resultados, todos os modelos apresentaram, em alguma medida, desempenhos muito discrepantes para determinados casos de teste. Para evitar que esses casos prejudicassem a análise estatística dos modelos, foi adotada uma estratégia de remoção de *outliers* baseada no Z-escore ($Z = (x - \mu)/\sigma$) no qual resultados cujo módulo do escore $|Z| > 3$ são removidos. Os casos removidos não representaram grande mudança quantitativa no conjunto de teste (<1%).

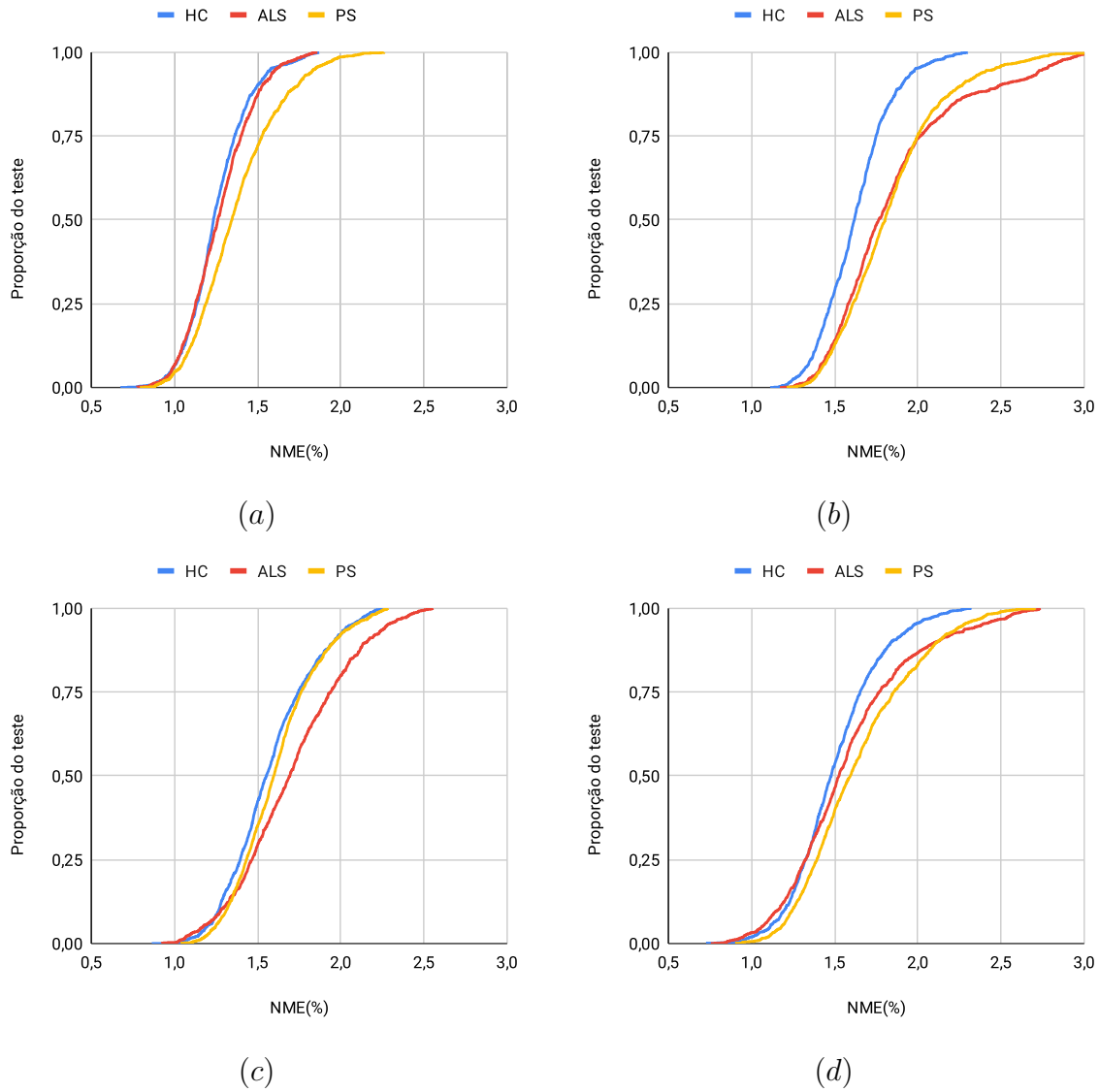
5.1 Cenário Ideal

Uma vez que a assimetria orofacial foi sugerida como a questão mais problemática em relação ao viés algorítmico na população clínica avaliada (BANDINI et al., 2021), este trabalho foi orientado pela divisão semântica da análise facial em subunidades, visando ao aumento da variabilidade do modelo. Portanto, antes de testar todo o método para demonstrar que essa noção poderia de fato reduzir o viés, os modelos de localização de pontos de cada subunidade foram aplicados individualmente. Essa abordagem inicial foi importante para iniciar a prova de conceito, além de permitir estimar os melhores resultados que a metodologia poderia alcançar.

Assim, foi simulado um cenário, baseado em anotações manuais, onde as subunidades são detectadas idealmente no conjunto de dados de teste. Cada modelo de localização de pontos das subunidades foi treinado individualmente e foram aplicados ao *Toronto Neuroface* separados por grupos (*HC*, *ALS*, e *PS*). Em seguida, foi calculada a média do erro normalizado como uma proporção da diagonal da caixa delimitadora da face. Os resultados da *FAN-2D* foram obtidos após a aplicação do modelo pré-treinado (*300W-LP-2D* com $\sim$61k imagens) disponibilizado por Bulat e Tzimiropoulos (2017b). Os modelos *ADNet* e *SPIGA* disponibilizados por Huang et al. (2021) e Prados-Torreblanca, Buenaposada e Baumela (2022) pré-treinados com o

300W também foram utilizados. Inicializamos todos os modelos com a caixa delimitadora da face e calculamos o erro da mesma forma. As Figuras 21 (b), (c), e (d) mostram o erro cumulativo dos modelos *FAN-2D*, *ADNet*, e *SPIGA* aplicados aos 3 grupos. Os resultados da *FAN-2D* e do *SPIGA* para o grupo *HC* indicam um desempenho significativamente melhor em relação aos grupos *ALS* e *PS*, indicando existência de viés algorítmico. O *ADNet* apresentou resultados mais próximos para os grupos, contudo, pode-se observar que o grupo *ALS* foi mais prejudicado.

Figura 21 – *CEDs* dos modelos de localização de pontos em cenário ideal e do estado da arte separados por grupo clínico do *Toronto Neuroface*. (a) Modelo de localização de pontos em cenário ideal. (b) *FAN-2D*. (c) *ADNet*. (d) *SPIGA*.



Fonte: Elaborado pelo autor.

A Figura 21 (a) mostra o erro cumulativo agregado dos modelos de localização de pontos das subunidades aplicados no cenário ideal[1]. Através das *CEDs*, é possível

[1] Os desempenhos de todos os modelos são ponderados e agregados.

observar que esses modelos são capazes de reduzir significativamente as diferenças de erro entre os 3 grupos quando comparados à *FAN-2D* e ao *SPIGA*. Essa abordagem aplicada ao grupo *PS* ainda demonstra um desempenho um pouco pior em relação ao *HC* e *ALS*, porém, as lacunas entre os grupos são mais próximas no gráfico.

Esse cenário experimental demonstrou que a abordagem multi-nível apresentada poderia idealmente reduzir o viés algorítmico entre os grupos clínicos avaliados, caso a detecção das subunidades não apresente falhas. Também mostrou que essa abordagem poderia potencialmente superar métodos de última geração para alinhamento facial sob o critério de redução do viés.

5.2 Configuração das Subunidades

O mecanismo de modelagem apresentado na metodologia deste trabalho representou uma grande vantagem na condução desta pesquisa por conta da maleabilidade para a criação de novos modelos de alinhamento facial. Durante o desenvolvimento da pesquisa vários modelos foram criados e seus resultados nortearam novas modelagens para os componentes faciais. Foi observado, por exemplo, que para uma modelagem facial em 8 subunidades como a apresentada no Capítulo 4, o elemento que representa a boca obtém bons resultados na maior parte do *dataset*. Contudo, ao avaliar expressões que afetam diretamente essa subunidade, observa-se que o desempenho do modelo tem piora significativa. A Tabela 8 mostra os desempenhos (*NME*) de um modelo avaliado por subunidade e expressão facial no *dataset*.

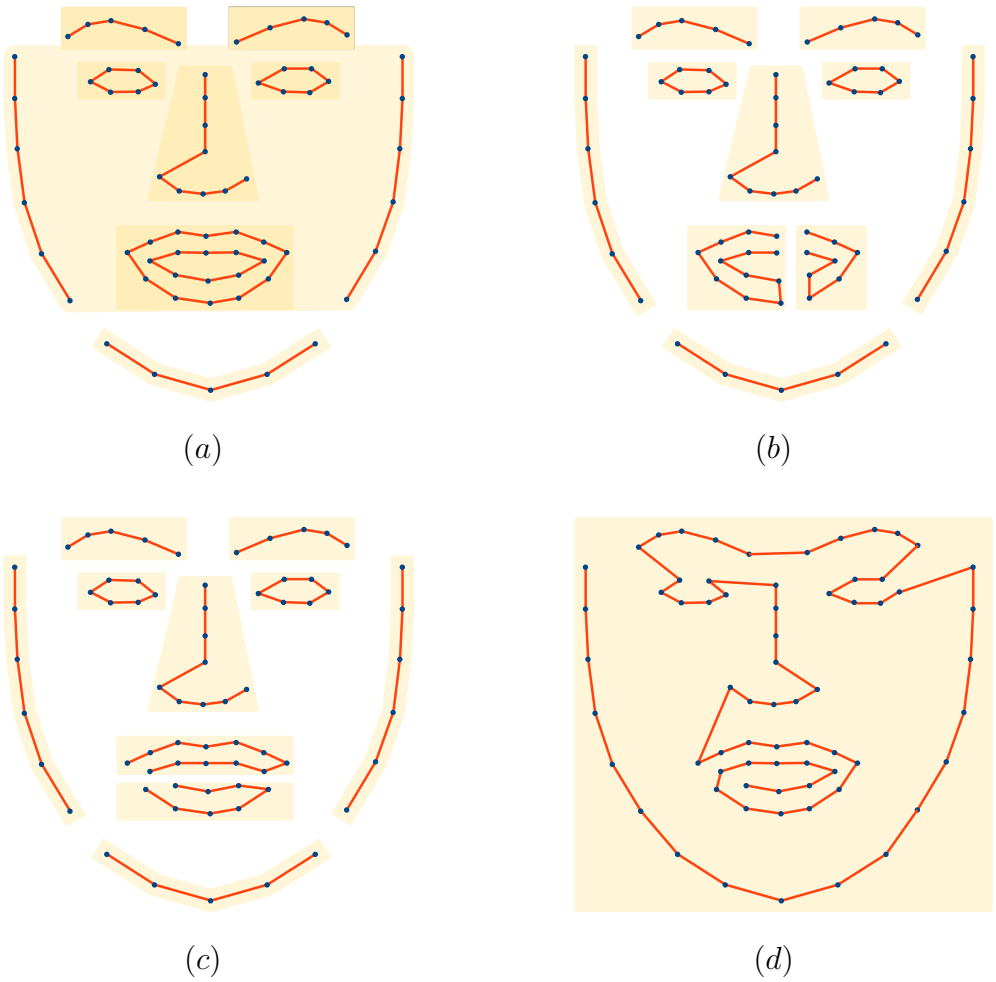
Tabela 8 – Relação do desempenho da localização de pontos (*NME*%) entre expressão e elemento facial agrupados por subunidade. A intensidade da coloração é proporcional ao desempenho relativo a todos os dados da tabela

	Subunidade Modelada							
Expressão/Tarefa	**S0**	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**
BBP_NORMAL	2,99	2,94	2,83	3,10	1,04	0,86	0,75	1,19
DDK_PA	3,18	2,76	2,93	3,14	1,16	0,79	0,74	1,12
DDK_PATAKA	3,08	2,69	2,98	3,34	0,99	0,89	0,68	1,22
NSM_BIGSMILE	2,25	2,36	2,15	2,08	0,80	0,60	0,45	1,23
NSM_BLOW	3,09	2,85	3,03	3,42	1,34	0,92	0,78	1,16
NSM_BROW	2,57	2,38	3,48	2,84	0,83	0,53	0,47	1,02
NSM_KISS	2,90	2,71	2,59	2,93	0,98	0,84	0,64	0,97
NSM_OPEN	3,03	2,82	3,24	3,37	2,95	1,21	1,22	2,76
NSM_SPREAD	2,74	2,73	2,52	2,87	0,83	0,80	0,62	1,37

Como observado por Bandini et al. (2021), as expressões faciais realizadas por pacientes com problemas neurológicos em exames clínicos potencializam a assimetria orofacial e impactam nos modelos de alinhamento facial. Durante a pesquisa foi observado que os resultados dos modelos para os pontos localizados na região compreendida por boca, queixo, e mandíbula foram inferiores, especialmente em imagens da expressão NSM_OPEN. Algumas configurações de subunidades foram concebidas

para tentar estimar melhor as localizações dos pontos nessas situações. A Figura 22 ilustra algumas configurações elaboradas neste trabalho.

Figura 22 – Configurações de subunidades faciais modeladas no trabalho. Cada configuração gerou um modelo de localização de pontos faciais. (a) Configuração do modelo M_0. (b) Configuração do modelo M_{S7h}. (c) Configuração do modelo M_{S7v}. (d) Configuração do modelo M_{68}



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

Fonte: Elaborado pelo autor.

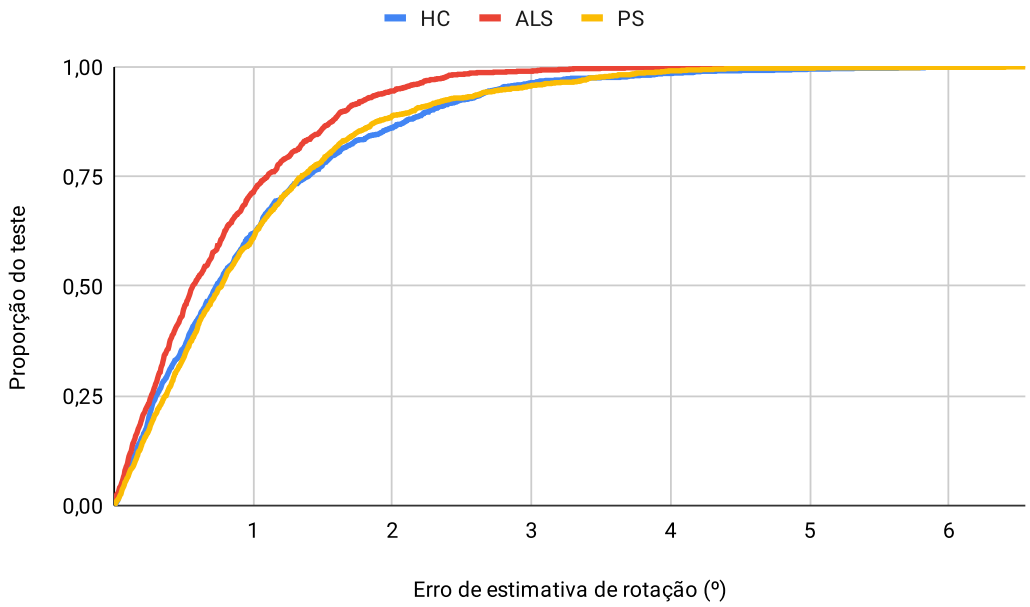
A configuração ilustrada na Figura 22 (a) é idêntica à apresentada no Capítulo 4. As configurações das Figuras 22 (b) e 22 (c) foram elaboradas para experimentação de diferentes modelos de localização de pontos para a boca, denominados M_{S7h} e M_{S7v} respectivamente. As Figuras 22 (a), 22 (b), e 22 (c) seguiram a ideia deste trabalho de divisão de subunidades faciais. A configuração da Figura 22 (d) foi elaborada sem a divisão facial, ou seja, considera todos os 68 pontos do padrão de anotação como uma única unidade. Ela serviu de referência ao teste da segunda hipótese apresentada no Capítulo 1.

5.3 Modelo de Rotação

A etapa inicial da metodologia consiste na uniformização de rotação das faces em torno do eixo longitudinal com o objetivo de potencializar os modelos de localização de pontos faciais. A ideia de uniformização da face para melhoria do desempenho de aplicações faciais já foi observada para outras tarefas dentro de um pipeline de aplicações faciais (JIN; TAN, 2017). Neste trabalho, essa ideia é aplicada para ajuste de rotação das faces a fim de padronizar as entradas. As faces presentes no *dataset Toronto Neuroface* não possuem grande variação de rotação, contudo, foi possível observar que alguns indivíduos apresentam alguma variação nessa característica, mesmo em posição facial natural.

O modelo construído a partir da metodologia estima a variação de rotação da face em relação à posição horizontal. Após o cálculo, a imagem é rotacionada no sentido contrário para concluir o ajuste. Em um cenário ideal, com a face perfeitamente ajustada, os modelos de localização de pontos são capazes de superar o desempenho do estado da arte. Entretanto, as estimativas do modelo, chamado de M_θ, apresentam algum nível de erro que é propagado para os modelos posteriores. A Figura 23 ilustra a distribuição do erro, em graus, no cálculo da rotação após aplicação no *dataset Toronto Neuroface* em cada grupo. É possível observar que menos de 5% dos casos de teste apresenta um erro superior a 3°, como exibido na Tabela 9.

Figura 23 – *CED* do modelo M_θ para estimativa de rotação em cada grupo.



Fonte: Elaborado pelo autor.

Uma vez que este trabalho está assentado na solução do problema do viés algo-

Tabela 9 – Erro da estimativa de rotação do modelo M_θ

Métrica	**Grupo**		
	HC	**ALS**	**PS**
Média	0,999	0,761	0,997
σ	0,936	0,659	0,888
AUC_{3°	0,681	0,750	0,680
TF_{3° (%)	3,60	1,00	4,30

rítmico em populações clínicas, e considerando que a abordagem é sedimentada em modelos de regressão em diferentes níveis, é interessante entender qual a contribuição de cada nível para o viés e para sua eventual redução. Os resultados atingidos com o modelo M_θ também foram avaliados nesse contexto. A Figura 23 e a Tabela 9 mostram que os resultados obtidos pelo modelo são bastante próximos, contudo, é necessária uma avaliação estatística para verificar se as diferenças de desempenho do modelo M_θ são significativas.

A primeira parte da Tabela 10 mostra os resultados das aplicações dos métodos de Kruskal-Wallis e Dunn para testar se os desempenhos do modelo nos grupos *HC*, *ALS*, e *PS* podem ser considerados originários da mesma distribuição, o que indicaria ausência de favorecimento pelo modelo. Verificou-se que os resultados falharam no teste de Kruskal-Wallis. Então, o teste de comparações de Dunn indicou que os pares *HC*-*ALS* e *ALS*-*PS* são avaliados distintamente pelo modelo. Contudo, não se pode rejeitar a hipótese de o modelo gerar resultados igualitários para os grupos HC-PS, uma vez que o valor p calculado após comparações dos resultados desses grupos está acima do limiar de significância de 5%.

Além disso, a terceira seção da tabela mostra os resultados dos mesmos testes aplicados aos desempenhos do modelo M_0 e de uma versão dele próprio treinado sem o ajuste de rotação (M_0 sem rotação). O modelo M_0, com correção de rotação, obteve resultados com diferenças insignificantes para o par *HC*-*ALS*, enquanto que o mesmo modelo sem o ajuste não produziu desempenhos igualitários entre nenhum dos grupos avaliados.

5.4 Modelo de Detecção de Subunidade

Após o ajuste de rotação, a imagem facial serve de entrada para o modelo de detecção de subunidades faciais. Esse modelo, chamado M_S, é dependente da configuração de subunidades faciais definida em fase de projeto dos modelos. Neste trabalho, algumas configurações distintas foram exploradas tanto no sentindo de testar as hipóteses sugeridas, quanto com o objetivo de encontrar melhores modelagens. A configuração de subunidades para a detecção abordada nesta seção é oriunda

Tabela 10 – Testes estatísticos para comparação das diferenças de desempenho entre grupos: 1) modelo de ajuste de rotação; 2) modelo de detecção de subunidades (gerado para M0); 3) modelos de localização de pontos faciais. Em negrito, resultados que indicam diferenças não significativas

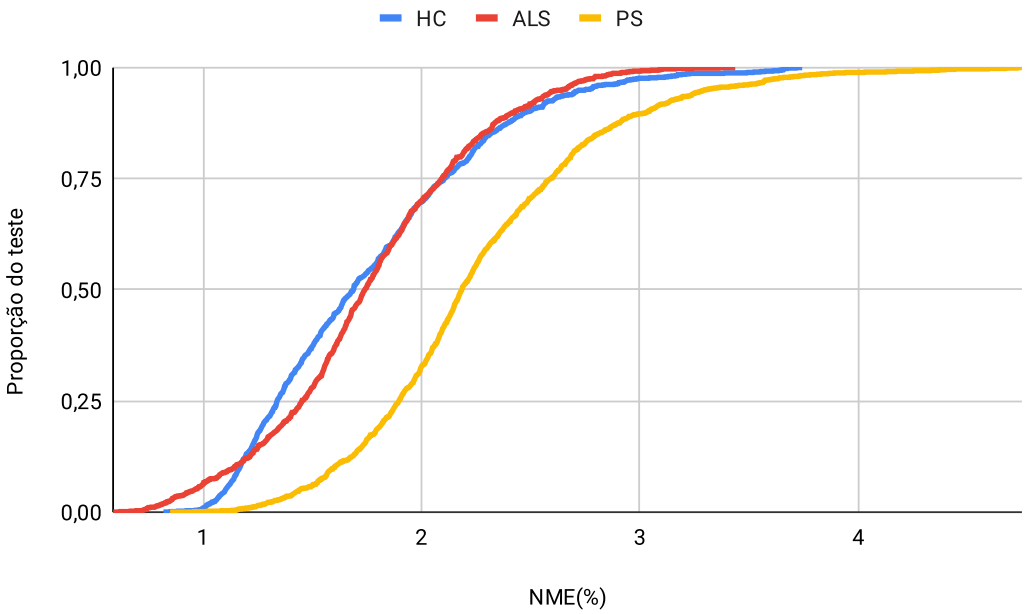
Modelo	**Kruskal-Wallis**		**Teste de Dunn**					
	H	**p**	**HC-ALS**		**HC-PS**		**ALS-PS**	
			p	**Signf.**	**p**	**Signf.**	**p**	**Signf.**
Mθ	40,98	1,26e-09	7,14e-07	****	**8,62e-1**	**ns**	8,67e-10	****
MS	514,09	< 2,20e-16	**1,38e-01**	**ns**	5,00e-78	****	8,88e-84	****
MS $\rightarrow S_0$	22,70	1,17e-05	9,52e-06	****	1,42e-04	***	**2,94e-01**	**ns**
MS $\rightarrow S_1$	63,53	1,59e-14	6,05e-03	*	1,11e-14	****	6,91e-07	****
MS $\rightarrow S_2$	245,88	< 2,20e-16	3,07e-06	****	2,88e-52	****	2,53e-22	****
MS $\rightarrow S_3$	50,02	1,37e-11	4,52e-12	****	**1,89e-02**	**ns**	3,12e-07	****
MS $\rightarrow S_4$	327,73	< 2,20e-16	**1,56e-01**	**ns**	1,07e-58	****	4,32e-45	****
MS $\rightarrow S_5$	67,73	1,95e-15	7,69e-03	*	9,22e-16	****	8,66e-07	****
MS $\rightarrow S_6$	200,19	< 2,20e-16	3,11e-17	****	4,48e-45	****	4,21e-06	****
MS $\rightarrow S_7$	71,91	2,42e-16	**2,47e-01**	**ns**	7,39e-15	****	3,43e-10	****
M0	152,82	< 2,20e-16	**6,98e-01**	**ns**	3,99e-20	****	1,88e-17	****
M0 (sem rotação)	215,11	< 2,20e-16	2,05e-17	****	1,69e-48	****	2,18e-07	****
M0 (*Toronto Neuroface*)	616,21	< 2,20e-16	8,85e-08	****	5,6e-122	****	1,11e-64	****
Ideal	152,82	< 2,20e-16	**7,54e-02**	**ns**	7,69e-30	****	5,34e-20	****
MS7h	48,86	2,45e-11	6,55e-08	****	3,95e-11	****	**5,15e-01**	**ns**
MS7v	77,62	< 2,20e-16	**4,57e-02**	**ns**	4,42e-10	****	1,72e-16	****
M68	183,75	< 2,20e-16	5,21e-07	****	4,95e-16	****	4,77e-40	****
FAN-2D	324,02	< 2,20e-16	9,99e-46	****	7,51e-64	****	**2,14e-01**	**ns**
ADNet	83,68	< 2,20e-16	2,67e-19	****	1,34e-03	**	1,12e-10	****
SPIGA	87,86	< 2,20e-16	1,14e-04	***	1,42e-20	****	8,57e-07	****

do modelo M0, que obteve os melhores resultados para a localização de pontos nos experimentos realizados com a metodologia. A configuração está ilustrada na Figura 22 (a).

Os resultados gerais do modelo MS estão ilustrados na Figura 24. É possível observar que ele apresenta resultados inferiores para o grupo *PS* em relação aos demais. Na segunda seção da Tabela 10 é possível confirmar que os resultados para os grupos *HC* e *ALS* podem ser considerados igualitários, enquanto o grupo PS é menos favorecido.

A análise dos resultados de detecção dos elementos faciais pode indicar possíveis deficiências no modelo apresentado e apontar para aprimoramentos que o conduzam para desempenhos mais igualitários, além de melhorar os resultados gerais. Assim, é importante entender onde o erro se manifesta mais intensamente para buscar soluções que atenuem o viés e seus efeitos. A Figura 25 revela como o modelo MS desempenha para cada subunidade da configuração escolhida em função dos grupos avaliados neste trabalho. Os resultados obtidos para a detecção das subunidades S0 e S7 indicam bastante proximidade entre os grupos. Isso é confirmado pelos testes de Dunn realizados, nos quais MS desempenhou estatisticamente semelhante na localização da subunidade S0 para os grupos *ALS* e *PS*, e na localização de S7 para os grupos *HC* e *ALS*.

Figura 24 – *CED* do modelo de detecção de subunidades para cada grupo.



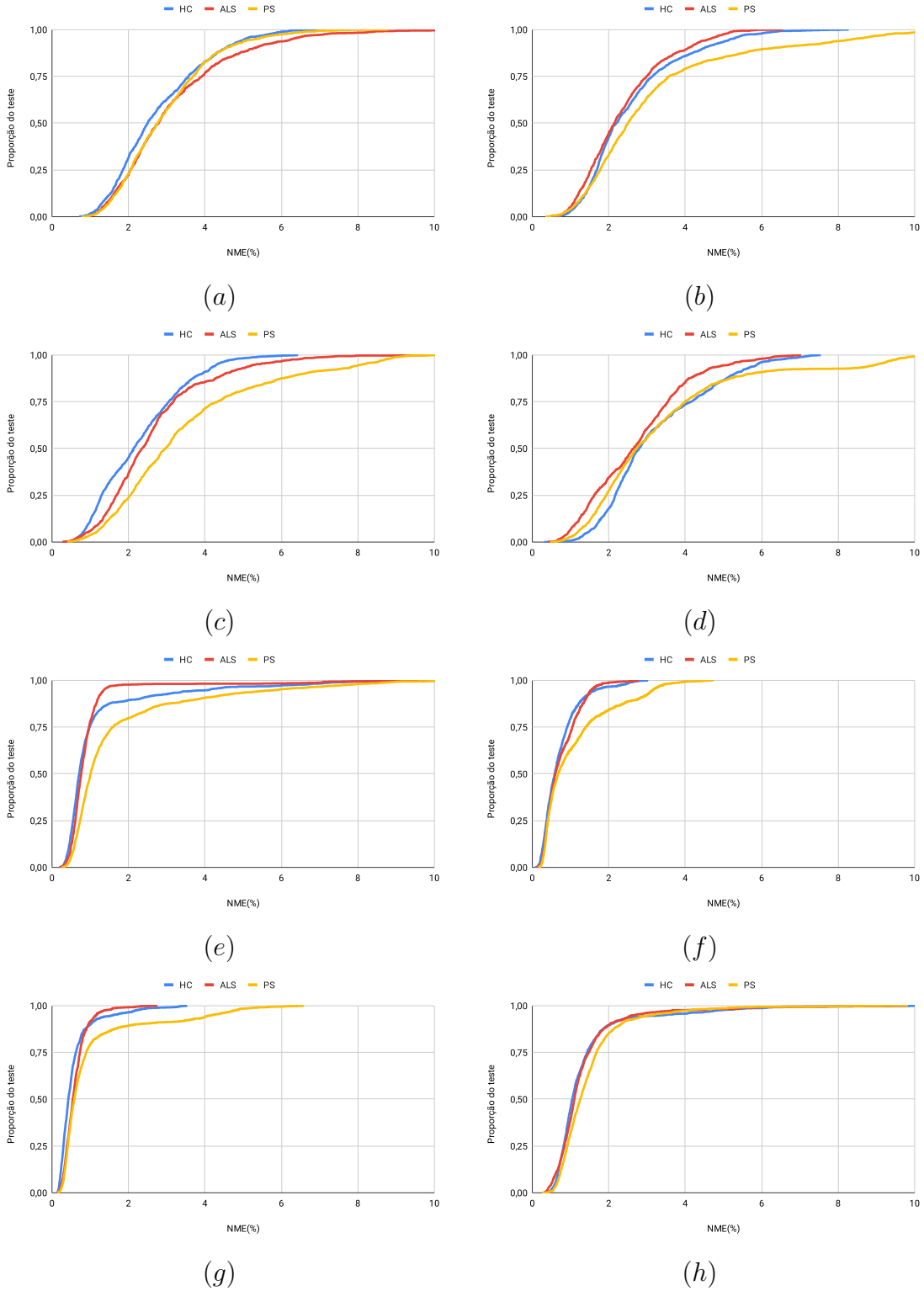
Fonte: Elaborado pelo autor.

Apesar de não ser visualmente aparente no gráfico, os testes também apontaram igualdade nos resultados do modelo para as detecções das subunidades S_3 e S_4, considerando os pares *HC-PS* e *HC-ALS* respectivamente. A Figura 26 mostra os resultados para o método N-Sigma, que quantifica as diferenças entre amostras. Nesse experimento, o conjunto de referência estabelecido para o cálculo da métrica foi o *HC*. Os valores encontrados indicam que existe grande disparidade nos resultados obtidos pelo modelo na localização das subunidades S_2, S_5, e S_6, especialmente em relação ao grupo *PS*.

Ao analisar o desempenho do modelo M_S pelos grupos clínicos, dessa vez, em função das subunidades, surge um aspecto interessante do modelo. A Figura 27 mostra o desempenho do modelo M_S em cada grupo clínico dividido por subunidade. É interessante notar que o modelo se comporta de maneira semelhante para todos os grupos. Nesse caso, os resultados de detecção das subunidades S_0, S_1, S_2, e S_3 são visivelmente piores que os resultados das demais em todos os grupos. Essas subunidades correspondem aos elementos mais extremos da face na configuração adotada: mandíbula, queixo, e sobrancelhas.

Considerando as rotulações fornecidas pelo *dataset Toronto Neuroface*, pode-se realizar também uma avaliação do modelo com base nas expressões ou tarefas faciais do conjunto. As expressões e tarefas faciais realizadas pelos participantes do *Toronto Neuroface* apresentam diferentes aspectos da avaliação do grau de severidade das enfermidades de cada indivíduo. Algumas tarefas exigem movimentos faciais mais

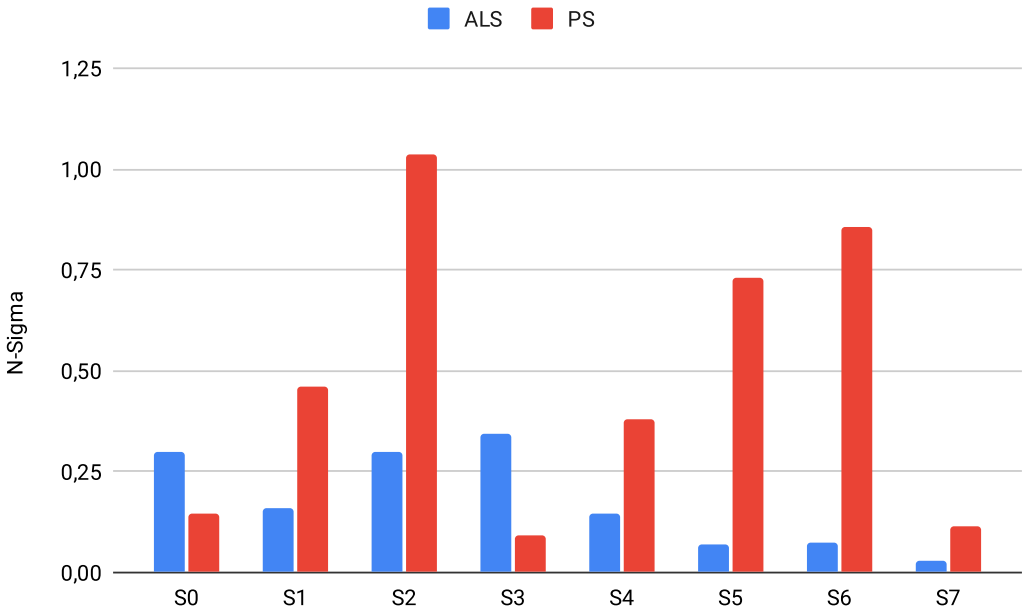
Figura 25 – *CEDs* do modelo M_S discriminadas por subunidade modelada e divididas por grupo clínico. Os gráficos ilustram os resultados das seguintes subunidades: (a) S_0. (b) S_1. (c) S_2. (d) S_3. (e) S_4. (f) S_5. (g) S_6. (h) S_7.



Fonte: Elaborado pelo autor.

extremos e contribuem para a piora no desempenho do detector de subunidades. Uma possível consequência negativa desses movimentos faciais extremos é a potencializa-

Figura 26 – Gráfico comparativo do valor N-Sigma calculado para os resultados de cada subunidade usando o grupo de controle como referência.



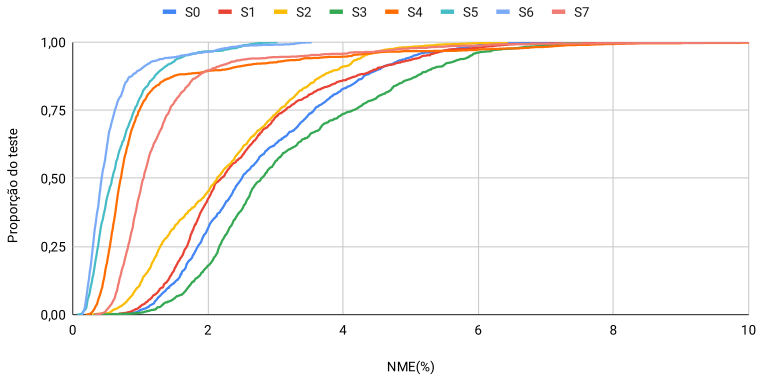
Fonte: Elaborado pelo autor.

ção da assimetria orofacial em indivíduos portadores de ELA ou vítimas de derrame. Fato que pode contribuir para o aumento das diferenças de desempenho. A Figura 28 compila os resultados do detector de subunidades faciais na localização de cada elemento com respeito à expressão executada. O primeiro gráfico mostra os resultados de localização das subunidades agregados por tarefa facial. É possível perceber que as tarefas NSM_OPEN e NSM_BLOW provocam os piores resultados gerais para a detecção das subunidades. Por outro lado, BIG_SMILE representa uma expressão onde o erro de localização das subunidades é menor. Os números podem ser melhor visualizados na Tabela 11.

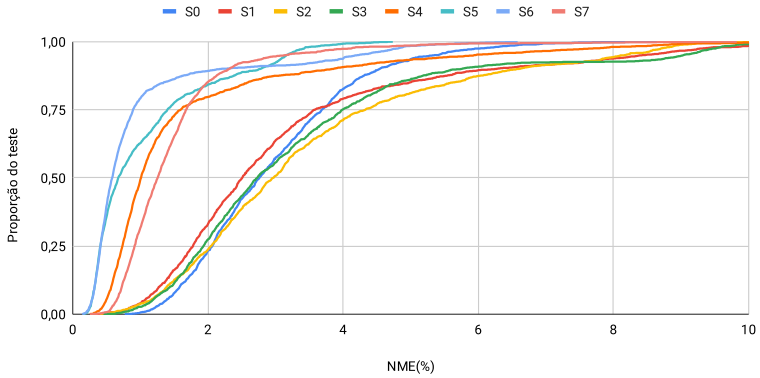
Contudo, é preciso ter em mente que cada subunidade localizada por M_S é composta por uma quantidade distinta de pontos faciais. Isso significa que a localização de cada subunidade tem uma influência particular no nível final do método apresentado. Após realizar uma ponderação dos resultados de localização das subunidades proporcional à quantidade de pontos que compõem cada uma, observa-se que, apesar de não apresentar um erro médio muito significativo, pode-se interpretar que as subunidades S_0 e S_7 correspondem praticamente por metade do erro do modelo, considerando que elas agrupam 32 dos 68 pontos do padrão de anotação (Figura 28 (b)).

Outro aspecto crucial para ponderação da análise é a quantidade de imagens faciais correspondentes a cada expressão. A Figura 28 (c) mostra o mesmo resultado da Figura 28 (a), porém, agora ponderado pela quantidade de imagens disponíveis de cada

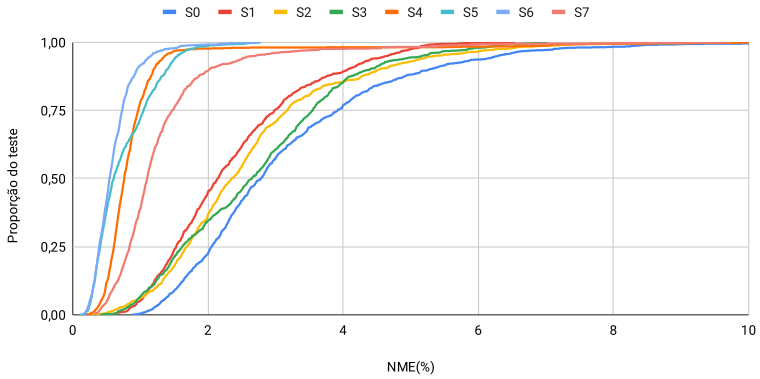
Figura 27 – *CEDs* do modelo Ms discriminadas por grupo e divididas por subunidade. Os gráficos ilustram os resultados dos seguintes grupos: (a) *HC*. (b) *ALS*. (c) *PS*.



(a)



(b)



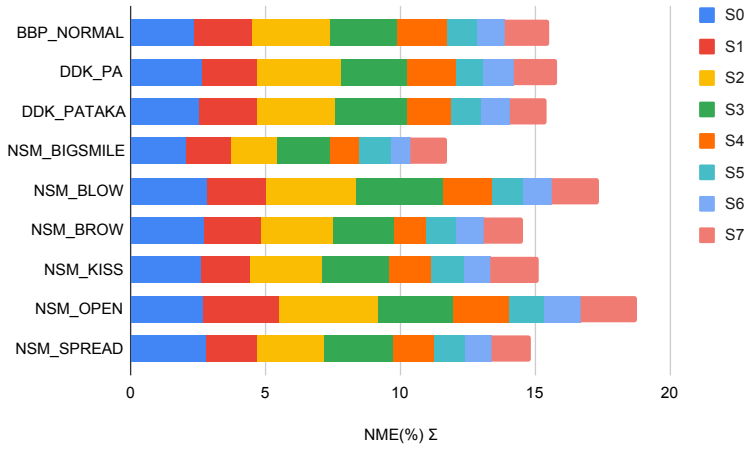
(c)

Fonte: Elaborado pelo autor.

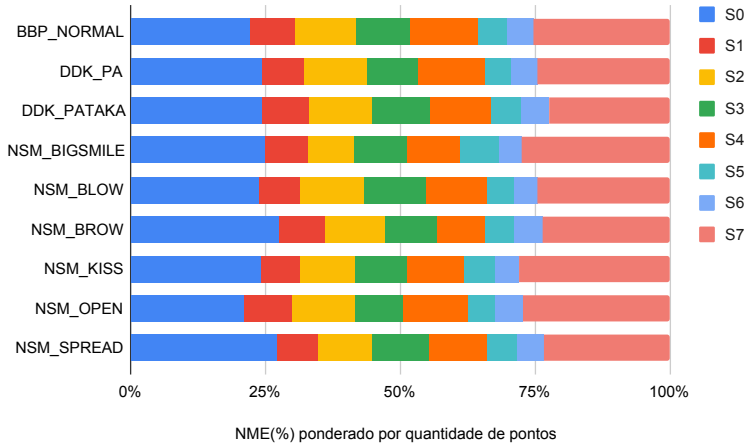
tarefa. Isso possibilita observar que os resultados apresentados para as expressões BIG_SMILE e NSM_BROW precisam ser relativizados em face do reduzido número de imagens.

Por último, foi realizada uma investigação sobre a caixa delimitadora da face de entrada. A metodologia descreve a realização de ajustes de escala de tal forma

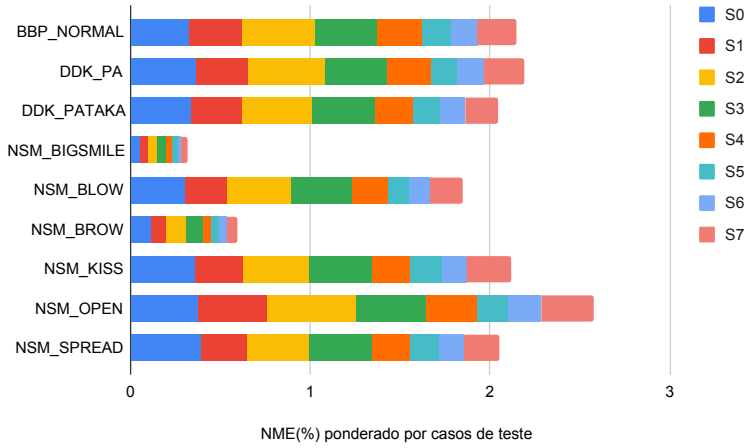
Figura 28 – Resultados gerais do modelo Ms separados por expressão e subunidade.



(a)



(b)



(c)

Fonte: Elaborado pelo autor.

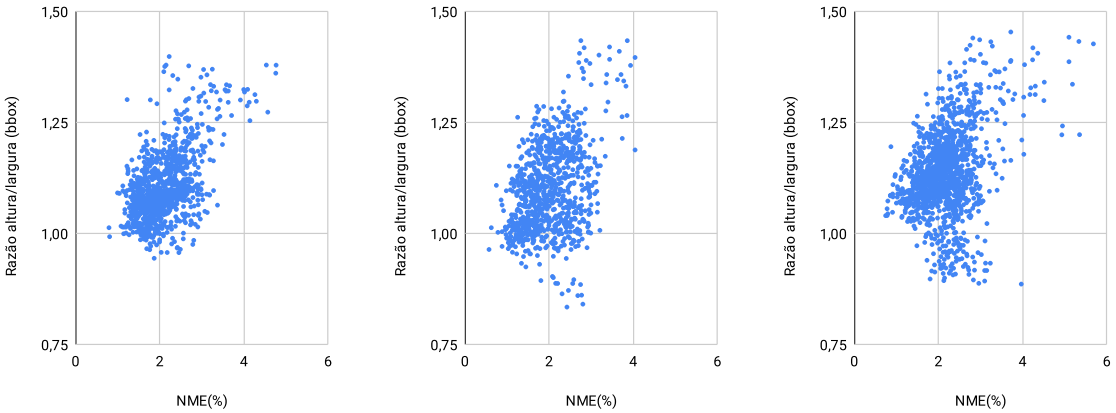
que as imagens de entrada estejam em conformidade com as estruturas de *CNNs* exploradas. Para isso, as imagens devem ser ajustadas de maneira a terem lados iguais

Tabela 11 – Relação do desempenho da detecção de subunidade (*NME*%) entre expressão e elemento facial. A intensidade da coloração é proporcional ao desempenho relativo a todos os dados da tabela

	Subunidade Modelada							
Expressão/Tarefa	S0	S1	S2	S3	S4	S5	S6	S7
BBP_NORMAL	2,36	2,12	2,90	2,52	1,81	1,12	1,06	1,61
DDK_PA	2,65	2,03	3,10	2,48	1,79	1,04	1,09	1,61
DDK_PATAKA	2,54	2,15	2,90	2,67	1,59	1,12	1,07	1,40
NSM_BIGSMILE	2,07	1,62	1,73	1,96	1,10	1,18	0,70	1,38
NSM_BLOW	2,83	2,19	3,36	3,23	1,81	1,15	1,04	1,76
NSM_BROW	2,76	2,06	2,68	2,29	1,17	1,11	1,03	1,43
NSM_KISS	2,60	1,87	2,65	2,49	1,52	1,24	0,97	1,80
NSM_OPEN	2,71	2,82	3,62	2,82	2,07	1,31	1,32	2,12
NSM_SPREAD	2,79	1,89	2,49	2,55	1,50	1,19	0,96	1,45

e incluírem todos os elementos faciais. Contudo, esse ajuste pode forçar a inclusão de atributos do plano de fundo na imagem facial, principalmente em imagens faciais cuja razão alturaxlargura sejam muito distantes de 1:1. É importante salientar que não foi realizado nenhum *padding* nas imagens faciais de entrada. A Figura 29 mostra a razão alturaxlargura em função do erro médio de localização das subunidades (*NME*). A partir do índice de Pearson, é possível verificar correlação moderada entre as duas variáveis nos resultados do grupo HC (Figura 29 (a)), e correlação fraca nos outros dois (*ALS* e *PS*).

Figura 29 – Gráficos de dispersão entre razão altura/largura da caixa delimitadora da face e o erro médio de localização de subunidades. (a) *HC*. (b) *ALS*. (c) *PS*.

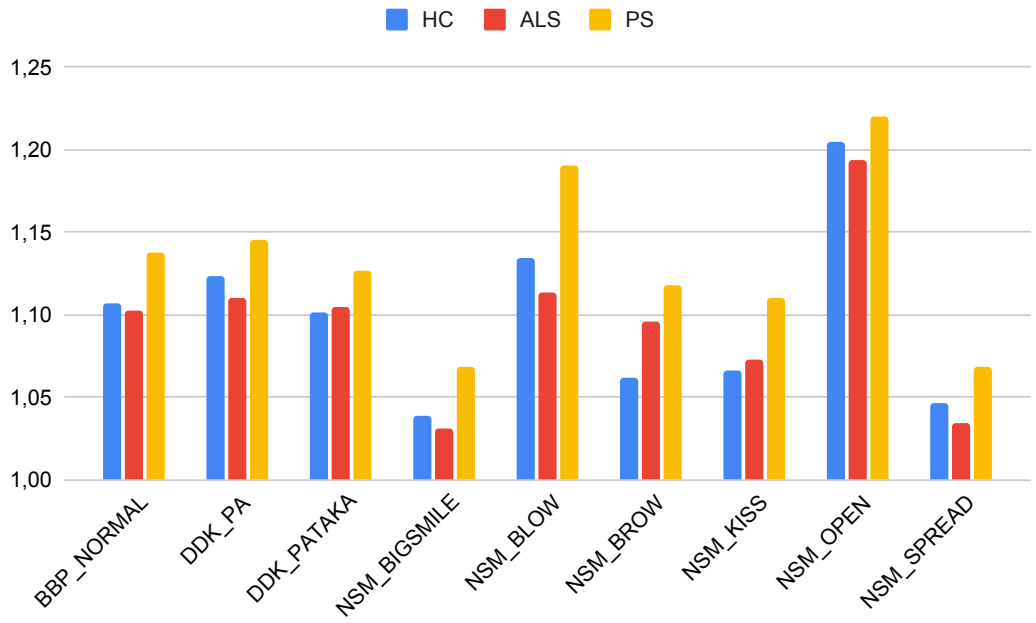


(a) (b) (c)

Fonte: Elaborado pelo autor.

Apesar de a correlação não ser muito forte, observou-se que houve impacto dessa característica na forma como o método realizou a detecção facial, em especial para as expressões NSM_OPEN e NSM_BLOW. O gráfico da Figura 30 mostra o valor médio da razão alturaxlargura para as imagens faciais de cada expressão. A Tabela 12 apresenta os números do gráfico em detalhes.

Figura 30 – Razão altura/largura da face de teste para cada grupo e discriminada por expressão.



Fonte: Elaborado pelo autor.

Tabela 12 – Razão altura/largura da face de teste para cada grupo e discriminada por expressão. A intensidade da coloração é proporcional a razões mais próximas de 1

	Grupos		
Expressão/Tarefa	**HC**	**ALS**	**PS**
BBP_NORMAL	1,107	1,102	1,138
DDK_PA	1,123	1,110	1,146
DDK_PATAKA	1,101	1,105	1,127
NSM_BIGSMILE	1,039	1,031	1,069
NSM_BLOW	1,134	1,113	1,190
NSM_BROW	1,062	1,096	1,118
NSM_KISS	1,067	1,073	1,111
NSM_OPEN	1,205	1,193	1,220
NSM_SPREAD	1,047	1,035	1,068

5.5 Modelos da Abordagem

O cenário apresentado na Seção 5.1 é ideal porque as subunidades foram detectadas anteriormente e apenas os modelos de localização de pontos das subunidades são necessários. No entanto, em um cenário real, o conhecimento prévio é apenas a caixa delimitadora da face. Nesse caso, o modelo de detecção de subunidades faciais é empregado para localizar o centro de cada subunidade que serve para inicializar o modelo de localização de pontos correspondente.

A Figura 31 (a) mostra os resultados do modelo M_0 para reduzir o viés algorítmico entre populações clínicas em um cenário real onde somente as faces foram

detectadas. Comparado ao cenário ideal, há uma clara perda de desempenho geral dessa abordagem multi-nível. No entanto, em comparação com os modelos do estado da arte *FAN-2D*, *ADNet*, e *SPIGA* (Figura 21) a abordagem pode reduzir de maneira mais significativa as diferenças de desempenho entre os grupos *HC* e *ALS*. Este modelo teve um desempenho igual nesses grupos, enquanto os resultados do *PS* ainda são visivelmente piores.

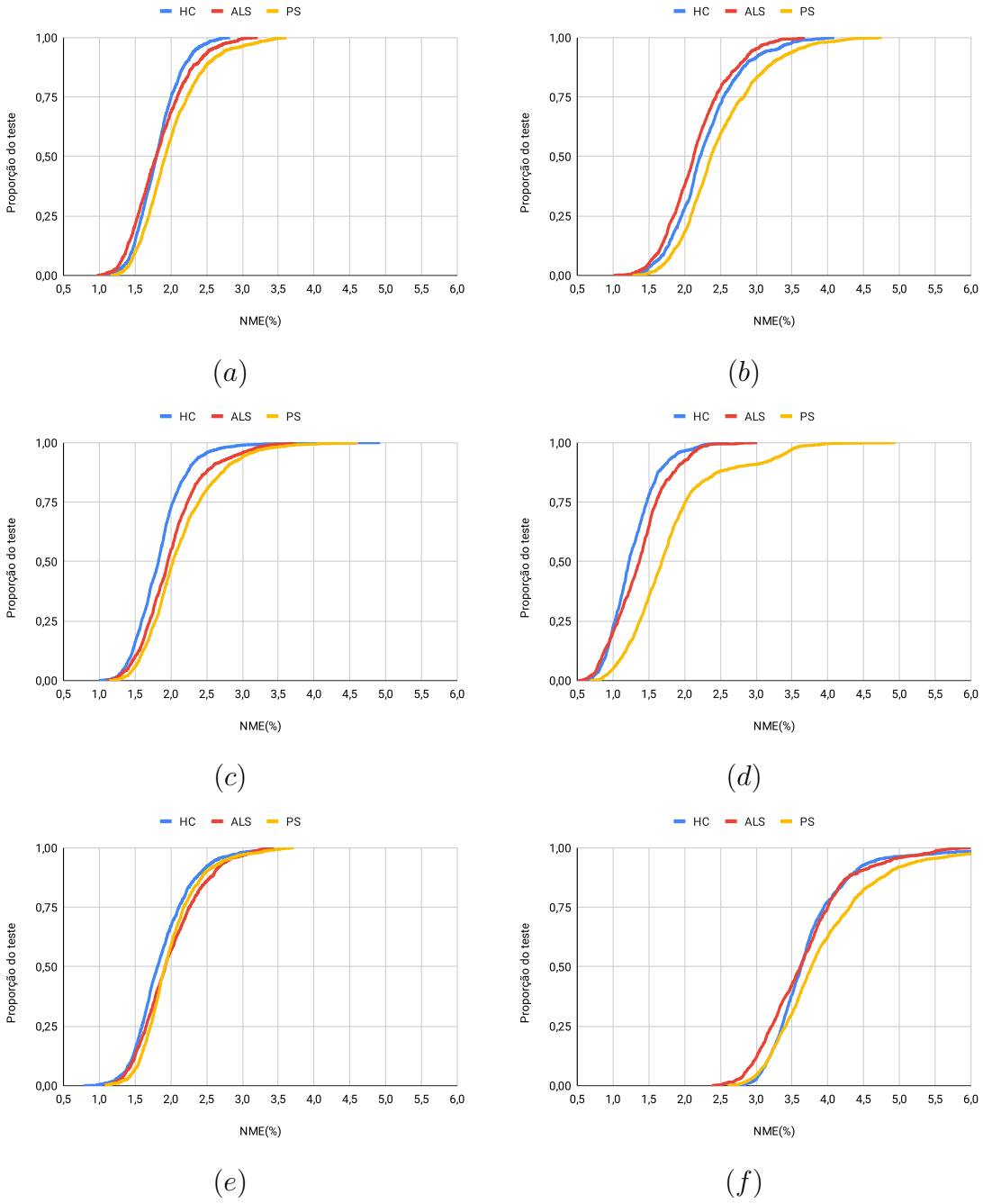
Em relação aos modelos nos quais a configuração da região da boca é divida em duas partes, também houve redução nas diferenças de desempenho entre grupos. O modelo M_{S7h}, que divide os pontos da boca em porções horizontais, obteve resultados muito próximos para os grupos *ALS* e *PS*. Contudo, o grupo *HC* continuou privilegiado em relação aos outros dois. O modelo M_{S7v}, que divide a modelagem da boca verticalmente, também obteve bons resultados na redução do viés. Nesse caso, o desempenho M_{S7v} é semelhante ao M_0, pois aproximou os resultados dos grupos *HC* e *ALS*. Entretanto, falhou em apresentar resultados para o grupo PS mais próximos dos outros dois.

As comparações do modelo em cenário ideal e dos modelos M_0, M_{S7h}, e M_{S7v} mostram que a metodologia apresentada neste trabalho conseguiu superar o estado da arte do alinhamento facial com respeito à redução do viés no *dataset Toronto Neuroface* (Tabela 10). Contudo, a partir apenas desses resultados, essa redução poderia ser vista como fruto unicamente do desempenho do *backbone* da rede utilizada para as modelagens. A segunda hipótese apresentada neste trabalho (Capítulo 1) é que a divisão da modelagem facial fundamentada nos elementos que compõem a face reduz as diferenças de performance entre grupos. Para afastar a ideia de que outros aspectos possam ser mais decisivos para essa redução, o modelo M_{68} foi elaborado a partir da mesma estrutura de rede, com os mesmos conjuntos e parâmetros para treino que os outros modelos gerados. A única diferença entre eles é a nível de configuração de subunidades. Enquanto que os modelos M_0, M_{S7h}, e M_{S7v} foram concebidos por configurações que levaram em conta a divisão da modelagem facial fundamentada em seus elementos, o M_{68} seguiu a ideia predominante de configuração observada na literatura do alinhamento facial. Ou seja, não há divisão da face em seus elementos, todos os pontos compõem uma única unidade. A Figura 31 (b) mostra que esse modelo, apesar de atingir desempenho geral razoável, mantém as diferenças de resultados entre todos os grupos avaliados, o que aponta para um viés algorítmico.

Em outro cenário de teste, o modelo M_0 deste trabalho foi aprimorado com inclusão 200 imagens (um pouco menos que 10% sobre o conjunto de treinamento inicial[2]) do conjunto de dados *Toronto Neuroface*. Para fazer uma comparação justa,

[2] O trabalho de Bandini et al. (2021) utilizou proporção similar do *Toronto Neuroface* ao realizar um ajuste fino na *FAN-2D*.

Figura 31 – *CEDs* dos modelos gerados a partir do *backbone* da Figura 14. (a) M_0. (b) M_{68}. (c) M_0 sem rotação. (d) M_0 + *Toronto Neuroface*. (e) M_{S7h}. (f) M_{S7v}



Fonte: Elaborado pelo autor.

cada novo modelo foi treinado com um subconjunto aleatório de dois grupos, deixando para teste somente o restante. A Figura 31 (d) mostra melhores resultados gerais, mas ao custo de um desempenho mais desigual comparando cada grupo.

Em certo casos, a análise dos gráficos de erro cumulativo é insuficiente para caracterizar as supostas diferenças de desempenho apresentadas pelos modelos avaliados. Mesmo quando as diferenças entre as curvas são aparentes nos gráfico de *CED*,

pode não ser possível afirmar, acima de dúvida, que os resultados são estatisticamente distintos, pois a escala do gráfico pode induzir ao erro. Assim, os testes estatísticos são necessários para comprovar se há, ou não, diferenças significativas de desempenho. A última parte da Tabela 10 exibe os resultados dos testes estatísticos realizados para verificar a existência de diferenças significativas no desempenho dos modelos avaliados. O primeiro teste realizado – Kruskal Wallis – identifica se os resultados dos modelos nos três grupos avaliados são originários da mesma distribuição, isto é, se não há diferenças significativas nos desempenhos. Quando a hipótese nula é rejeitada, ou seja, existem desempenhos significativamente distintos, aplica-se o teste de Dunn para identificar em quais grupos se localiza a diferença.

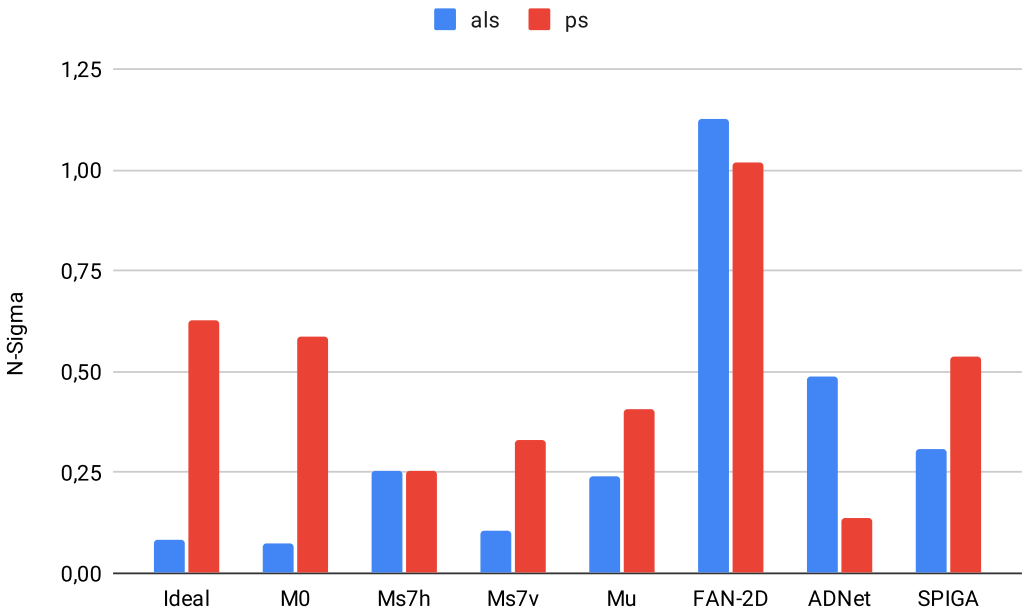
É possível perceber que a *FAN-2D* de fato apresentou diferenças significativas de desempenho que privilegiam o grupo de controle, o que pode caracterizar um viés desse modelo na direção do grupo de adultos saudáveis. Contudo, os resultados que ela apresentou para os grupos clínicos foram bastante próximos. O teste de múltiplas comparações de Dunn indica que ela desempenhou de forma igualitária nos grupos *ALS* e *PS*. No caso dos modelos avaliados no cenário ideal, houve a redução da diferença de desempenho entre os grupos *HC* e *ALS* a níveis insignificantes. Os modelos M_0 e M_{S7v} também foram capazes de reduzir as diferenças entre *HC* e *ALS* ao mesmo patamar. O modelo M_{S7h} atuou de maneira muito semelhante à *FAN-2D* e igualou os desempenhos nos grupos *HC* e *ALS*. Contudo, as diferenças entre os desempenhos nos grupos *HC* e *PS* persistiram para todos os modelos.

Os outros modelos do estado da arte avaliados – *ADNet* e *SPIGA* – não apresentaram resultados igualitários para nenhuma combinação de grupos. O resultado do valor p, calculado pelo teste de Dunn com os desempenhos da *ADNet* nos grupos *HC* e *PS*, indicam uma proximidade maior do que os outros modelos avaliados. Contudo, segundo o teste de Dunn, não se pode afirmar que houve redução do viés pois o p calculado é muito menor que o limiar de significância. O mesmo pode-se constatar dos desempenhos do *SPIGA* nos grupos *HC* e *ALS*. Considerando todos os modelos avaliados, o M_{68} foi o único que apresentou resultados muito significativamente diferentes para todas as combinações do teste de Dunn.

Após serem constatadas, através dos testes de Kruskall-Wallis e Dunn, as significativas diferenças entre os desempenhos, o método N-Sigma foi utilizado para quantificar as distâncias entre os resultados alcançados para os grupos. Para isso, escolheu-se o grupo de controle como referência, uma vez que ele é o grupo mais privilegiado pelos modelos de alinhamento facial. A Figura 32 mostra os resultados do método para os desempenhos de cada modelo utilizado. Nesse gráfico, quanto menor o tamanho da coluna (valor N-Sigma), mais próximo o conjunto correspondente está da referência. É possível estabelecer um paralelo entre os resultados do teste de Dunn

e o método N-Sigma. No gráfico, as três menores colunas relativas aos desempenhos do grupo *ALS* correspondem aos modelos cujas diferenças de resultados entre *HC* e *ALS* são insignificantes, apontando que os grupos estão bem próximos. A *FAN-2D* apresentou valores elevados para os grupos *ALS* e *PS* a partir da referência. Muito embora as diferenças de desempenho da *FAN-2D* nos grupos *ALS* e *PS* sejam consideradas insignificantes, ela apresentou resultados muito discrepantes desses dois grupos em relação ao controle. A interpretação do N-Sigma para esses resultados da *FAN-2D* indica que os resultados para os grupos *ALS* e *PS* estão a mais de um desvio padrão de distância dos resultados de *HC*. O valores do método N-Sigma calculados para os principais modelos apresentados neste trabalho estão na Tabela 13.

Figura 32 – Gráfico comparativo do valor N-Sigma calculado para os resultados dos modelos de localização de pontos usando o grupo de controle como referência.



Fonte: Elaborado pelo autor.

A Tabela 13 resume os resultados de algumas métricas utilizadas na avaliação dos desempenhos dos melhores modelos – considerando *NME*, *AUC*, e taxa de falha – nos grupos *HC*, *ALS*, e *PS*. É possível observar desempenho geral ligeiramente pior dos modelos gerados com a metodologia do trabalho em relação ao estado da arte. Contudo, os modelos de localização de pontos no cenário ideal ainda superam com margem consideravelmente alta todos os modelos avaliados, estabelecendo um patamar de desempenho a ser perseguido pelos modelos deste trabalho.

A Figura 33 ilustra alguns resultado qualitativos de quatro modelos: localização de pontos em cenário ideal; M_0; *FAN-2D*; e *SPIGA*. Essas amostras faciais correspon-

Tabela 13 – Resumo dos melhores desempenhos separados por grupos clínicos

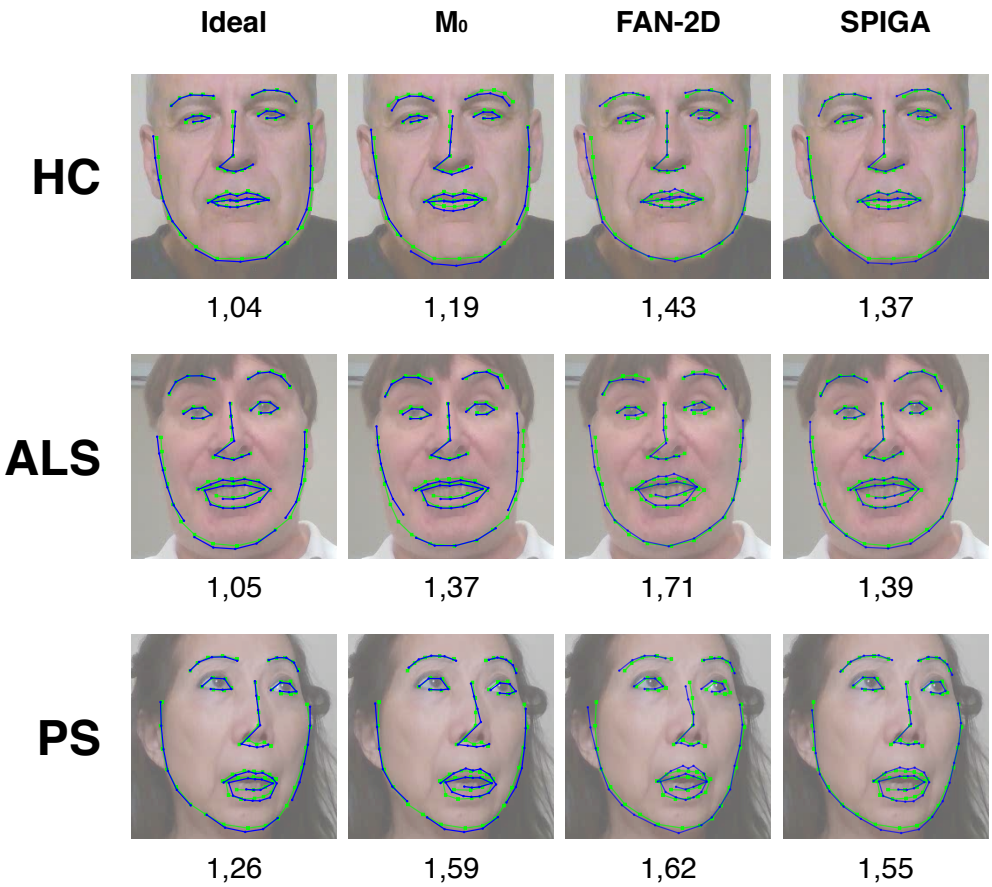
Métrica	**Modelo**	**Grupo**		
		HC	***ALS***	***PS***
NME%	Ideal	1,26	1,27	1,37
	M_0	1,82	1,84	1,99
	M_{S7h}	1,88	1,99	1,99
	FAN-2D	1,63	1,86	1,84
	ADNet	1,58	1,71	1,61
	SPIGA	1,50	1,58	1,64
$AUC_{2,5\%}$	Ideal	0,497	0,491	0,450
	M_0	0,275	0,271	0,219
	M_{S7h}	0,257	0,222	0,220
	FAN-2D	0,349	0,266	0,268
	ADNet	0,369	0,318	0,355
	SPIGA	0,401	0,370	0,345
$TF_{2,5\%}$	Ideal	0,0	0,0	0,0
	M_0	2,4	6,5	11,5
	M_{S7h}	7,8	13,9	9,8
	FAN-2D	0,0	9,7	4,1
	ADNet	0,0	0,6	0,0
	SPIGA	0,0	3,2	1,0
N-Sigma calculado usando *HC* como referência				
N-Sigma	Ideal	-	0,082	0,627
	M_0	-	0,074	0,584
	M_{S7h}	-	0,252	0,256
	FAN-2D	-	1,124	1,016
	ADNet	-	0,488	0,135
	SPIGA	-	0,306	0,538

dem a resultados com erro abaixo da média para cada um dos modelos. É possível observar que cada modelo apresenta uma deficiência específica, que se manifesta mesmo em exemplares nos quais eles desempenharam bem. O modelo M_0 apresentou um problema de continuidade da forma da mandíbula por conta da segmentação dessa unidade facial. A *FAN-2D* apresentou formas faciais mais irregulares, especialmente na região dos lábios. O *SPIGA* enfrentou um problema na localização vertical dos pontos laterais da mandíbula, principalmente em imagens com alguma expressão facial mais acentuada. Apesar dos problemas de cada modelo, a análise qualitativa dos resultados mostra que as estimativas são bastante próximas das anotações.

5.6 Análise de Resultado por Expressões Faciais

Além da rotulação por grupo clínico, o conjunto de dados *Toronto Neuroface* também é dividido em expressões faciais. Os participantes foram solicitados a realizar

Figura 33 – Amostras de resultados de quatro modelos avaliados para três indivíduos. O erro médio normalizado (*NME*%) se encontra abaixo da face. As marcações em azul representam as estimativas dos respectivos modelos.



Fonte: Elaborado pelo autor.

tarefas faciais, como sorrir ou levantar as sobrancelhas, geralmente empregadas para avaliar o nível de gravidade da doença considerando pacientes com ELA ou pós-derrame. As tarefas anotadas foram descritas na Seção 4.8 e detalhadas no trabalho de Bandini et al. (2021). A análise dos principais modelos gerados neste trabalho, assim como dos modelos do estado da arte, foi dividida pelas expressões faciais presentes no *dataset*. A Figura 34 ilustra as *CEDs* dos resultados de cada modelo para cada grupo e tarefa facial. Os *NMEs* dessa análise estão resumidos na Tabela 14. Os melhores resultados de cada modelo, para cada grupo e separado por expressão estão em fundo verde. Analogamente, os piores estão em vermelho.

Em geral, os modelos apresentaram desempenhos bastante similares em relação à maioria das tarefas faciais. Os desempenhos mais positivos dos modelos se manifestaram nos grupos de imagens relacionadas às tarefas NSM_KISS, NSM_BROW, NSM_BLOW, e NSM_BIGSMILE. Com exceção de NSM_BIGSMILE, essas tarefas não possuem grande impacto na forma facial, ou seja, durante a execução da expressão

Tabela 14 – Desempenho dos principais modelos separados por grupo e tarefa facial (*NME*%)

		Expressão / Tarefa								
Grupo	**Modelo**	**DDK_PA**	**NSM_KISS**	**NSM_BROW**	**DDK_PATAKA**	**NSM_SPREAD**	**NSM_BIGSMILE**	**BBP_NORMAL**	**NSM_OPEN**	**NSM_BLOW**
HC	Ideal	1,25	1,19	**1,13**	1,27	1,27	1,21	1,22	**1,42**	1,25
	M0	1,81	**1,71**	**1,71**	1,85	1,84	1,73	1,77	**2,10**	1,74
	FAN-2D	1,64	**1,53**	**1,75**	1,64	1,68	1,66	1,61	1,69	1,60
	ADNet	1,59	1,52	1,49	1,57	1,69	**1,35**	1,54	**1,72**	1,60
	SPIGA	1,49	1,44	1,40	1,55	1,48	**1,37**	1,42	**1,70**	1,55
ALS	Ideal	1,30	1,18	**1,15**	1,25	1,27	1,26	1,25	**1,43**	1,25
	M0	1,89	**1,69**	1,97	1,85	1,83	1,88	1,92	**2,06**	1,76
	FAN-2D	**2,04**	**1,69**	1,94	2,00	1,89	1,82	1,80	1,84	1,75
	ADNet	**1,80**	**1,62**	1,76	1,78	1,77	1,54	**1,62**	1,74	1,73
	SPIGA	**1,74**	**1,45**	1,68	1,69	1,53	1,47	1,56	1,64	1,49
PS	Ideal	1,37	1,33	**1,17**	1,42	1,32	1,32	1,42	**1,51**	1,34
	M0	1,95	1,95	2,07	1,99	1,92	**1,79**	2,07	**2,27**	1,91
	FAN-2D	1,83	1,82	1,80	1,86	1,83	**1,75**	1,87	**1,94**	1,76
	ADNet	1,62	1,57	1,55	1,60	1,67	1,66	1,66	**1,83**	**1,53**
	SPIGA	1,65	1,63	1,47	1,64	1,56	**1,45**	1,69	**1,86**	1,62

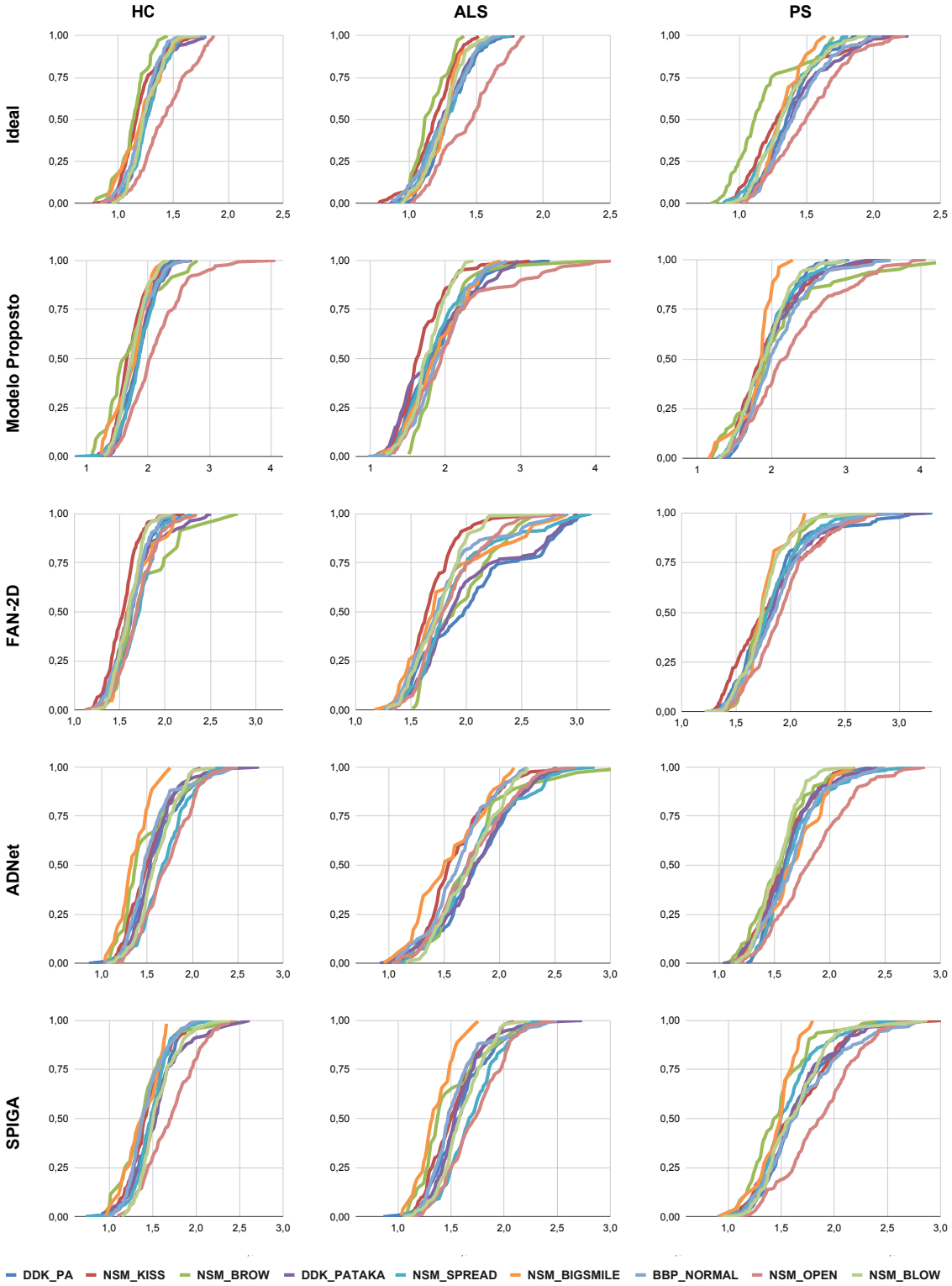


Figura 34 – *CED* dos principais modelos aplicados ao *Toronto Neuroface* separados por grupo clínico e discriminados por expressão/tarefa presente no *dataset*.

a face permanece razoavelmente neutra comparadas às outras tarefas presentes no *dataset*.

As expressões faciais que apresentaram maiores desafios aos modelos en-

volvem movimentos mais acentuados com a boca ou a mandíbula. Esse aspecto foi negativo para todos os modelos avaliados, inclusive do estado da arte. É importante notar que o cálculo do desempenho é realizado com base em todos os pontos faciais de maneira indistinta. Isto é, todos os pontos que compõem o modelo contribuem igualmente para a composição do resultado. Nesse contexto deve-se observar com atenção a composição do padrão de anotação facial. A subunidade facial com maior densidade de pontos corresponde à região dos lábios. Assim, uma detecção incorreta da boca possui impacto maior no desempenho do modelo que uma detecção incorreta das sobrancelhas.

É possível observar que os modelos avaliados, em especial o modelo M_0, obtiveram bons resultados quando avaliados sob a tarefa NSM_BROW que impacta na posição natural das sobrancelhas. Isso ocorre a despeito dos resultados abaixo da média que os modelos deste trabalho apresentaram para as subunidades que modelam os pontos das sobrancelhas (Tabela 8). A subunidade que representa a boca, por outro lado, foi responsável pelos piores resultados apresentados em todos os modelos quando avaliadas as imagens da tarefa NSM_OPEN. Isso pode ser explicado parcialmente pela falta de imagens de treinamento onde esse tipo de expressão facial está presente. Adicionalmente é preciso estar atento às observações do final da Seção 5.4 que dizem respeito ao impacto que a forma facial de entrada pode ter na detecção do nível superior. Contudo, é importante notar que essa expressão foi desafiadora para todos os modelos, incluindo modelos do estado da arte. A Figura 35 ilustra alguns dos piores resultados para a localização dos pontos da boca estimados pelo modelo M_0 em imagens relativas à tarefa NSM_OPEN. Em contraponto, a Figura 36 mostra alguns dos melhores resultados obtidos pelo modelo em imagens neutras, nesse caso, relativas à tarefa NSM_KISS.

5.7 Desempenho Geral

O desempenho geral no problema de alinhamento facial não é o foco principal deste trabalho. Contudo, é importante que a abordagem seja capaz de atingir o objetivo traçado sem que haja perda significativa no desempenho geral. Assim, foram compilados os resultados de cada modelo com base na análise de todos os grupos agregados. A Figura 37 mostra as *CEDs* do modelos gerados pela metodologia deste trabalho. Os desempenhos gerais dos modelos M_0 e M_{S7h} se mostraram bastante próximos, enquanto que os modelos M_{S7v} e M_{68} desempenharam claramente pior, apresentando resultados insatisfatórios perante o limite estabelecido pelos modelos de localização de pontos no cenário ideal.

A Figura 38 mostra compilados os resultados gerais: dos modelos em cenário ideal; do modelo M_0, que apresentou o menor *NME* para os modelos gerados a

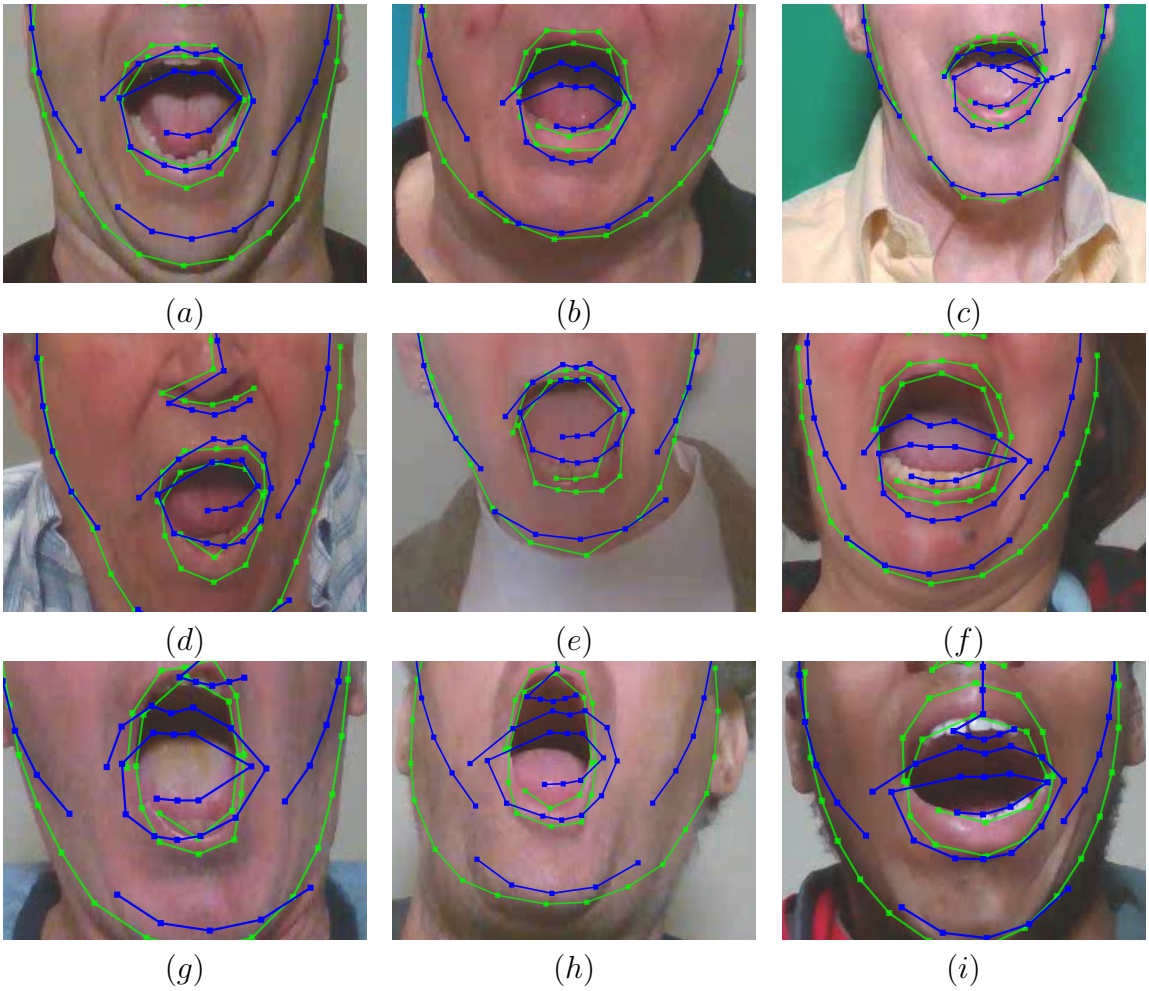


Figura 35 – Amostras dos piores resultados no *dataset Toronto Neuroface* com indivíduos realizando a expressão NSM_OPEN. Os resultados do modelo M_0 estão em azul. (a, b, c) *HC*. (d, e, f) *ALS*. (g, h, i) *PS*.

partir da metodologia apresentada; e dos modelos do estado da arte. O modelo M_0 tem desempenho geral ligeiramente pior quando comparado à *FAN-2D*, *ADNet*, e *SPIGA*. Entretanto, os modelos de localização de pontos em cenário ideal se mostram claramente superiores ao estado da arte. O objetivo da medição dos resultados gerais dos modelos em cenário ideal não busca estabelecer um comparativo direto, visto que tais modelos se encontram em vantagem competitiva (detecção ideal de subunidades). Porém, esses resultados estabelecem um patamar de desempenho a ser alcançado pelos modelos gerados a partir da metodologia. Isso pode indicar um trabalho futuro de aperfeiçoamento nos modelos de detecção de subunidades. A Tabela 15 resume os desempenhos gerais atingidos pelos principais modelos gerados pela metodologia e os modelos do estado da arte.

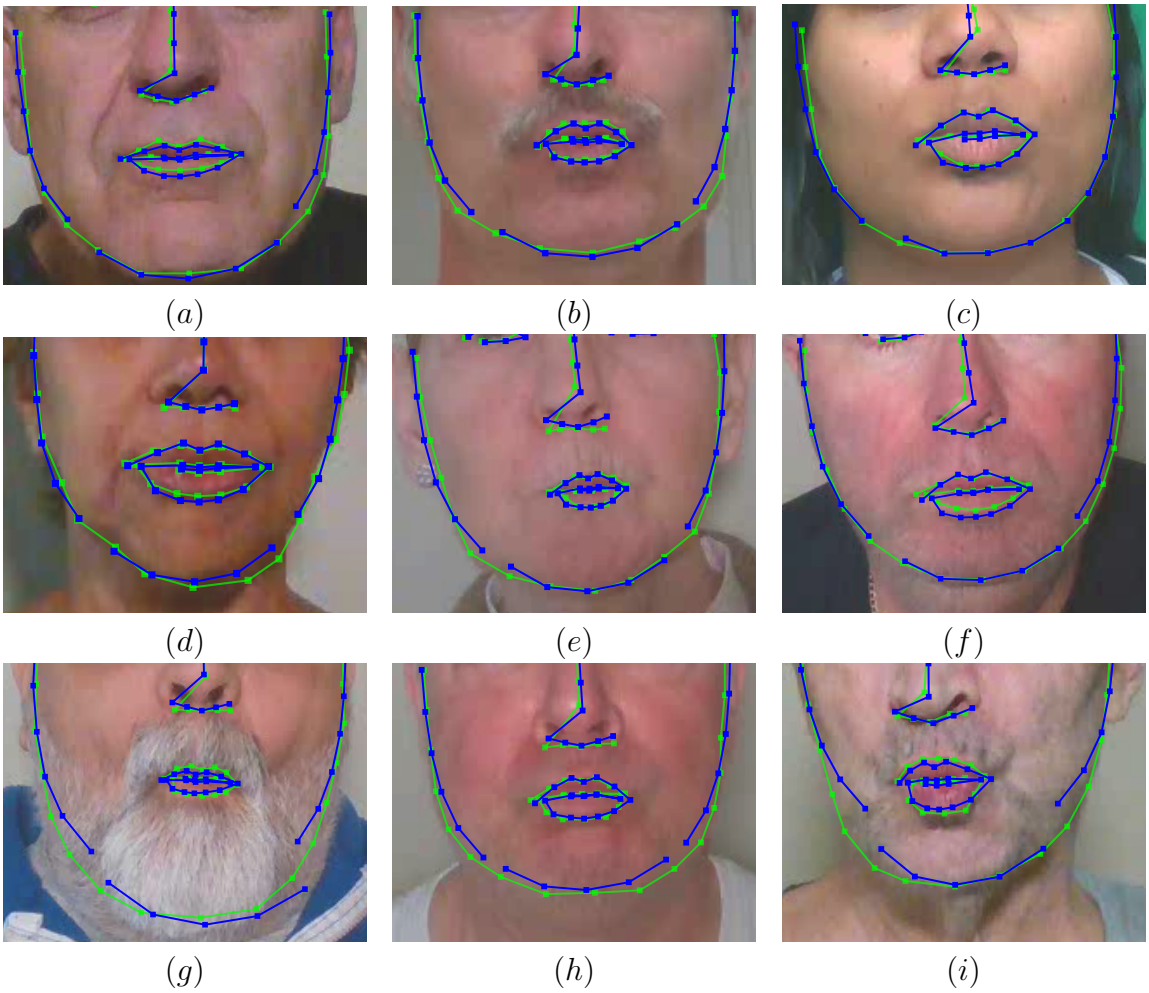


Figura 36 – Amostras dos melhores resultados no *dataset Toronto Neuroface* com indivíduos realizando a expressão NSM_KISS. Os resultados do modelo M_0 estão em azul. (a, b, c) *HC*. (d, e, f) *ALS*. (g, h, i) *PS*.

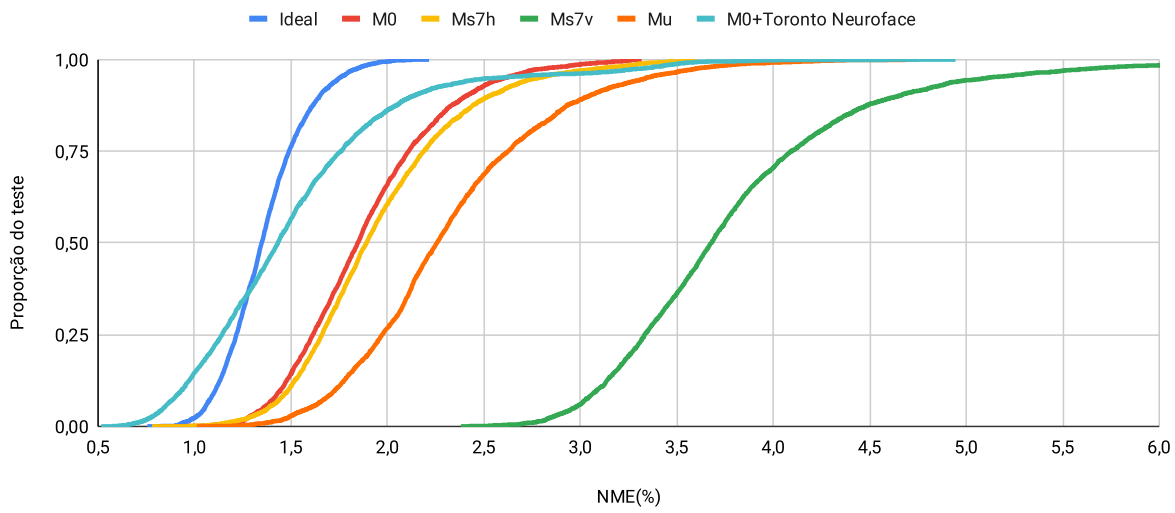
Tabela 15 – Desempenhos gerais dos principais modelos avaliados

Modelo	*NME*%	σ	*AUC*2%	*AUC*2,5%	*AUC*3%	TF2%	TF2,5%	TF3%
Ideal	1,36	0,21	0,319	0,456	0,546	0,5	0,0	0,0
M_0	1,89	0,39	0,110	0,252	0,371	34,0	7,1	1,3
M_{S7h}	1,96	0,43	0,096	0,232	0,350	39,3	10,6	3,0
FAN-2D	1,78	0,32	0,141	0,295	0,409	18,7	3,9	0,0
ADNet	1,64	0,30	0,192	0,345	0,453	12,8	0,5	0,0
SPIGA	1,58	0,33	0,224	0,369	0,474	12,0	1,1	0,0

5.8 Considerações Finais

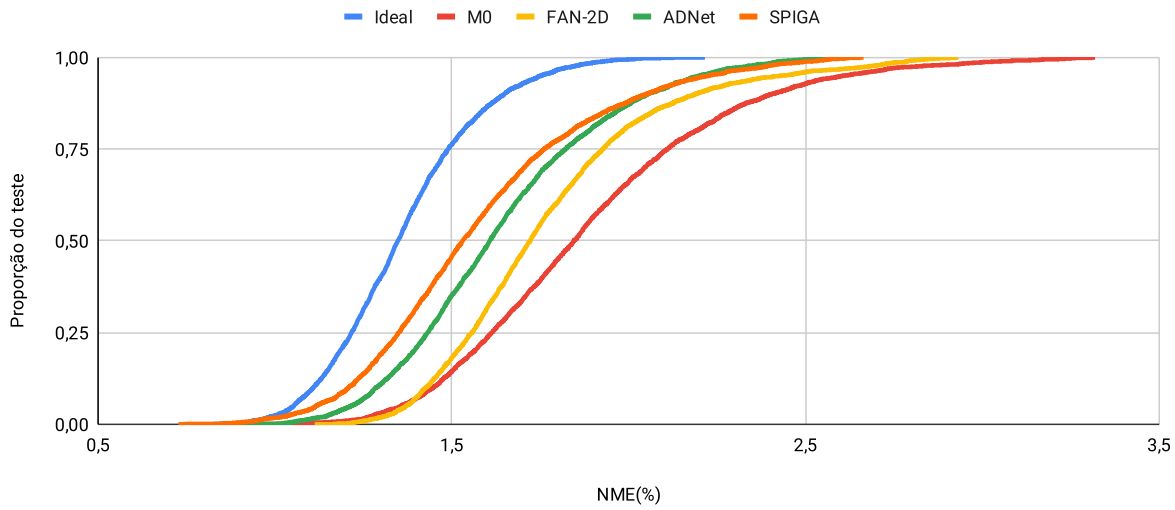
Este capítulo apresentou os resultados observados com a aplicação dos modelos gerados a partir da metodologia. Foi possível observar que a ideia de divisão da análise facial em subunidades se mostrou eficaz para a solução do problema apresentado. Além disso, a discriminação da análise de resultados com base nas expressões

Figura 37 – *CEDs* dos resultados gerais dos modelos da metodologia.



Fonte: Elaborado pelo autor.

Figura 38 – *CEDs* dos resultados gerais dos melhores modelos da metodologia e do estado da arte.



Fonte: Elaborado pelo autor.

faciais e nas subunidades apontou para problemas que podem ser corrigidos futuramente. Por fim, a comparação com o estado da arte do alinhamento facial mostrou que a abordagem desenvolvida nesta pesquisa, além de diminuir as diferenças de desempenho entre grupos, mantém o resultado geral em níveis comparáveis. No próximo capítulo é feita uma breve retomada da abordagem, são apontados problemas a serem resolvidos, e elencados caminhos a serem seguidos a partir da metodologia e das análises estabelecidas.

6 Conclusão

Este trabalho apresentou uma abordagem multi-nível para o alinhamento facial que visa à redução do viés algorítmico quando aplicada em populações clínicas. O primeiro passo para justificar essa abordagem foi a caracterização do viés dos modelos do estado da arte de alinhamento facial. Isso foi buscado a partir de um *dataset* balanceado com indivíduos de populações clínicas, o *Toronto Neuroface*. Foi exposto que os modelos *ADNet* e *SPIGA* apresentaram significativas diferenças de desempenho entre todos os grupos avaliados. A *FAN-2D* não apresentou desempenho que fosse capaz de aproximar os resultados de alinhamento facial obtidos nos grupos clínicos aos resultados do grupo de controle. De fato, ela apresentou diferenças de desempenho entre os grupos *ALS* e *PS* a níveis insignificantes. Por outro lado, os resultados desse modelo para o grupo *HC* mostraram disparidades ainda maiores que os outros modelos avaliados.

Para a construção da metodologia, levou-se em consideração as análises apontadas por Bandini et al. (2021). Em seu trabalho, eles observaram que o problema da assimetria orofacial amplifica as diferenças de desempenho entre os grupos. Isso indica que o problema do viés algorítmico nesse contexto pode estar mais relacionado a uma questão de morfologia do que propriamente de textura. Assim, foram desenvolvidos três modelos preditivos: um para uniformização da entrada; e dois para localização de coordenadas que operam em diferentes níveis no sentido de atenuar as diferenças de desempenho apresentadas por conta da forma facial:

- Modelo de uniformização facial que tem como objetivo padronizar a imagem facial com relação a rotação;
- Modelo de detecção de subunidades faciais que determina as coordenadas de cada elemento facial;
- Modelos de localização de pontos faciais que realiza a detecção dos pontos que compõem cada subunidade.

Os modelos foram integrados em um único método que determina as coordenadas de um conjunto de pontos faciais com base em uma imagem facial de entrada. Esse método foi comparado com o estado da arte do alinhamento facial sob o critério de redução do viés algorítmico já observado em grupos clínicos. Os resultados obtidos indicaram que a abordagem apresentada foi capaz de reduzir as diferenças de desempenho entre os grupos clínicos avaliados. Em relação aos grupos *HC* e *ALS*, as diferenças foram reduzidas a níveis insignificantes. No caso do grupo *PS*, as diferenças

ainda são consideráveis, porém, os desempenhos apresentados são mais próximos quando comparados ao estado da arte.

6.1 Análise das Hipóteses

Para demonstrar que o método de alinhamento facial apresentado consegue reduzir o viés algorítmico, foram formuladas três hipóteses: 1) a padronização das imagens por rotação contribui na redução das diferenças de desempenho entre os grupos; 2) a divisão da modelagem facial em elementos faciais reduz as diferenças de desempenho entre os grupos; 3) a adição de amostras do *dataset Toronto Neuroface* reduz as diferenças de desempenho entre os grupos. A primeira hipótese foi confirmada pelos resultados apresentados na Seção 5.3 onde foi realizada uma comparação do melhor modelo gerado, utilizando a padronização de rotação, com outra versão dele próprio sem a padronização. A versão que não faz ajuste de rotação foi incapaz de reduzir as diferenças de desempenho a níveis insignificantes considerando todas as combinações dos grupos. Os resultados foram avaliados a partir de testes de hipótese adequados. A segunda hipótese também foi confirmada de maneira semelhante. Um modelo foi treinado utilizando as mesmas características do melhor modelo gerado pela abordagem, ou seja, utilizou-se o mesmo *backbone*, mesmo conjunto de treino, e mesmas configurações de treinamento. Esse modelo, além de ter um desempenho geral significativamente pior, apresentou resultados bastante distintos para o alinhamento facial dos grupos clínicos.

A terceira hipótese visou investigar a redução do viés por meio da adição de representatividade no modelo gerado a partir da metodologia. Uma vez que o conjunto de treinamento utilizado não possui informação de indivíduos com problemas neurológicos (indicando um viés de representatividade), foram retiradas amostras aleatórias do *dataset Toronto Neuroface* para treinamento com o modelo M_0. Cada novo modelo foi treinado com faces de indivíduos de dois grupos e testado no grupo restante. Esperava-se uma redução nos desempenhos entre os grupos, contudo, observou-se que as discrepâncias aumentaram entre todos eles, muito embora o desempenho geral tenha melhorado significativamente. Isso indica que a hipótese pode ser rejeitada, contudo, é possível que outras formas de adição de representatividade consigam atingir melhores resultados.

6.2 Trabalhos Futuros e Problemas Observados

Os experimentos realizados na pesquisa mostraram algumas deficiências dos modelos gerados com a metodologia apresentada. Considerando que essa metodologia se mostrou eficaz em reduzir o viés algorítmico entre grupos distintos no problema

do alinhamento facial, é importante que os trabalhos futuros observem inicialmente os principais caminhos para aprimoramento da mesma. A seguir, são elencados os principais e mais imediatos aspectos que devem fazer parte da investigação do viés algorítmico sob a ótica da metodologia apresentada neste trabalho:

- **Configuração de subunidades faciais** - Nesta pesquisa foram apresentadas três diferentes configurações para as subunidades faciais que foram elaboradas com o objetivo de ampliar a capacidade de expressão do modelo. Uma vez que a metodologia é flexível para novas implementações nesse aspecto, um dos caminhos mais imediatos para futuras pesquisas concentra-se na proposta e experimentação de novas formas de configuração. Pode-se buscar dividir a forma facial em ainda mais elementos, ou realizar diferentes combinações de agrupamentos com duas ou mais subunidades. Outra possibilidade, porém mais complexa, é a elaboração de um método analítico para concepção de novas configurações que levasse em conta o conjunto de formas faciais observadas nos *datasets*. Isso pode ser guiado por um método de agrupamento aliado a uma análise Procrustes das formas faciais;

- **Adição de representatividade no treino** - Um dos caminhos de investigação seguidos nesta pesquisa levou à tentativa de contornar o problema do viés observado nos desempenhos dos modelos com a adição de elementos do *dataset* alvo do trabalho. O objetivo foi tentar atenuar o viés de representatividade presente nos *datasets* de treinamento. Contudo, a adição experimentada não surtiu efeito na redução do viés, apenas os resultados gerais foram afetados positivamente. Muito embora essa tentativa não tenha sido bem sucedida, essa questão deve ser melhor investigada. Neste trabalho, foram utilizadas amostras aleatórias dos grupos clínicos para aprimoramento dos modelos. Outro caminho imediato para trabalhos futuros seria investigar novas formas de aprimoramento dos *datasets* de treino com amostras do teste, principalmente levando em consideração que o *Toronto Neuroface* apresentou resultados distintos a depender da expressão facial avaliada.

- **Experimentação de outros *backbones*** - Todos os modelos gerados pela metodologia apresentada nesta pesquisa enquadram-se na categoria de modelo de regressão de coordenadas. Foi demonstrado que o modelo de localização de pontos faciais reduziu consideravelmente o viés algorítmico entre os grupos apresentados. Em relação ao desempenho geral, esse modelo superou o estado da arte do alinhamento facial quando operou em cenário ideal, estabelecendo um limite a ser buscado pela metodologia. Isso mostrou que a arquitetura usada nesses modelos é competitiva nos dois aspectos. Contudo, o modelo de detecção

de subunidades com a mesma estrutura de regressão de coordenadas não conseguiu sustentar desempenho semelhante, e ficou aquém do limite estabelecido. É importante notar que a detecção de subunidades envolve uma quantidade maior de atributos em comparação com a localização dos pontos em elementos faciais. Um caminho promissor para o aprimoramento dos modelos é a utilização de outros *backbones* na etapa de detecção de subunidades, especialmente arquiteturas de regressão de *heatmaps*. Uma melhor detecção de subunidades aliada aos modelos de localização de pontos dos elementos faciais, que já obtiveram resultados superiores ao estado da arte, tem grandes possibilidades de elevar os resultados dos modelos, tanto na redução do viés quanto no desempenho geral.

Além desses aspectos principais, os experimentos realizados lançaram luz sobre diversos aspectos do *dataset Toronto Neuroface* e dos métodos do estado da arte, que também devem servir de guia para continuidade das investigações sobre o viés algorítmico no alinhamento facial:

- a exemplo do que foi observado por Bandini et al. (2021), as expressões faciais amplificam as dificuldades enfrentadas pelos modelos, especialmente em indivíduos com problemas neurológicos. Esse problema mostrou-se desafiador tanto para os modelos gerados pela abordagem desenvolvida, quanto para os modelos do estado da arte;
- um outro problema, que está diretamente vinculado às expressões, é a localização de pontos de referência que são afetados espacialmente na imagem pelas próprias expressões faciais. Pontos das sobrancelhas e do queixo apresentaram uma maior dificuldade na localização;
- observou-se correlação leve/moderada entre os resultados de detecção de subunidades e a razão altura/largura da face de entrada. Seria interessante verificar se uma estratégia de *padding* para redução de atributos do fundo na imagem de entrada tem algum efeito no desempenho, especialmente para imagens com maior relação altura/largura. Outra ideia para esse problema seria aplicar uma transformação de escala não-uniforme na entrada deformando a face para eliminar o fundo;
- o modelo M_0 apresentou problemas relacionados à continuidade da forma da mandíbula por conta da divisão do contorno facial na configuração adotada. Seria interessante introduzir restrições de forma entre as subunidades de tal modo a conferir maior coesão no resultado final;
- a metodologia realiza a detecção das subunidades com base em textura global, contudo a localização dos pontos utiliza amostras de textura do elemento facial.

Em caso de oclusão do elemento, o modelo fica bastante prejudicado. Esse tipo de desafio não está presente no *Toronto Neuroface*, contudo, é importante ser observado em trabalhos futuros, uma vez que a oclusão é um aspecto comum nos *datasets* de alinhamento facial, podendo impactar também na convergência dos modelos em treinamento.

6.3 Contribuições e Publicações

A pesquisa realizada teve como principal contribuição o desenvolvimento de um novo método de alinhamento facial com foco na redução das diferenças de desempenho entre grupos clínicos. Ademais, os experimentos realizados mostraram alguns caminhos e entendimentos que devem ser explorados em trabalhos futuros para aperfeiçoamento do método com o objetivo de reduzir ainda mais as diferenças observadas.

Alguns dos conceitos elaborados e explorados durante o desenvolvimento desta pesquisa foram publicados na forma de artigos originais em conferência ou periódico. Esses artigos estão listados a seguir:

- ***Improving Active Shape Models robustness towards locating facial landmarks in profile contour***.
  International Conference on Systems, Signals and Image Processing, 2020;
- ***A CNN-based multi-level face alignment approach for mitigating demographic bias in clinical populations***.
  Computational Statistics, 2023.

Referências

BANDINI, A.; REZAEI, S.; GUARÍN, D. L.; KULKARNI, M.; LIM, D.; BOULOS, M. I.; ZINMAN, L.; YUNUSOVA, Y.; TAATI, B. A new dataset for facial motion analysis in individuals with neurological disorders. **IEEE Journal of Biomedical and Health Informatics**, v. 25, n. 4, p. 1111–1119, 2021.

BELHUMEUR, P. N.; JACOBS, D. W.; KRIEGMAN, D. J.; KUMAR, N. Localizing parts of faces using a consensus of exemplars. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 35, n. 12, p. 2930–2940, 2013.

BULAT, A.; SANCHEZ, E.; TZIMIROPOULOS, G. Subpixel heatmap regression for facial landmark localization. In: **32nd British Machine Vision Conference 2021**. BMVA Press, 2021. Disponível em: <https://www.bmvc2021-virtualconference.com/assets/papers/1405.pdf>.

BULAT, A.; TZIMIROPOULOS, G. Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources. In: **2017 IEEE International Conference on Computer Vision (ICCV)**. [S.l.: s.n.], 2017. p. 3726–3734.

BULAT, A.; TZIMIROPOULOS, G. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In: **2017 IEEE International Conference on Computer Vision (ICCV)**. [S.l.: s.n.], 2017. p. 1021–1030.

BUOLAMWINI, J.; GEBRU, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In: FRIEDLER, S. A.; WILSON, C. (Ed.). **Proceedings of the 1st Conference on Fairness, Accountability and Transparency**. New York, NY, USA: PMLR, 2018. (Proceedings of Machine Learning Research, v. 81), p. 77–91. Disponível em: <http://proceedings.mlr.press/v81/buolamwini18a.html>.

BURGOS-ARTIZZU, X.; PERONA, P.; DOLLAR, P. **Caltech Occluded Faces in the Wild (COFW)**. [S.l.]: CaltechDATA, 2022.

CAO, D.; CHEN, C.; PICCIRILLI, M.; ADJEROH, D.; BOURLAI, T.; ROSS, A. Can facial metrology predict gender? In: **2011 International Joint Conference on Biometrics (IJCB)**. [S.l.: s.n.], 2011. p. 1–8.

CITRON, D.; PASQUALE, F. The scored society: Due process for automated predictions. **Washington Law Review**, v. 89, p. 1–33, 2014.

COOTES, T.; TAYLOR, C.; COOPER, D.; GRAHAM, J. Active shape models-their training and application. **Computer Vision and Image Understanding**, v. 61, n. 1, p. 38 – 59, 1995. ISSN 1077-3142.

COOTES, T. F.; TAYLOR, C. J. Active shape models - 'smart snakes'. In: **Proceedings of the British Machine Vision Conference**. [S.l.]: BMVA Press, 1992. p. 28.1–28.10. ISBN 3-540-19777-X.

COOTES, T. F.; TAYLOR, C. J. Statistical models of appearance for medical image analysis and computer vision. In: SONKA, M.; HANSON, K. M. (Ed.). **Medical Imaging 2001: Image Processing**. SPIE, 2001. v. 4322, p. 236 – 248. Disponível em: <https://doi.org/10.1117/12.431093>.

COOTES, T. F.; TAYLOR, C. J.; LANITIS, A. Active shape models: Evaluation of a multi-resolution method for improving image search. In: **Proceedings of the British Machine Vision Conference**. [S.l.]: BMVA Press, 1994. p. 32.1–32.10. ISBN 952-1898-1-X.

DEALCALA, D.; SERNA, I.; MORALES, A.; FIERREZ, J.; ORTEGA-GARCIA, J. Measuring bias in ai models: An statistical approach introducing n-sigma. In: **2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC)**. [S.l.: s.n.], 2023. p. 1167–1172.

DROZDOWSKI, P.; RATHGEB, C.; DANTCHEVA, A.; DAMER, N.; BUSCH, C. Demographic bias in biometrics: A survey on an emerging challenge. **IEEE Transactions on Technology and Society**, v. 1, n. 2, p. 89–103, 2020.

DU, C.; WU, Q.; YANG, J.; WU, Z. Svm based asm for facial landmarks location. In: **2008 8th IEEE International Conference on Computer and Information Technology**. [S.l.: s.n.], 2008. p. 321–326.

DUNN, O. J. Multiple comparisons among means. **Journal of the American Statistical Association**, American Statistical Association, Taylor & Francis, Ltd., v. 56, n. 293, p. 52–64, 1961. ISSN 01621459. Disponível em: <http://www.jstor.org/stable/2282330>.

GARCIA, R. V.; WANDZIK, L.; GRABNER, L.; KRUEGER, J. The harms of demographic bias in deep face recognition research. In: **2019 International Conference on Biometrics (ICB)**. [S.l.: s.n.], 2019. p. 1–6. ISSN 2376-4201.

GEORGOPOULOS, M.; OLDFIELD, J.; NICOLAOU, M. A.; PANAGAKIS, Y.; PANTIC, M. Mitigating demographic bias in facial datasets with style-based multi-attribute transfer. **International Journal of Computer Vision**, v. 129, n. 7, p. 2288–2307, 2021.

GEORGOPOULOS, M.; PANAGAKIS, Y.; PANTIC, M. Investigating bias in deep face analysis: The kanface dataset and empirical study. **Image and Vision Computing**, v. 102, p. 103954, 2020.

GOGIĆ, I.; AHLBERG, J.; PANDŽIĆ, I. S. Regression-based methods for face alignment: A survey. **Signal Processing**, v. 178, p. 107755, 2021. ISSN 0165-1684. Disponível em: <https://www.sciencedirect.com/science/article/pii/S016516842030298X>.

GONZALEZ, R.; WOODS, R. **Digital Image Processing, Global Edition**. Pearson Education, 2018. ISBN 9781292223070. Disponível em: <https://books.google.com.br/books?id=P8AoEAAAQBAJ>.

GONZALEZ, R. C.; WOODS, R. E. **Digital Image Processing**. Pearson, 2018. ISBN 9780133356724. Disponível em: <https://books.google.com.br/books?id=0F05vgAACAAJ>.

GROSS, R.; MATTHEWS, I.; COHN, J.; KANADE, T.; BAKER, S. Multi-pie. **Image and Vision Computing**, v. 28, n. 5, p. 807–813, 2010. ISSN 0262-8856. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0262885609001711>.

GUO, Y.; ZHANG, L.; HU, Y.; HE, X.; GAO, J. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In: LEIBE, B.; MATAS, J.; SEBE, N.; WELLING, M. (Ed.). **Computer Vision – ECCV 2016**. Cham: Springer International Publishing, 2016. p. 87–102. ISBN 978-3-319-46487-9.

HAYKIN, S. S. **Neural networks and learning machines**. Third. Upper Saddle River, NJ: Pearson Education, 2009.

HE, K.; ZHANG, X.; REN, S.; SUN, J. Deep residual learning for image recognition. In: **2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. [S.l.: s.n.], 2016. p. 770–778.

HSU, C.-F.; LIN, C.-C.; HUNG, T.-Y.; LEI, C.-L.; CHEN, K.-T. **A Detailed Look At CNN-based Approaches In Facial Landmark Detection**. 2020.

HUANG, X.; DENG, W.; SHEN, H.; ZHANG, X.; YE, J. Propagationnet: Propagate points to curve to learn structure information. In: **2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**. Los Alamitos, CA, USA: IEEE Computer Society, 2020. p. 7263–7272. Disponível em: <https://doi.ieeecomputersociety.org/10.1109/CVPR42600.2020.00729>.

HUANG, Y.; YANG, H.; LI, C.; KIM, J.; WEI, F. Adnet: Leveraging error-bias towards normal direction in face alignment. In: **2021 IEEE/CVF International Conference on Computer Vision (ICCV)**. [S.l.: s.n.], 2021. p. 3060–3070.

HUPONT, I.; FERNÁNDEZ, C. Demogpairs: Quantifying the impact of demographic imbalance in deep face recognition. In: **2019 14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019)**. [S.l.: s.n.], 2019. p. 1–7.

JESORSKY, O.; KIRCHBERG, K. J.; FRISCHHOLZ, R. W. Robust face detection using the hausdorff distance. In: BIGUN, J.; SMERALDI, F. (Ed.). **Audio- and Video-Based Biometric Person Authentication**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2001. p. 90–95. ISBN 978-3-540-45344-4.

JIN, H.; LIAO, S.; SHAO, L. Pixel-in-pixel net: Towards efficient facial landmark detection in the wild. **International Journal of Computer Vision**, v. 129, n. 12, p. 3174–3194, 2021.

JIN, X.; TAN, X. Face alignment in-the-wild: A survey. **Computer Vision and Image Understanding**, v. 162, p. 1–22, 2017. ISSN 1077-3142. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1077314217301455>.

JOHNSTON, B.; CHAZAL, P. A review of image-based automatic facial landmark identification techniques. **EURASIP Journal on Image and Video Processing**, v. 2018, p. 86, 2018.

KHABARLAK, K.; KORIASHKINA, L. Fast facial landmark detection and applications: A survey. **Journal of Computer Science and Technology**, v. 22, n. 1, p. e02, 2022. Disponível em: <https://journal.info.unlp.edu.ar/JCST/article/view/1972>.

KRUSKAL, W. H.; WALLIS, W. A. Use of ranks in one-criterion variance analysis. **Journal of the American Statistical Association**, Taylor & Francis, v. 47, n. 260, p. 583–621, 1952. Disponível em: <https://www.tandfonline.com/doi/abs/10.1080/01621459.1952.10483441>.

KUMAR, A.; MARKS, T. K.; MOU, W.; WANG, Y.; JONES, M.; CHERIAN, A.; KOIKE-AKINO, T.; LIU, X.; FENG, C. Luvli face alignment: Estimating landmarks' location, uncertainty, and visibility likelihood. In: **IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**. [S.l.: s.n.], 2020.

KÖSTINGER, M.; WOHLHART, P.; ROTH, P. M.; BISCHOF, H. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In: **2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)**. [S.l.: s.n.], 2011. p. 2144–2151.

LE, V.; BRANDT, J.; LIN, Z.; BOURDEV, L.; HUANG, T. S. Interactive facial feature localization. In: FITZGIBBON, A.; LAZEBNIK, S.; PERONA, P.; SATO, Y.; SCHMID, C. (Ed.). **Computer Vision – ECCV 2012**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012. p. 679–692. ISBN 978-3-642-33712-3.

LECUN, Y.; BOSER, B.; DENKER, J. S.; HENDERSON, D.; HOWARD, R. E.; HUBBARD, W.; JACKEL, L. D. Backpropagation applied to handwritten zip code recognition. **Neural Computation**, v. 1, n. 4, p. 541–551, 1989.

LECUN, Y.; BOTTOU, L.; BENGIO, Y.; HAFFNER, P. Gradient-based learning applied to document recognition. **Proceedings of the IEEE**, v. 86, n. 11, p. 2278–2324, 1998.

MA, N.; ZHANG, X.; ZHENG, H.-T.; SUN, J. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In: FERRARI, V.; HEBERT, M.; SMINCHISESCU, C.; WEISS, Y. (Ed.). **Computer Vision – ECCV 2018**. Cham: Springer International Publishing, 2018. p. 122–138. ISBN 978-3-030-01264-9.

MENEZES, H. F.; FERREIRA, A. S. C.; PEREIRA, E. T.; GOMES, H. M. Bias and fairness in face detection. In: **2021 34th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)**. [S.l.: s.n.], 2021. p. 247–254.

MILBORROW, S. **Locating Facial Features with Active Shape Models**. Dissertação (Mestrado) — Faculty of Engineering, University of Cape Town, 2007.

MOGHADDAM, B.; YANG, M.-H. Gender classification with support vector machines. In: **Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)**. [S.l.: s.n.], 2000. p. 306–311.

MORETTIN, P. A.; SINGER, J. d. M. **Estatística e Ciência de Dados**. [S.l.]: LTC, 2022.

PANTIC, M.; ROTHKRANTZ, L. J. M. Automatic analysis of facial expressions: the state of the art. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 22, n. 12, p. 1424–1445, 2000. ISSN 0162-8828.

PRADOS-TORREBLANCA, A.; BUENAPOSADA, J. M.; BAUMELA, L. Shape preserving facial landmarks with graph attention networks. In: **33rd British Machine Vision Conference 2022**. BMVA Press, 2022. Disponível em: <https://bmvc2022.mpi-inf.mpg.de/0155.pdf>.

PRATT, W. K. **Digital Image Processing: PIKS Scientific Inside**. USA: Wiley-Interscience, 2007. ISBN 0471767778.

SAGONAS, C.; ANTONAKOS, E.; TZIMIROPOULOS, G.; ZAFEIRIOU, S.; PANTIC, M. 300 faces in-the-wild challenge: database and results. **Image and Vision Computing**, v. 47, p. 3–18, 2016. ISSN 0262-8856. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0262885616000147>.

SAGONAS, C.; TZIMIROPOULOS, G.; ZAFEIRIOU, S.; PANTIC, M. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In: **2013 IEEE International Conference on Computer Vision Workshops**. [S.l.: s.n.], 2013. p. 397–403.

SESHADRI, K.; SAVVIDES, M. Robust modified active shape model for automatic facial landmark annotation of frontal faces. In: **IEEE 3rd International Conference on Biometrics: Theory, Applications, and Systems (BTAS '09)**. [S.l.: s.n.], 2009. p. 1–8.

SUN, Y.; WANG, X.; TANG, X. Deep convolutional network cascade for facial point detection. In: **2013 IEEE Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2013. p. 3476–3483.

SURESH, H.; GUTTAG, J. A framework for understanding sources of harm throughout the machine learning life cycle. In: **Equity and Access in Algorithms, Mechanisms, and Optimization**. New York, NY, USA: Association for Computing Machinery, 2021. (EAAMO '21). ISBN 9781450385534. Disponível em: <https://doi.org/10.1145/3465416.3483305>.

TAATI, B.; ZHAO, S.; ASHRAF, A. B.; ASGARIAN, A.; BROWNE, M. E.; PRKACHIN, K. M.; MIHAILIDIS, A.; HADJISTAVROPOULOS, T. Algorithmic bias in clinical populations—evaluating and improving facial analysis technology in older adults with dementia. **IEEE Access**, v. 7, p. 25527–25534, 2019.

TANG, Z.; PENG, X.; LI, K.; METAXAS, D. N. Towards efficient u-nets: A coupled and quantized approach. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 42, n. 8, p. 2038–2050, 2020.

VENKATESAN, R.; LI, B. **Convolutional Neural Networks in Visual Computing: A Concise Guide**. CRC Press, 2018. (Convolutional Neural Networks in Visual Computing: A Concise Guide). ISBN 9781138747951. Disponível em: <https://books.google.com.br/books?id=Y2xSAQAACAAJ>.

WANG, C. The development and challenges of face alignment algorithms. **Journal of Physics: Conference Series**, IOP Publishing, v. 1335, n. 1, p. 012009, 2019. Disponível em: <https://doi.org/10.1088/1742-6596/1335/1/012009>.

WANG, M.; DENG, W.; HU, J.; TAO, X.; HUANG, Y. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In: **2019 IEEE/CVF International Conference on Computer Vision (ICCV)**. [S.l.: s.n.], 2019. p. 692–702.

WANG, N.; GAO, X.; TAO, D.; YANG, H.; LI, X. Facial feature point detection: A comprehensive survey. **Neurocomputing**, v. 275, p. 50–65, 2018. ISSN 0925-2312. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0925231217308202>.

WANG, X.; BO, L.; FUXIN, L. Adaptive wing loss for robust face alignment via heatmap regression. In: **2019 IEEE/CVF International Conference on Computer Vision (ICCV)**.

Los Alamitos, CA, USA: IEEE Computer Society, 2019. p. 6970–6980. Disponível em: <https://doi.ieeecomputersociety.org/10.1109/ICCV.2019.00707>.

WU, W.; PROTOPAPAS, P.; YANG, Z.; MICHALATOS, P. Gender classification and bias mitigation in facial images. In: **12th ACM Conference on Web Science**. New York, NY, USA: Association for Computing Machinery, 2020. (WebSci '20), p. 106–114.

WU, W.; QIAN, C.; YANG, S.; WANG, Q.; CAI, Y.; ZHOU, Q. Look at boundary: A boundary-aware face alignment algorithm. In: **2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2018. p. 2129–2138.

XU, Z.; LI, B.; GENG, M.; YUAN, Y. Anchorface: An anchor-based facial landmark detector across large poses. In: **Proceedings of the AAAI Conference on Artificial Intelligence, 35**. [S.l.: s.n.], 2021. (AAAI Technical Track on Computer Vision III, v. 4), p. 3092–3100.

YANG, J.; LIU, Q.; ZHANG, K. Stacked hourglass network for robust facial landmark localisation. In: **2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)**. [S.l.: s.n.], 2017. p. 2025–2033. ISSN 2160-7516.

ZAFEIRIOU, S.; ZHANG, C.; ZHANG, Z. A survey on face detection in the wild: Past, present and future. **Computer Vision and Image Understanding**, v. 138, p. 1–24, 2015. ISSN 1077-3142. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1077314215000727>.

ZHU, X.; RAMANAN, D. Face detection, pose estimation, and landmark localization in the wild. In: **2012 IEEE Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2012. p. 2879–2886.