UNIVERSIDADE DE SÃO PAULO
FACULDADE DE FILOSOFIA, CIÊNCIAS E LETRAS DE RIBEIRÃO PRETO
DEPARTAMENTO DE COMPUTAÇÃO E MATEMÁTICA

LEANDRO PERSONA

Reconhecimento de emoções por meio da geometria facial com coordenadas normalizadas dos *landmarks*

Ribeirão Preto-SP

2022

LEANDRO PERSONA

Reconhecimento de emoções por meio da geometria facial com coordenadas normalizadas dos *landmarks*

Versão Original

Dissertação apresentada à Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto (FFCLRP) da Universidade de São Paulo (USP), como parte das exigências para a obtenção do título de Mestre em Ciências.

Área de Concentração: Computação Aplicada.

Orientador.....: Professora Doutora Alessandra Alaniz Macedo
Coorientador..: Professor Doutor Fernando Meloni

Ribeirão Preto-SP

2022

Persona, Leandro

Reconhecimento de emoções por meio da geometria facial com coordenadas normalizadas dos *landmarks*. Ribeirão Preto-SP, 2022.
102p. : il.; 30 cm.

Dissertação apresentada à Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto da USP, como parte das exigências para a obtenção do título de Mestre em Ciências,

Área: Computação Aplicada.

Orientador: Professora Doutora Alessandra Alaniz Macedo

1. Aprendizado de Máquina. 2. Tecnologias Assistivas. 3. Reconhecimento de emoções. 4. Padrões faciais.

Leandro Persona

Reconhecimento de emoções por meio da geometria facial com coordenadas normalizadas dos *landmarks*

Modelo canônico de trabalho monográfico acadêmico em conformidade com as normas ABNT.

Trabalho aprovado. Ribeirão Preto-SP, 18 de Abril de 2022:

Professora Doutora Alessandra Alaniz Macedo - FFCLRP / USP

Professor Doutor José Francisco de Magalhães Netto - UFAM

Professor Doutor Marcelo Garcia Manzatto - ICMC / USP

Ribeirão Preto-SP
2022

Agradecimentos

Primeiramente a Deus, por estar sempre guiando os meus passos.

Aos meu pais, por todo o carinho, amor e preocupação ao longo da minha vida. Meu pai, como maior exemplo de honestidade, ética e caráter. Minha mãe, como exemplo de proteção, força e determinação.

Aos meus irmãos, mesmo distantes por consequências da vida, sempre me ensinaram a não desistir dos meus objetivos. Em especial para a minha irmã, pela sua valiosa contribuição no início da escrita deste trabalho, por suas dicas e contribuições.

Á minha professora orientadora, Alessandra. Obrigado por acreditar no meu sonho e permitir que eu pudesse chegar ao final de mais essa etapa da minha vida. Pela maneira tão clara e simples de demonstrar qual a essência de ser um pesquisador. Obrigado pela orientação e ensinamentos que levarei por toda a minha vida.

Ao meu coorientador, Fernando. Obrigado por todas as incansáveis vezes que guiou o meu trabalho. Por todos os ensinamentos, paciência e clareza das informações. Obrigado por sempre estar ao meu lado nas horas difíceis.

À Empresa Brasileira de Correios e Telégrafos, por permitir o início e conclusão deste trabalho.

Pesquisa desenvolvida com o auxílio dos recursos de HPC disponibilizados pela Superintendência de Tecnologia da Informação da Universidade de São Paulo.

Tudo parece ser impossível, até que seja feito.
Nelson Mandela

Resumo

PERSONA, L. **Reconhecimento de Emoções por meio da geometria facial com coordenadas normalizadas dos *landmarks*.** 2022. 99f. Dissertação (Mestrado) – Departamento de Computação e Matemática, FFCLRP, Universidade de São Paulo, Ribeirão Preto, 2022.

O reconhecimento de emoções é parte intrínseca das relações sociais humanas e está associado com comportamentos que resultam em padrões faciais distintos. Devido ao fato da expressão facial ser um indicador de contexto emocional, existe grande interesse científico, artístico, médico e comercial sobre o assunto e isso tem estimulado o desenvolvimento de técnicas e métodos computacionais de reconhecimento automático de emoções. Apesar dos métodos atuais apresentarem resultados satisfatórios, há ainda grandes desafios relacionados ao reconhecimento de emoções. Este trabalho apresenta um novo método, denominado REGL, cujo encadeamento de etapas tem o objetivo de aprimorar o reconhecimento de expressões faciais e emoções humanas. O método visa diminuir a variabilidade amostral, permitindo um melhor ajuste das informações que definem os padrões faciais. Dentre as técnicas exploradas, inclui-se a normalização dos pontos fiduciais faciais, chamados de *landmarks*, para a construção de classificadores destinados ao reconhecimento das emoções. Como resultado, obteve-se uma acurácia média de 90% com a utilização de algoritmos de Aprendizado de Máquina com diferentes arquiteturas. Esses resultados indicam que: (i) o método REGL é mais aprimorado do que os investigados em termos de taxa de acerto e (ii) produziu resultados mais resilientes, considerando menor dependência do conjunto de treino e da arquitetura do classificador. Destaca-se que o reconhecimento de emoções faciais é um passo importante em diferentes áreas, possibilitando o desenvolvimento de tecnologias assistivas mais robustas e permitindo a melhoria das técnicas de síntese computacional.

Palavras-chave: Aprendizado de Máquina. Tecnologias assistivas. Reconhecimento de emoções. Padrões faciais.

Abstract

PERSONA, L. **Recognition of Emotions through facial geometry with normalized landmarks coordinates.** 2022. 99f. Dissertation (Master's degree) – Department of Computing and Mathematics, FFCLRP, University of São Paulo, Ribeirão Preto, 2022.

Recognition of emotions is an intrinsic act of human social relationships, and it is associated with behaviors that result in distinct facial patterns. Considering that facial expression is an indicator of emotional context, there is scientific, artistic, medical, and marketing interest in the subject, which has stimulated the development of techniques and computational methods for automatic emotion recognition. Our work presents the method REGL. REGL is a sequence of steps to recognize facial expressions and human emotions in images. The method aims to reduce sample variability, allowing for a better adjustment of the information that defines facial patterns. Among the techniques, we included the normalization of facial fiducial points (landmarks). We also configured classifiers for the recognition of facial emotions. As a result, we obtained an average accuracy of 90% using Machine Learning algorithms with different architectures. These results demonstrate that: (i) the REGL method is an improved method in terms of hit rate, and (ii) it produced more resilient results, considering less dependence on the training set and the classifier architecture. It is noteworthy that the facial emotion recognition of facial emotions is an essential step in different areas, enabling the development of more robust assistive technologies and allowing the improvement of computational synthesis techniques.

Keywords: Machine Learning. Assistive technologies. Recognition of emotions. Facial Patterns.

Lista de figuras

Lista de tabelas

Lista de Algoritmos

Lista de abreviaturas e siglas

AEHZ	Algoritmo de equalização de histograma e zoom
AM	Aprendizado de máquina
CNN	Convolucional neural network (Rede neural convolucional)
FACS	*Facial Action Coding System*
FFCLRP	Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto
FN	False negative (Falso negativo)
FP	False positive (Falso positivo)
HOG	Histogram of oriented gradients (Histograma de gradientes orientados)
PDI	Processamento digital de imagens
REGL	Reconhecimento de emoções pela geometria dos *landmarks*
RGB	Red (Vermelho), Green (Verde) e Blue (Azul)
ROI	Region Of interest (Região de interesse)
SRE	Sistema de reconhecimento de emoções
TP	True positive (Verdadeiro positivo)
TN	True negative (Verdadeiro negativo)
USP	Universidade de São Paulo

Lista de símbolos

$\leftarrow$	Atribuição
$\equiv$	Equivalente
$\subset$	Está contido
$\rightarrow$	Implica, se..então
$\cap$	Intersecção
Δ	Letra grega maiúscula Delta
Ω	Letra grega maiúscula Ômega
Π	Letra grega maiúscula Pi
Θ	Letra grega maiúscula Teta
ξ	Letra grega maiúscula Xi
$\leq$	Menor ou igual
$\geq$	Maior ou igual
$\in$	Pertence
$\sqrt{x}$	Raiz quadrada de x
$\Leftrightarrow$	Se e somente se
$\cup$	União

Sumário

1

Introdução

O reconhecimento de emoções é parte intrínseca das relações humanas. Ainda na infância, as crianças aprendem a mapear e a interpretar as emoções das outras pessoas para utilizar essas informações como indicadores do contexto ao qual estão inseridas (DARWIN, 2013; HESS, 2001; LANGNER, 2010a). Este mapeamento é possível porque os seres humanos, quase sempre, traduzem suas emoções em movimentos físicos detectáveis, tais como as expressões faciais, que são fundamentais para as interações sociais. Expressões faciais são comportamentos ubíquos e dependem fracamente dos fatores culturais (EKMAN; FRIESEN, 1971). De acordo com (ALVAREZ, 2013), pode-se identificar sete tipos diferentes de emoções universais: medo, raiva, tristeza, felicidade, surpresa e nojo, além da emoção neutra. Como apresentam padrões marcantes e estáveis, as expressões faciais podem ser identificadas mesmo em pessoas desconhecidas. Por consequência, o reconhecimento pode ser sistematizado com grande potencial para automatização, de forma que máquinas podem ser empregadas na tentativa de interpretar as emoções humanas.

Nos últimos anos, os métodos de reconhecimento de emoções, a partir de expressões faciais, tiveram uma rápida evolução, devido ao crescente interesse científico, médico e comercial sobre o tema (CHENG; LING, 2008; JIA J; ZHANG; CAI, 2010; LAHIRI, 2011; XIE, 2015). Os métodos mais comuns de reconhecimento de pessoas e expressões faciais focam nos processos automatizados de detecção de padrões em imagens digitais. Por exemplo, ferramentas presentes em redes sociais, aparelhos móveis e máquinas fotográficas são capazes de identificar se uma pessoa está sorrindo ou não (CHAUGULE, 2016). Há também ferramentas assistivas que auxiliam portadores de síndromes comportamentais (e.g., autismo e distúrbio do humor) a identificar emoções em outras pessoas (PICARD, 2016; CHENG; LING, 2008; LAHIRI, 2011). Portanto, o reconhecimento automático de expressões faciais oferece novos meios para que humanos e máquinas interajam entre si, seja pelo uso de movimentos voluntários ou pelo reconhecimento de movimentos involuntários.

Em geral, a interpretação das informações em imagens envolve automatização via Aprendizado de Máquina (CHOLLET, 2017). Porém, a manipulação de imagens

exige etapas de pré-processamento inicial (VAILLANCOURT, 2010) para possibilitar a detecção de formas específicas, neste caso, uma face humana. Modelos previamente treinados realizam uma varredura em segmentos da imagem, usando probabilidades para confirmação de padrões (WU; JI, 2018). Alguns modelos são capazes de detectar a presença de faces humanas em imagens com altas taxas de acerto, acima de 90% (WU; JI, 2018; VIOLA; JONES, 2001). Assim que a identificação da face é confirmada, o próximo passo para avaliação das expressões faciais é mapear as Regiões de Interesse (em inglês *Regions of Interest* - ROI), estruturas específicas, tais como olhos, nariz, boca, queixo, etc. Em seguida, as posições relativas das estruturas recebem marcações, os *landmarks*, aos quais são atribuídas coordenadas bi ou tridimensionais, inserindo uma camada adicional de informações.

O uso de *landmarks* oferece uma maneira simples e objetiva de reconhecer as expressões faciais, por meio da comparação de padrões, para a identificação de pessoas (reconhecimento facial) e para alterações nos padrões faciais (reconhecimento de expressões e emoções). Embora as implementações sejam diferentes, ambas consideraram as coordenadas dos landmarks como variáveis do problema (valores de referência) para o cálculo de distâncias (e.g. euclidiana, cosseno, etc) (Li, 2012; GARRIDO; JOSHI, 2018; MEHTA, 2018). No caso das emoções, os movimentos musculares responsáveis por cada expressão facial produzem mudanças geométricas na posição relativa dos ROIs (XIE W.; SHEN; JIANG, 2017). Essas mudanças alteram as coordenadas dos *landmarks* e os valores das distâncias entre eles. Dado que uma face em repouso apresenta distâncias relativas diferentes de uma face assustada ou sorrindo, o conjunto de variações pode ser utilizado para identificar as expressões (MEHTA, 2018). A generalidade das respostas emocionais entre os seres humanos tende a simplificar o problema, permitindo identificar quais distâncias são mais afetadas por cada tipo de expressão facial. Ao final da análise dessas variações, classificadores de Aprendizado de Máquina podem ser treinados para o reconhecimento das emoções humanas (WU; JI, 2018).

Apesar dos métodos atuais apresentarem resultados satisfatórios para contextos específicos, como para fotos adquiridas em ambientes controlados, há ainda grandes desafios relacionados ao tema. O principal problema envolve a variabilidade de padrões encontrados em imagens obtidas em condições de campo (variáveis não-controladas como luz, brilho, distância do equipamento de captura, etc) (MEHTA, 2018) (TESTA, 2019). Por exemplo, diferenças entre a forma das faces tendem a agregar ruído ao problema, tornando o treino dependente de raça, idade, sexo, etc. Além disso, aspectos como distância focal, luminosidade, enquadramento, pose das pessoas (imagem da face rotacionada e expressão de sentimento) e de configurações de hardware afetam de forma significativa a concentração e a localização dos pixels, adicionando ruído aos dados. Essas fontes de variabilidade interferem negativamente na performance dos classificadores, dificultando o processo de Aprendizado de Máquina (CHOLLET, 2017) (MEHTA, 2018). Similar a outros problemas,

vislumbra-se que a diminuição da variabilidade nos dados pode contribuir para a melhoria da performance computacional. Contudo, o domínio das informações nas imagens são os pixels, de forma que as técnicas clássicas de compatibilização e de normalização dos dados necessitam ser adaptadas para este contexto, exigindo estratégias computacionalmente criativas e conceitualmente elaboradas.

De acordo com (FILHO OGE MARQUES; VIEIRA NETO, 1999; GONZALES; WOODS, 2008), um dos grandes desafios dos algoritmos de Processamento Digital de Imagens (PDI) em ambientes não controlados está relacionado às dificuldades de se obter bons resultados em diferentes condições de luminosidade e contraste, uma vez que essas condições produzem ruído e variabilidade nos dados. Além disso, o posicionamento relativo do objeto na cena tende a produzir variabilidade, reduzindo a eficiência dos algoritmos.

As ideias investigadas, criadas e descritas nesta dissertação foram discutidas no projeto SofiaFala (SOFIAFALA, 2019), apoiado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), por meio do processo 442533/2016-0, que busca o desenvolvimento de um aplicativo para dispositivos móveis que auxilia crianças, com Síndrome de Down, nos treinamentos de fonoaudiologia para o desenvolvimento e aprimoramento da fala.

1.1 Objetivo

Este trabalho apresenta um novo método de reconhecimento de emoções em imagens digitais, denominado REGL, que funciona sob diversas condições de variabilidade, entre elas: escala, rotação, questões raciais e de aquisição das imagens. O método emprega algumas técnicas já conhecidas de manipulação de imagens, como a equalização do histograma e o alinhamento facial, a fim de compatibilizar as informações, permitindo a adequada normalização. Em seguida, foram utilizados classificadores adequados à metodologia do trabalho, os quais permitiram avaliar o possível ganho de performance.

Os passos de processamento e de normalização visaram a redução da variabilidade nos dados e a simplificação do problema. Como hipótese, a redução da variabilidade tende a produzir classificadores com melhor performance, independente do método de aprendizado de máquina empregado. Assim, desenvolver métodos que compatibilizem informações de imagens, em tese, possibilita a obtenção de melhores resultados para a área como um todo.

Ao final das etapas que antecedem a utilização dos modelos preditivos, obteve-se classificadores que (i) foram mais eficientes do que os atuais em termos de taxas de acerto; (ii) produziram resultados mais resilientes (menor dependência do conjunto de treino) e (iii) produziram respostas qualitativas em tempo real. Argumenta-se que este trabalho

tem o potencial de aprimorar e expandir o uso do reconhecimento de padrões faciais para a inferência de emoções, auxiliando diversas áreas como o desenvolvimento de tecnologias assistivas.

Como resultado, espera-se em uma futura versão do aplicativo SofiaFala que as análises comparativas entre os diversos padrões faciais sejam indicadores responsivos para os desenvolvedores sobre a utilização do sistema.

1.2 Organização do documento

O conteúdo desta Dissertação está dividido em seis capítulos. As referências encontram-se nas páginas finais. A seguir, um resumo dos próximos capítulos:

- Capítulo 2: Revisão da literatura recente na área de processamento digital de imagens e as principais técnicas para detecção facial e extração dos *landmarks*.
- Capítulo 3: Descrição dos principais conceitos envolvendo o reconhecimento de padrões e as métricas utilizadas para avaliação dos resultados dos algoritmos de Aprendizado de Máquina utilizados neste trabalho.
- Capítulo 4: Apresentação do método REGL de reconhecimentos de emoções e os bancos de dados de expressões faciais utilizados neste trabalho.
- Capítulo 5: Os experimentos realizados e os resultados obtidos com o método REGL. Também são apresentados diversos dados estatísticos, além do tempo de processamento dos algoritmos. Esses resultados são analisados e discutidos.
- Capítulo 6: As conclusões obtidas com os experimentos realizados no Capítulo 5. Também são apresentadas as contribuições, as limitações e os trabalhos futuros.

2

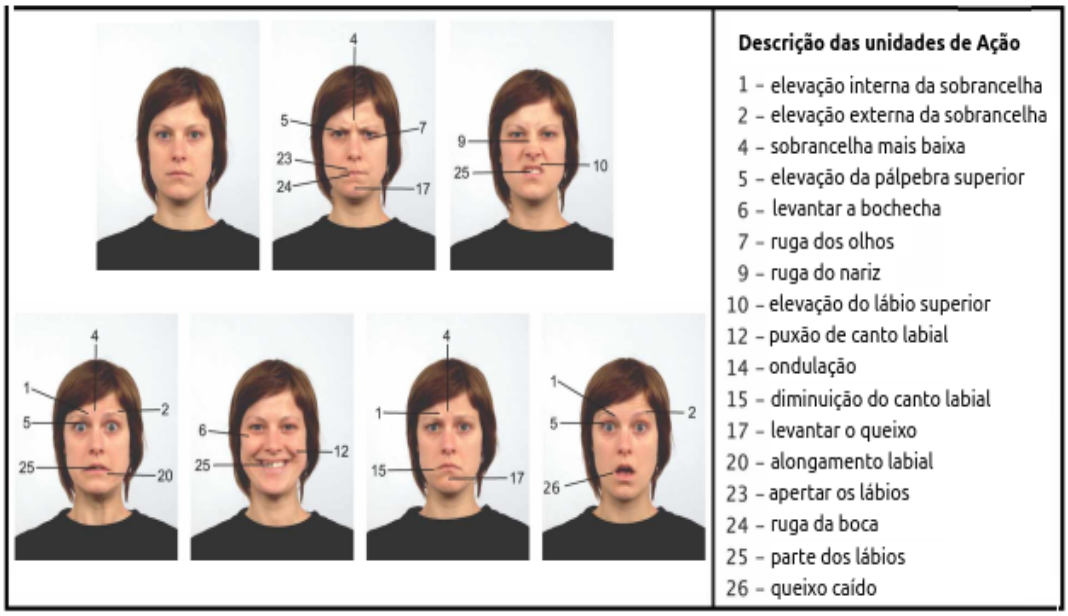
Técnicas de Aquisição e Processamento de Imagens para o Reconhecimento de Emoções

Na início da década de 70, Paul Ekman (EKMAN; FRIESEN, 1971) realizou um experimento científico que se tornou um marco no reconhecimento das emoções humanas. Na época, acreditava-se que as pessoas utilizavam seus músculos faciais de acordo com um conjunto de convenções sociais e expressões formadas pela convivência em sociedade, de forma similar ao que ocorre com os idiomas, diversificando cada região do planeta. Ekman registrou inúmeras imagens de homens e mulheres com diversas expressões faciais. Depois viajou para o Brasil, Argentina e Japão. Para a sua surpresa, as pessoas dos diferentes países que participaram dos experimentos, obtiveram os mesmos resultados na classificação das imagens. O experimento se estendeu para as florestas de Papua-Nova Guiné, na Oceania, para as vilas mais remotas e afastadas da civilização. Descobriu-se que, mesmo com os habitantes dessas regiões, os resultados não foram diferentes, concluindo que as emoções humanas, manifestadas na forma de expressões faciais são universais e independentes de fatores étnicos e sociais.

Outra contribuição importante do trabalho de Ekman foi a criação do Facial Action Coding System (FACS), um sistema para classificar emoções humanas. É um padrão para categorizar sistematicamente a expressão física das emoções, permitindo rotular qualquer expressão facial anatomicamente possível.

A Figura 1 apresenta as seis principais emoções, juntamente com a expressão neutra, subdivididas em unidades de ação, que são os elementos formadores dos FACS e suas as principais diferenças.

Figura 1 – Divisão por unidade de ação dos FACS



Fonte: traduzido de (LANGNER, 2010b)

Com os crescentes avanços no início deste século, nas áreas de robótica e automatização, a procura por sistemas de reconhecimento de expressões faciais e emoções tem se tornado cada vez maior. Os seres humanos são responsivos aos estados emocionais uns dos outros e a automatização de processos criou a expectativa de que computadores e máquinas também adquiram essa habilidade. Com os avanços no estudo da interação homem-máquina, pesquisadores têm conseguido melhorar significativamente essa interação com o uso de sensores (ZHANG, 2012). Consoles de videogames, como o Kinect da Microsoft, podem detectar o movimento humano e reagir de acordo com este, conectando o mundo físico com o mundo virtual. Sensores de detecção de expressões faciais em automóveis podem identificar quando um motorista está sonolento e agir para reduzir o risco de acidentes (CHELLAPPA, 2018). Para o desenvolvimento destes sistemas artificiais, é necessário realizar a transição do mundo físico e contínuo para o mundo computacional e discreto, cujos dados possam ser armazenados e processados em computadores.

Segundo (VAILLANCOURT, 2010), as melhores práticas para o Aprendizado de Máquina podem ser divididas e ordenadas como: (a) adequação e normalização dos dados, (b) filtro das amostras, (c) seleção de variáveis, (d) treino do modelo preditivo (classificador) e (e) teste de validação. Apesar da relação direcional entre os passos, a implementação prática envolve recursividade do processo como um todo, e os passos podem ser revisitados múltiplas vezes.

No contexto de reconhecimento de padrões e classificação em imagens, os passos (a) e (b) são dificultados pela própria natureza da fonte das informações, uma vez que são

suscetíveis a ruídos adquiridos na aquisição ou manipulação das imagens. A concentração e a distribuição dos pixels em uma imagem dependem não apenas do tipo de objeto, mas também das suas características dimensionais, pela intensidade e direção da luz incidente, pela posição relativa do objeto em relação ao equipamento de captura (câmera) e pelas configurações de hardware utilizadas para a aquisição da imagem (GARRIDO; JOSHI, 2018; MEHTA, 2018). Portanto, a qualidade das informações contidas em imagens é afetada pela maneira com que a imagem é adquirida. Em ambientes não controlados, as amostras tendem a conter dados com grande variabilidade e ruído. Em termos estatísticos, a variabilidade excessiva dificulta a compatibilização das informações e a normalização das variáveis, convergindo em classificadores com baixa performance (RASCHKA, 2015).

Assim, a manipulação das informações e o filtro amostral tornam-se essenciais ainda na fase de análise das imagens, mesmo que o processamento dessas informações seja diferente dos vetores comuns, utilizados em outros contextos. Contudo, não há garantias sobre a compatibilidade das informações adquiridas em imagens diferentes, aumentando o nível de complexidade do processo. Esse fato tem estimulado o uso crescente de métodos de aprendizado profundo para gerar classificadores de imagens (LI; DENG, 2018). Esses métodos utilizam baixo nível de manipulação das informações, ao custo da exigência de um número massivo de amostras para o treinamento dos modelos preditivos. Apesar dos bons resultados alcançados por alguns modelos em contextos específicos, tais como o reconhecimento de lista de objetos, pedestres e faces humanas, os resultados tendem a ser muito dependentes do conjunto amostral (GARRIDO; JOSHI, 2018; MEHTA, 2018). Além disso, há pouco controle sobre os mecanismos de classificação, o que por vezes dificulta sua manipulação, escalabilidade e a generalização para outros contextos.

No contexto do reconhecimento de emoções humanas, os fundamentos teóricos sobre Processamento Digital de Imagens, detecção facial, extração de *landmarks* e frontalização são imprescindíveis para o entendimento do método REGL e são apresentados neste capítulo.

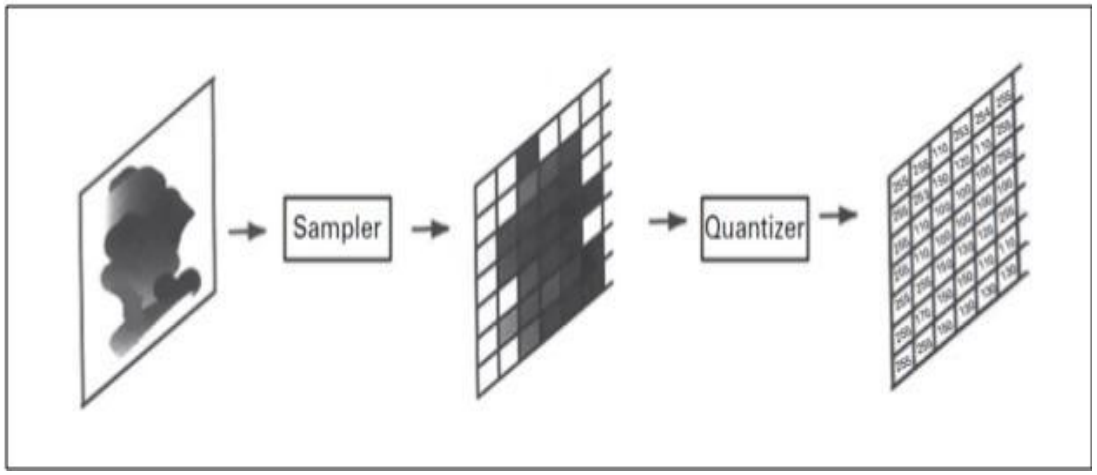
2.1 Imagens digitais

Em síntese, uma imagem digital representa a conversão do mundo real para o mundo artificial, digitalizado em coordenadas espaciais que representam a sua resolução em largura e altura. Em geral, Ω é um retângulo de lados m x n, discretizado com uma malha bidimensional regular. Desse modo, representamos uma imagem digital em escala de cinza, com profundidade de cor de 8 bits, por meio de uma matriz m x n de entradas $u_{ij} \in Z$, em que $u_{ij} = \mathrm{u}(x_i, y_j)$ para $(i, j) \in \mathrm{I} = \{1, 2, 3, 4, 5, 6, 7, ..., m\}$ x $\{1, 2, 3, 4, 5, 6, 7, ..., n\}$ e $0 \leq u(i, j) \leq 255$. Já uma imagem colorida é composta por outras matrizes correspondentes aos respectivos canais de cor. Cada elemento da matriz é chamado de pixel. Uma

imagem bidimensional pode ser definida como uma função $u : \Omega \subset R^2 \rightarrow R^c$, em que $c = 1$ para imagens em escala de cinzas e $c = 3$ para imagens coloridas. De acordo com (BURGER; BURGE, 2008) a transformação radiométrica de uma imagem colorida possibilita a normalização do brilho e do contraste em apenas um canal de cor, em escala de cinza, por meio da média aritmética dos valores dos canais R (vermelho), G (verde) e B (azul), quando o sistema de cor representado for o RGB. Para uma imagem em escala de cinza bidimensional, os valores de $u(x, y)$ correspondem à intensidade dos pixels nas coordenadas $(x, y) \in \Omega$. Entretanto, em uma imagem colorida, tem-se $u(x, y) = (r \cup (x, y), g \cup (x, y), b \cup (x, y))$, em que $r(u), g(u)$ e $b(u)$ representam a intensidade dos canais vermelho, verde e azul, respectivamente. Note que $r(u), g(u)$ e $b(u)$ também são imagens em escala de cinza, uma vez que $r(u), g(u), b(u) : \Omega \subset R^2 \rightarrow R$.

A Figura 2 ilustra a aquisição de uma imagem digital por meio do processo de amostragem (*sampling*), que cria a matriz de resolução (largura x altura), e do processo de quantização (*quantization*), que discretiza a cor de cada pixel da imagem em um valor numérico inteiro de 8 bits, ou seja, estabelece a sua intensidade de brilho.

Figura 2 – Aquisição de uma imagem digital: amostragem e quantização



Fonte: adaptado de (GONZALES; WOODS, 2008)

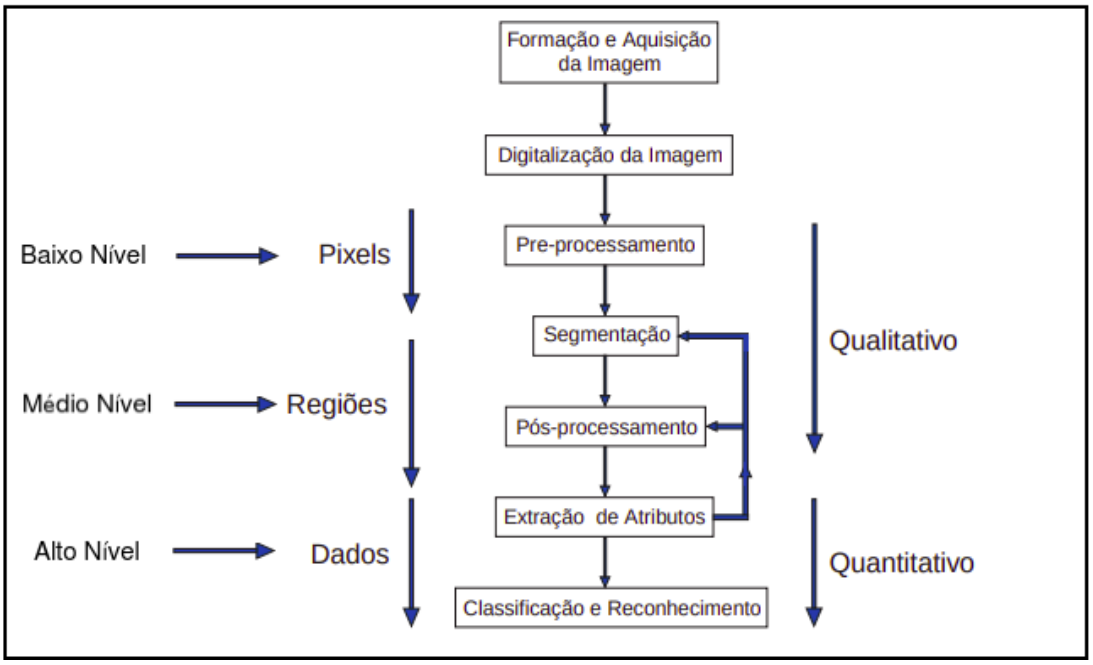
2.2 Processamento digital de imagens

Entende-se como Processamento Digital de Imagens (PDI), toda operação computacional caracterizada por ter uma entrada e uma saída de dados na forma de imagem (FILHO OGE MARQUES; VIEIRA NETO, 1999). Sua importância para os sistemas que utilizam Aprendizado de Máquina está relacionada ao preparo e a limpeza de ruídos, obtidos na fase de aquisição das imagens. Esses ruídos podem influenciar a performance dos classificadores. Como exemplo, os detectores de borda podem confundir possíveis regiões de interesse,

caso não sejam tratadas adequadamente. Como consequência, o ruído pode ocasionar uma classificação incorreta ao final do processamento.

De acordo com (GONZALES; WOODS, 2008), existem três principais subdivisões envolvendo o PDI, conforme ilustra a Figura 3: (i) baixo nível, que consiste em operações básicas dos pixels, como melhoria no contraste, redução de ruídos e recortes, tendo uma imagem como entrada e outra como saída, (ii) médio nível, cujo foco é a segmentação, ou seja, um processo no qual a imagem original é dividida em diversas ROIs, para que seja possível subsidiar a próxima etapa de mais alto nível, como o reconhecimento de objetos ou faces e (iii) alto nível, que tenta dar sentido aos objetos extraídos na etapa anterior, com base na análise dos ROIs em associação com técnicas de Aprendizado de Máquina (árvores de decisão, máquinas de vetores de suporte, etc). Em síntese, essa subdivisão permite a interpretação e a identificação de objetos em uma imagem, por exemplo, se é um cachorro ou um gato, além de inseri-lo em um contexto factível, ou seja, o cachorro pode estar sentado ou correndo no quintal, com medo, dentre outras situações, identificadas de modo automatizado.

Figura 3 – Subdivisão das técnicas de processamento digital de imagens



Fonte: (GONZALES; WOODS, 2008)

Normalmente, os algoritmos de PDI de baixo nível trabalham com imagens em escala de cinza, com apenas um canal de cor. A justificativa está relacionada à simplificação da codificação e o tempo necessário para execução, além de ser possível identificar e segmentar as ROIs pela intensidade de brilho presente neste único canal.

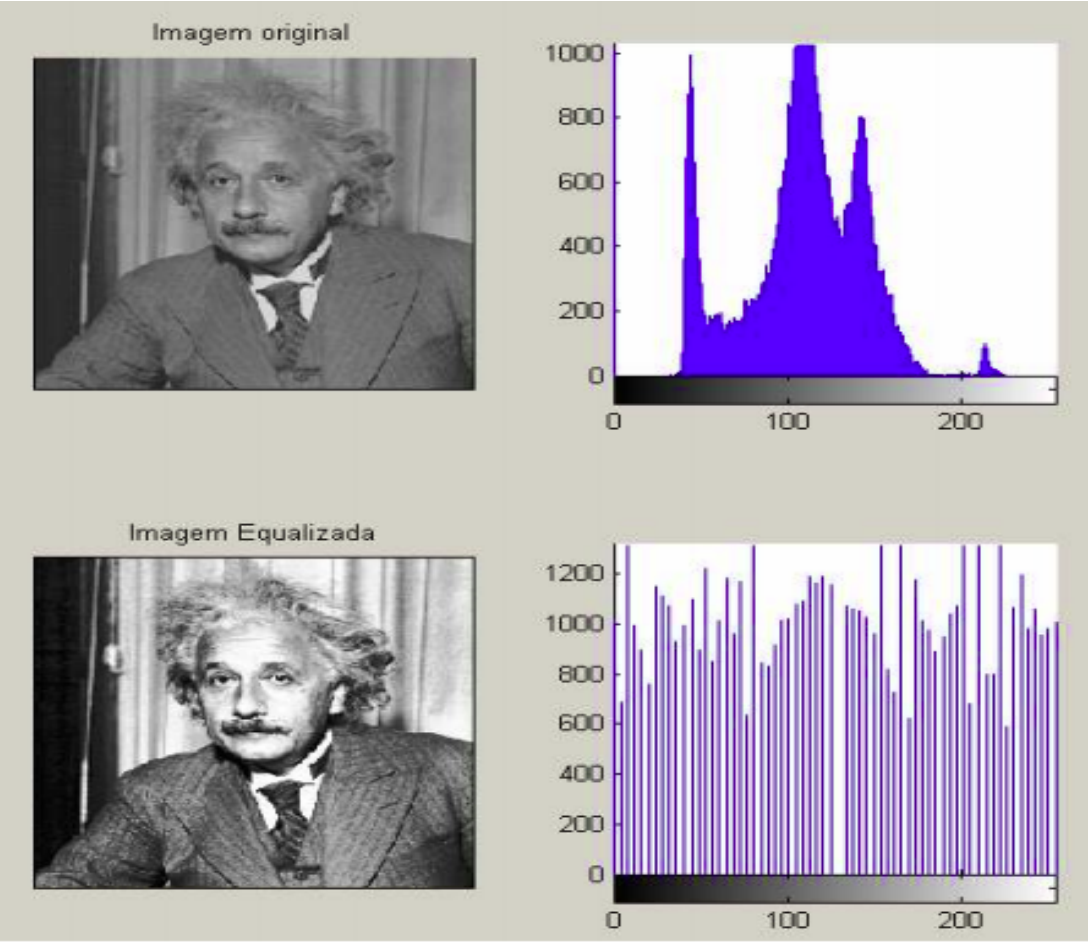
2.2.1 Equalização de Histograma

A equalização de histograma é uma técnica de PDI de baixo nível que se propõe a redistribuir a intensidade dos pixels de uma imagem, de modo a obter um padrão uniforme em toda a escala de cores (FILHO OGE MARQUES; VIEIRA NETO, 1999).

Quando uma imagem possui pouco contraste e brilho, torna-se difícil a sua interpretação e extração das informações desejadas. O aumento do intervalo dinâmico entre os níveis de cores possibilita uma melhora no contraste e no brilho das imagens, principalmente quando adquiridas sob condições de iluminação não ideais, proporcionando resultados mais otimizados na etapa de nível médio do PDI, responsável pela segmentação.

Este trabalho utiliza imagens obtidas sob diversas condições de aquisição, portanto, a equalização do histograma é fundamental para a obtenção de bons resultados. A Figura 4 ilustra as diferenças entre uma imagem original e o resultado após a aplicação da equalização do histograma.

Figura 4 – Processo de equalização de histograma



Fonte: disponível em: <https://web.fe.up.pt/TD-histograms.pdf>. Acesso em 23/11/2020

2.2.2 Detecção facial

A detecção facial é uma técnica de PDI de nível médio com o propósito de determinar a existência ou não de uma ou mais faces humanas em uma determinada imagem. Em caso positivo, deve-se retornar as suas respectivas localizações, por meio de coordenadas cartesianas. Apesar de ser uma tarefa trivial e simples para os seres humanos, a detecção facial pode ser considerada um problema desafiador para os computadores. Muitas variáveis podem estar envolvidas no processo, dificultando a execução. Dentre essas variáveis pode-se destacar a cor, a iluminação, o posicionamento do objeto na imagem e a escala.

A detecção facial tem fundamental importância na redução do espaço de busca das diversas soluções que envolvem esse tema. A diferença do significado semântico de duas palavras da língua inglesa, traduzidas como "exploração", resumem e ilustram a importância da detecção facial: *Exploration* e *Exploitation*. A primeira retrata uma busca global e exaustiva em toda região do problema, neste caso a imagem toda é analisada. Na segunda, passamos para uma busca local, mais otimizada e com possibilidade de obtenção de resultados mais consistentes. Portanto, a busca por informações faciais limita-se apenas na região de interesse e evita processamento em regiões desnecessárias.

A detecção facial automatizada se popularizou no início dos anos 2000, quando (VIOLA; JONES, 2001) desenvolveram o método Haar, capaz de obter resultados acima de 93% de precisão utilizando câmeras e processadores populares no mercado naquela época, tudo isso em tempo real. Foi um método revolucionário e muito utilizado ainda nos dias atuais, embora possua algumas limitações, como a alta taxa de Falsos Positivos retornados pelo algoritmo.

Outras soluções foram desenvolvidas ao longo dos últimos anos, com acurácia e desempenho superior. Dentre elas destacam-se o aperfeiçoamento do Histograma de Gradientes Orientados (do inglês HOG - Histogram of Oriented Gradients) (MONZO D; ALBIOL; MOSSI, 2010; GHORBANI G; TARGHI; DEHSHIBI, 2015) e das Redes Neurais Convolucionais (do inglês CNN - Convolutional Neural Network) (Ma; Wang, 2018).

Para o desenvolvimento deste trabalho, implementou-se um novo método de detecção facial, detalhado com mais profundidade no Capítulo 4 e denominado AEHZ, que utiliza a estrutura do algoritmo HOG.

2.2.3 Histograma de gradientes orientados (HOG)

O HOG é uma técnica de PDI muito utilizada para detecção e localização de diversos tipos de objetos, inclusive faces humanas. Sua principal funcionalidade é retornar um descritor de característica de forma. Por meio de um vetor de gradientes, ocorre a redução da dimensionalidade das imagens, eliminando informações desnecessárias, como exemplo, fundo com cor constante (DALAL; TRIGGS, 2005). Em vez de usar as informações de cada pixel, o HOG utiliza um agrupamento em blocos e usa classificadores de Aprendizado de Máquina para classificar e detectar objetos. A ideia principal do algoritmo é que a aparência e a forma dos objetos mudam conforme sua descrição de agrupamento. Essa informação pode ser bem caracterizada pelo gradiente de intensidade e pela distribuição na direção dessas variações.

Por definição, um gradiente é a variação da intensidade dos pixels em um determinado sentido, no caso de uma imagem em escala de cinza, ou variação da intensidade de cor de cada canal em uma imagem colorida (BURGER; BURGE, 2008). É um conceito bastante utilizado em detectores de borda. As variações de gradiente, muitas vezes bruscas, são decisivas e determinantes para a segmentação de regiões de interesse em imagens.

Uma característica interessante do HOG é a capacidade de ser invariante à rotação das imagens, com isso o deslocamento de um objeto durante a aquisição não adiciona ruído ao problema estudado.

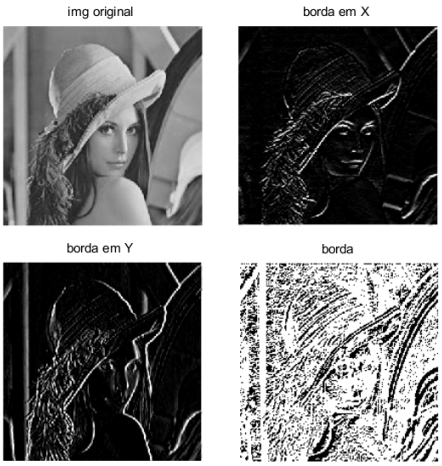
As etapas do processo de extração dos descritores de forma com o histograma de gradientes orientados são:

1. Primeiro passo: Cálculo do gradiente de todos os pixels da imagem, geralmente com a utilização de máscaras de derivação na vertical (Equação 2.1) e na horizontal (Equação 2.2), semelhante ao filtro de detecção de bordas de Sobel (LUBNA F.; KHAN; MUFTI, 2016) ilustrado na Figura 5.

$$G_x = \begin{bmatrix} 1 \\ 0 \\ \text{-}1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ \text{-}1 & \text{-}1 & \text{-}1 \end{bmatrix} \tag{2.1}$$

$$G_y = \begin{bmatrix} 1 & 0 & \text{-}1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & \text{-}1 \\ 1 & 0 & \text{-}1 \\ 1 & 0 & \text{-}1 \end{bmatrix} \tag{2.2}$$

Figura 5 – Cálculo do gradiente X e Y com o filtro de Sobel



Fonte: disponível em <http://www.epischisto.com/lena.jpg> Acesso em 02/12/2020

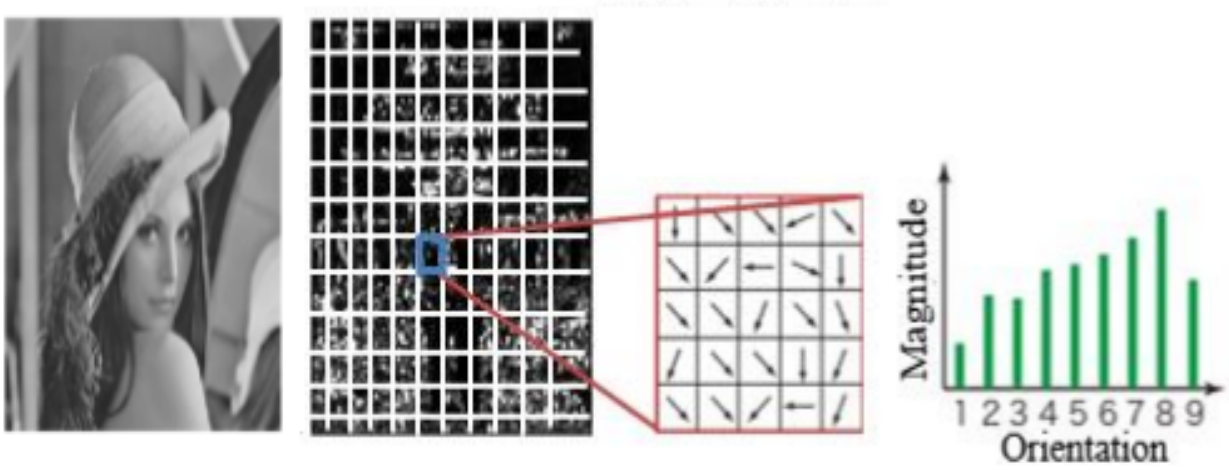
2. Segundo passo: Cálculo da magnitude (Equação 2.3) e orientação do gradiente (Equação 2.4) para todos os pixels da imagem.

$$G = \sqrt{G_x^2 + G_x^2} \tag{2.3}$$

$$\Theta = arctan\frac{G_x}{G_y} \tag{2.4}$$

3. Terceiro passo: construção do histograma de gradientes. A imagem é dividida em várias células, normalmente exponenciais de 2 (Ex: 8 x 8 ou 16 x 16). Para cada uma delas, um histograma de gradientes orientados é construído contando as ocorrências da magnitude e o sentido de orientação, conforme a Figura 6.

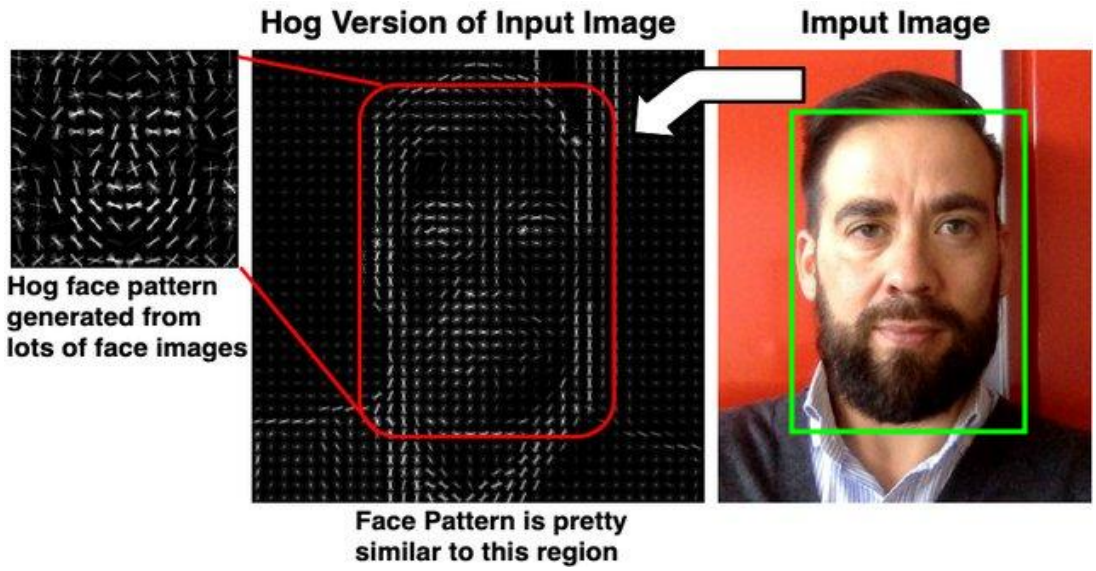
Figura 6 – Processo de extração de descritores de forma com HOG



Fonte: elaborada pelo autor

A Figura 7 ilustra um modelo treinado para detecção facial (esquerda) e uma imagem a ser analisada (direita), após a extração do seu HOG (centro).

Figura 7 – Detecção facial utilizando HOG



Fonte:<https://datascience.stackexchange.com/application of-histogram-of-oriented-gradients>. Acesso em: 02/10/2020.

2.3 Extração dos landmarks e redução da dimensionalidade

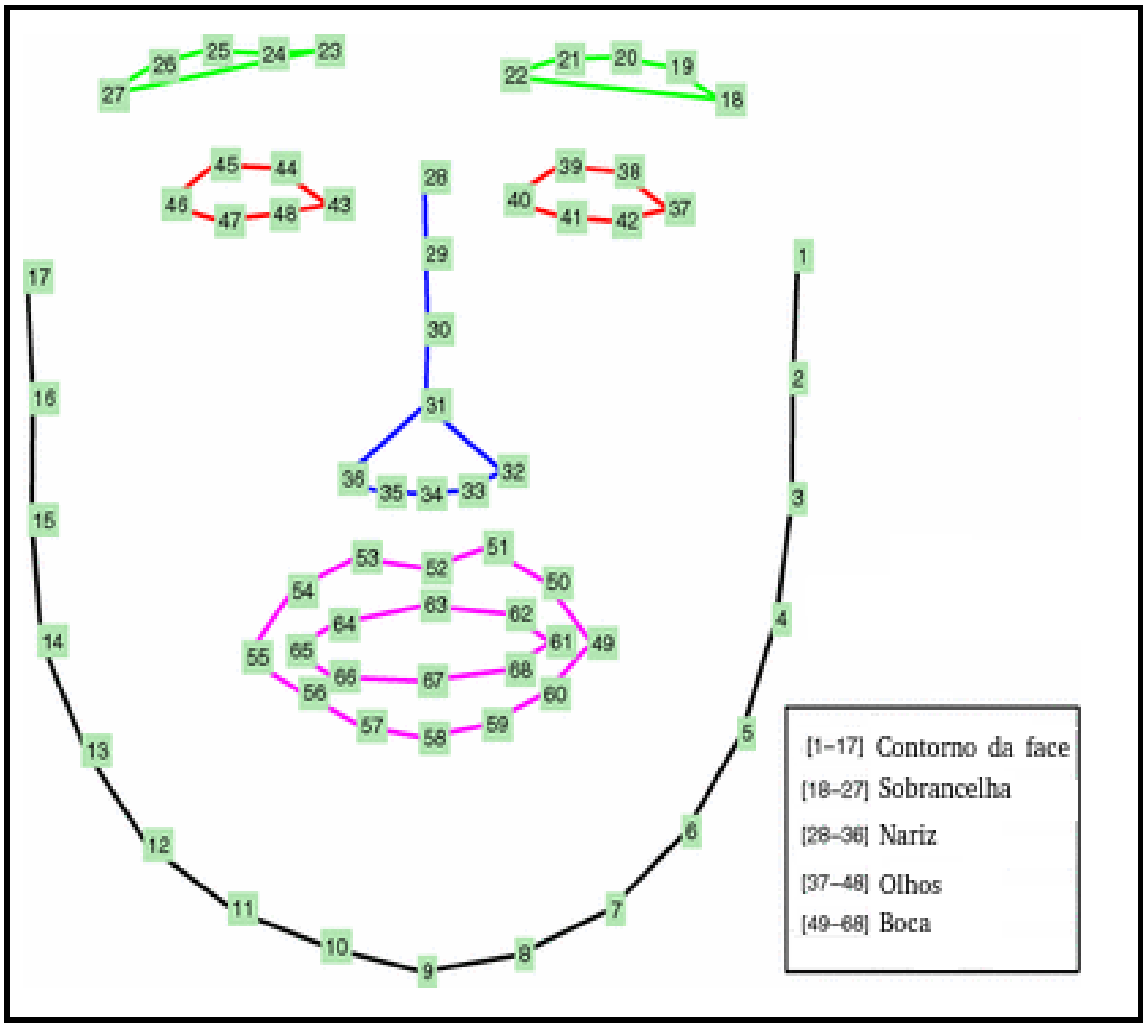
A capacidade de reconhecimento e identificação das expressões faciais, por parte dos seres humanos, está diretamente ligada com a identificação das diversas mudanças musculares que ocorrem nessa região (EKMAN; FRIESEN, 1971). Mudando para o contexto artificial, essas variações precisam ser mensuradas para que seja possível a comparação. Com isso torna-se imprescindível a identificação de diversos pontos de referência, que sejam universais e possibilitem essa diferenciação. Esses pontos de referência são os *landmarks*.

Como os *landmarks* são compatíveis para todas as faces, o uso de uma distância fixa comum, como a distância entre os olhos, permite normalizar os dados obtidos em imagens com diferentes escalas (ISMAIL; SABRI, 2009). No entanto, essa normalização é pouco eficaz para reduzir a variabilidade relacionada às condições de aquisição da imagem, principalmente com relação à rotação e as variações raciais. Um determinado grupo de indivíduos pode ter características sobressalentes, capazes de influenciar negativamente os resultados dos algoritmos de Aprendizado de Máquina, como exemplo podemos citar a largura da boca como um fator trivial para a identificação da emoção de felicidade, manifestada na forma de sorriso. Nesse cenário, um classificador treinado com o personagem Coringa, vilão do super-herói Batman, encontraria dificuldades para convergir e diferenciar

uma pessoa em relação à emoção de felicidade, quando utilizada a métrica de distância euclidiana para normalizar os *landmarks*.

A Figura 8 ilustra a localização automática de 68 *landmarks*; o objetivo é a identificação de vários componentes estruturais da face humana, entre eles: contorno do rosto, sobrancelha, olhos, nariz, boca, etc.

Figura 8 – Relação de todos os 68 *landmarks* faciais



Fonte: elaborada pelo autor

A partir da extração dos *landmarks*, a etapa de PDI é encerrada. Os dados processados migram dos pixels da imagem para as coordenadas bidimensionais (x e y) da localização dos landmarks faciais, reduzindo a dimensionalidade do problema estudado.

Do ponto de vista matemático, rotação é uma transformação linear envolvendo coordenadas espaciais, neste caso, em duas dimensões e que mantém invariante o módulo do comprimento de seus vetores, além de preservar a orientação no espaço físico (WINTERLE, 2014).

Portanto, o uso de coordenadas tende a produzir uma variabilidade menor que a

utilização de distâncias entre os *landmarks*, além de descartar um tempo de processamento adicional para computar as métricas de distância.

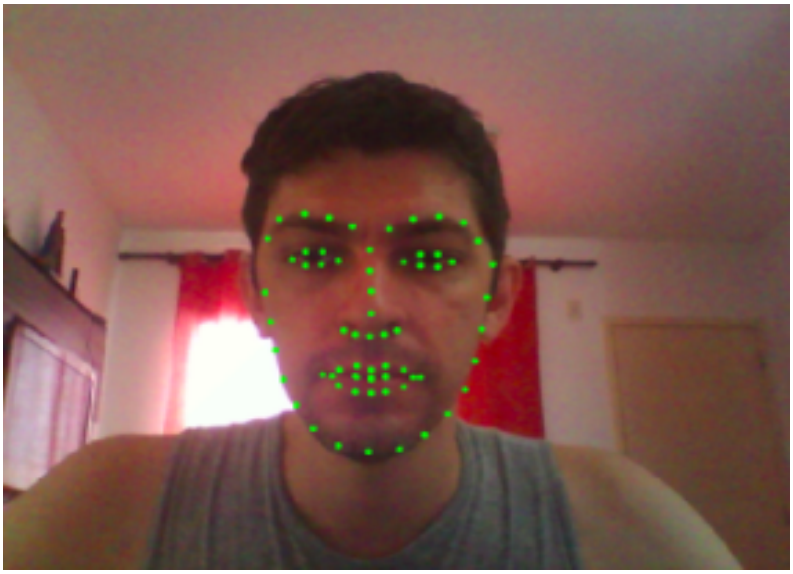
Outro fator importante, causador de variabilidade nos dados, nesse caso nas coordenadas dos *landmarks* é o fator de escala, relacionado com a proximidade do ator com o equipamento de captura da imagem.

O sucesso do uso de *landmarks* para a detecção de padrões faciais depende da implementação proposta. A detecção facial pode ser processada por meio de formas geométricas ou inferência prévia, pela detecção de regiões de borda e contraste (HOG) ou por métodos que utilizam as duas implementações (TESTA, 2019). Já a marcação dos *landmarks* utiliza inferência matemática, detecção automatizada de formas e também manualmente. O uso de diferentes métodos de marcação de *landmarks* produz implicações não triviais para a análise de expressões faciais e a escolha do método mais adequado deve ser norteada pelos objetivos almejados. Nos métodos manuais, o pesquisador define e ajusta os pontos de interesse em cada imagem (LI; DENG, 2018; LEE, 2009; LEE J. H.AND LEE, 2007). Esse processo tende a ser mais preciso, porém fracamente replicável. Já os métodos completamente automatizados utilizam algoritmos inteligentes e modelos estatísticos preditivos previamente treinados (HASTIE, 2001; VAILLANCOURT, 2010). Seus resultados são dependentes do conjunto amostral, porém tendem a ser mais rápidos e facilmente replicáveis. Por fim, as técnicas semi-automatizadas utilizam rotinas que intercalam passos automatizados com etapas supervisionadas.

O detector de *landmarks* faciais incluído na biblioteca Dlib (KING, 2009) possui implementação que utiliza um conjunto de treinamento de pontos fiduciais faciais manualmente demarcados e rotulados. As coordenadas bidimensionais (x e y) das regiões ao redor de cada estrutura facial são calculadas. Com esses dados de treinamento, um conjunto de árvores de regressão, um algoritmo de Aprendizado de Máquina, projeta as posições dos *landmarks* diretamente por meio da intensidades dos pixels. O resultado final é um detector de *landmarks* faciais que pode ser utilizado em tempo real com previsões de alta confiabilidade e assertividade.

A Figura 9 ilustra uma aplicação prática da extração dos *landmarks* faciais de forma automática, com o auxílio da biblioteca DLib, implementada em linguagem de programação Python 3, em um sistema operacional Linux Ubuntu 18.04, com um processador Intel i5 8225U e 8 Gb de memória RAM.

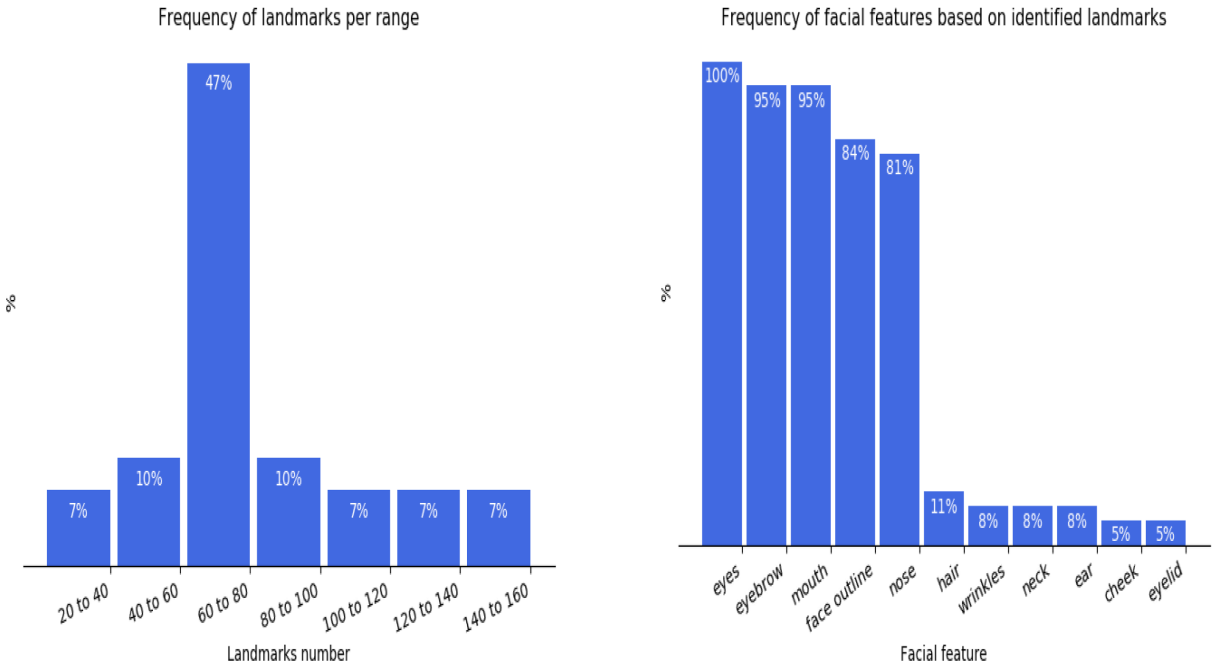
Figura 9 – Extração dos 68 landmarks faciais em linguagem Python



Fonte: elaborada pelo autor

Em (TESTA, 2019), os autores realizaram uma revisão bibliográfica de todos o trabalhos que utilizam o reconhecimento de estruturas faciais por meio da extração dos *landmarks*. A Figura 10 apresenta os resultados sintetizados. Nota-se que as estruturas predominantes na detecção dos *landmarks* são os olhos, as sobrancelhas, a boca, o contorno do rosto e o nariz. Em relação à quantidade de *landmarks* detectados, predomina o intervalo entre 60 e 80 coordenadas, presentes em quase metade dos trabalhos da literatura.

Figura 10 – Estrutura de *landmarks* encontrados na literatura



Fonte: (TESTA, 2019).

Outro fator causador de ruído, relacionado com o processamento de imagens faciais, concentra-se na posição relativa do ator com o equipamento de captura e eventuais rotações, tanto na horizontal quanto na vertical. Como solução para esse problema pode-se utilizar as técnicas de frontalização.

2.4 Técnicas de frontalização

Técnicas de frontalização possibilitam sintetizar artificialmente vistas frontais para diversos tipos de objetos, incluindo faces humanas, a partir de fontes originais rotacionadas. Sua utilização tende a aumentar substancialmente o desempenho dos sistemas de classificação e reconhecimento que utilizam imagens, normalizando as variações que podem ocorrer durante o processo de aquisição (HASSNER, 2015). Além disso, as etapas de treinamento e teste, utilizadas por algoritmos de Aprendizado de Máquina, podem ser realizadas com as amostras padronizadas em uma posição única.

As técnicas de frontalização de imagens classificam-se de duas formas distintas, segundo (Vonikakis; Winkler, 2020) e listadas a seguir:

- **Frontalização pela aparência:** tenta-se produzir uma vista frontal completa de uma face humana, por meio do ajuste com um modelo 3D frontal médio, previamente treinado. A imagem de entrada precisa ser renderizada (e.g. construída novamente) para projetar os pixels no novo formato. Esse processo ocasiona alterações na imagem final e que são proporcionais ao ângulo de rotação da imagem original.

Figura 11 – Frontalização facial por aparência

(a) (b) (c)

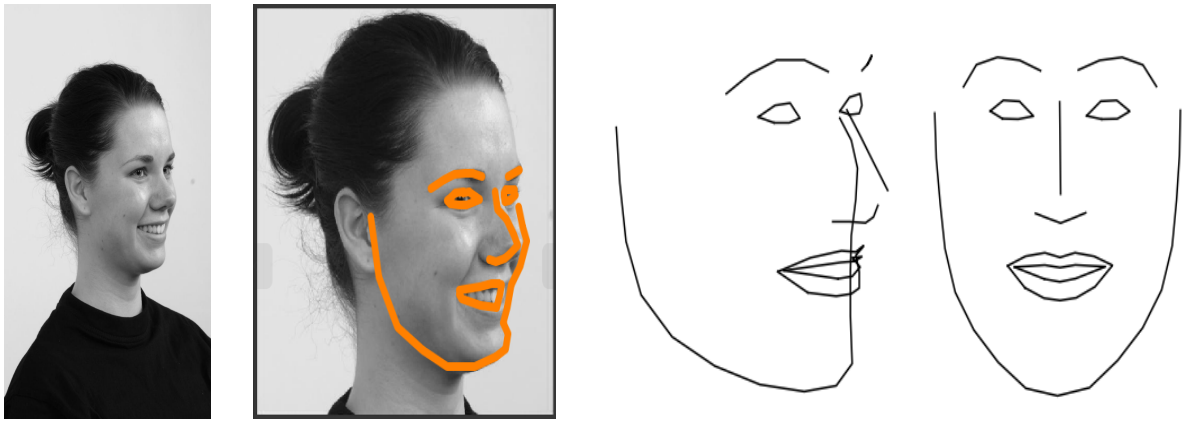
Fonte: adaptada de (HASSNER, 2015)

A Figura 11 exibe o resultado de um método que utiliza a Frontalização pela aparência: (a) Imagem original, (b) Imagem Frontalizada e (c) Extração dos *landmarks* da imagem (b). O resultado final comprova que não houve perda de informação das coordenadas após as transformações realizadas. Um aspecto negativo dessa técnica

é o custo computacional de processamento. Sua utilização em sistemas de tempo real é inviável já que o processo de renderização da imagem frontalizada é lento.

- **Frontalização das coordenadas (*landmarks*):** concentra-se em simular a localização dos landmarks na posição frontal, por meio do cálculo de uma face média durante a etapa de treinamento, descartando qualquer processamento em nível de pixel. Essa abordagem tende a ser consideravelmente menos dispendiosa do ponto de vista de tempo de processamento computacional, uma vez que não requer renderização. Nos últimos anos, a conversão de coordenadas 3D, a partir de imagens 2D tornou-se possível e viável (ZHAO, 2018), simplificando e produzindo vistas frontais de diversos tipos de objetos em tempo real. O processo produz uma matriz de pesos espaciais com as mesmas dimensões do problema original, neste caso, 68 coordenadas em duas dimensões $[68, 2]$ que são multiplicados pelas coordenadas faciais originais que, na grande maioria dos casos, possuem algum tipo de rotação. Segundo (Valstar, 2017), as diferenças de pose e rotação são responsáveis por mais da metade de toda a variação em sistemas que processam faces humanas. Essas variações podem interferir sistematicamente nos resultados dos algoritmos de Aprendizado de Máquina. A Figura 12 ilustra o processo de Frontalização das coordenadas na seguinte ordem: (a) imagem original, (b) detecção das coordenadas dos landmarks faciais, (c) isolamento das coordenas dos landmarks faciais e (d) coordenadas faciais após o processo de Frontalização.

Figura 12 – Frontalização facial de coordenadas (*landmarks*)



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

Fonte: elaborada pelo autor

Considerações finais

Neste capítulo, foram apresentadas as principais técnicas de aquisição e processamento de imagens utilizadas neste trabalho, além do detalhamento das principais técnicas de detecção facial. As etapas de redução de variabilidade e de dimensionalidade, por meio da extração dos *landmarks* faciais também foram descritas, juntamente com as técnicas de frontalização facial.

Todo o referencial teórico apresentado possibilita preparar os dados para indução nos algoritmos de Aprendizado de Máquina, discutidos no próximo capítulo e responsáveis pela obtenção do conhecimento e classificação das emoções humanas.

3

Reconhecimento de Padrões

O Aprendizado de Máquina é um segmento da área de Inteligência Artificial, modelado na ideia de que sistemas computacionais automatizados possam aprender com dados, identificar padrões e tomar decisões com o mínimo de intervenção humana (HASTIE, 2001; DOMINGOS, 2015; VAILLANCOURT, 2010).

Os algoritmos de Aprendizado de Máquina suportam um mecanismo denominado indução ou inferência para obter conclusões e classificações genéricas, a partir de um conjunto particular de exemplos. Em síntese, o conhecimento está intrínseco nos dados. Por meio de uma etapa inicial de treinamento, é possível criar um modelo preditivo, capaz de gerar conhecimento e convergir quando recebe novos exemplos na fase de teste (MITCHELL, 1997).

Em linhas gerais, segundo (NORVIG; RUSSELL, 2013), o Aprendizado de Máquina consiste em intercalar métodos estatísticos e processos lógicos em algoritmos capazes de solucionar problemas de alta complexidade, os quais são utilizados para gerar aplicações inteligentes. Em seguida, essas aplicações são projetadas para resolver problemas reais e apoiar na tomada de decisões. Alguns exemplos são: reconhecimento de voz e de faces humanas, detecção de fraude e análise de riscos financeiros, etc. Os principais algoritmos de Aprendizado de Máquina estão disponíveis para as mais diversas linguagens de programação (C, C++, Java, Python, R, dentre outras) e sistemas operacionais (Windows, Linux, IOS, Android, dentre outros).

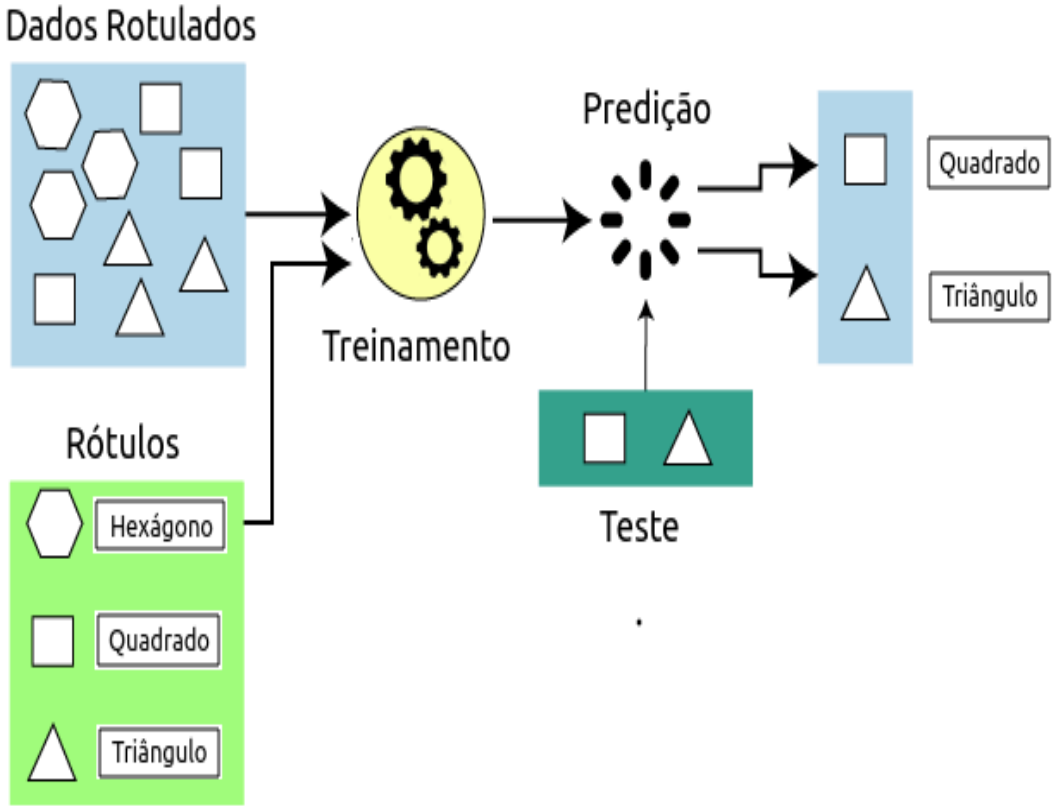
Existem detalhes a serem considerados antes da elaboração de softwares que utilizam Aprendizado de Máquina em sua execução. A escolha dos métodos e das bibliotecas deve ser guiada primariamente pelas características do problema e do projeto. As implementações específicas, a performance computacional durante a execução e os requisitos mínimos do sistema são determinantes e devem ser analisados durante a fase de planejamento. Embora haja particularidades em cada projeto, pode-se delinear um fluxo geral de passos para a criação de algoritmos que utilizam Aprendizado de Máquina (VAILLANCOURT, 2010) como: (i) extração de características e pré-processamento de dados, (ii) redução de

dimensionalidade, (iii) classificação das amostras e geração do classificador e (iv) teste do classificador gerado.

Os algoritmos de Aprendizado de Máquina podem ser divididos em duas categorias distintas:

- **Aprendizado supervisionado:** os dados de treinamento são rotulados, conforme a saída desejada. Por meio dessas informações, os computadores indicam a saída em um processo semelhante ao aprendizado humano auxiliado por um professor. Um algoritmo de aprendizado supervisionado, ilustrado na Figura 13, fornece dados de entrada e seus respectivos rótulos de saída para o indutor (classificador), responsável por identificar toda a sistemática que liga essas informações. O objetivo é encontrar uma função de mapeamento (F) para unir a variável (ou variáveis) de entrada (x) com a sua respectiva variável de saída (y), conforme a equação $F(x) \rightarrow y$. Os problemas que utilizam Aprendizado de Máquina supervisionado podem ser subdivididos em Regressão, quando existe uma aproximação numérica da função alvo, similar à previsão do valor de uma determinada ação na bolsa de valores; e Classificação, quando o resultado final é uma classe específica, similar ao processo de reconhecimento de emoções humanas.

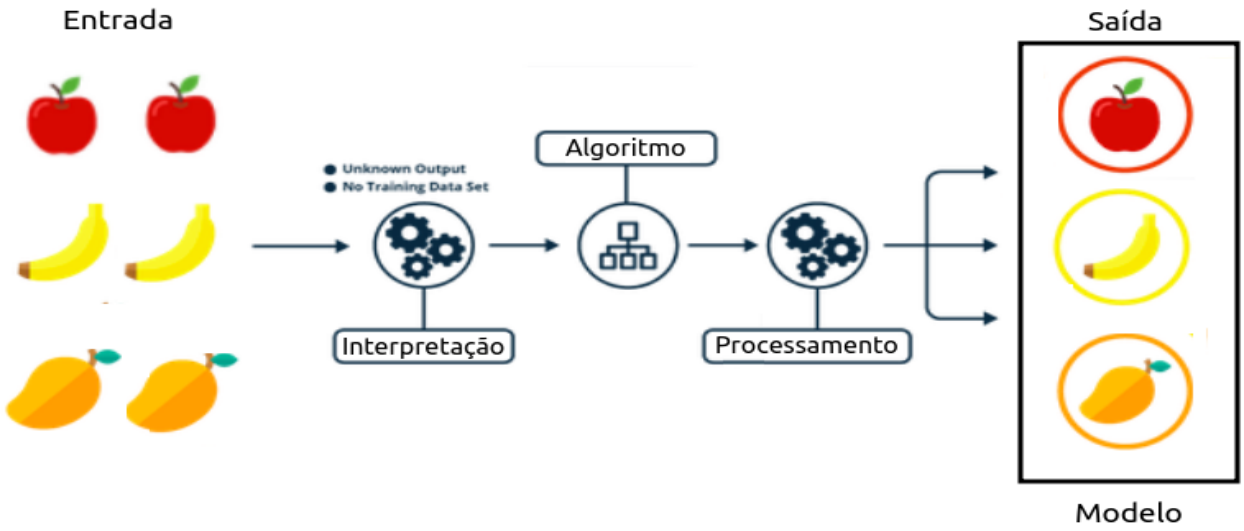
Figura 13 – Aprendizado de máquina supervisionado



Fonte: adaptado de <www.javatpoint.com/supervised-ML>. Acesso em 26/12/2020.

- **Aprendizado não supervisionado:** os dados de treinamento não são rotulados. Como o nome sugere, no aprendizado não supervisionado, ilustrado na Figura 14, os próprios modelos identificam os padrões e associações ocultas nos dados fornecidos. Pode ser comparado ao aprendizado que ocorre no cérebro humano quando aprendemos informações novas. Para encontrar os padrões nos dados, os algoritmos utilizam agrupamentos, normalmente por meio de métricas de semelhança. Uma das mais utilizadas é a menor distância euclidiana entre dois elementos, ou seja, a menor distância em linha reta entre eles.

Figura 14 – Aprendizado de máquina não supervisionado



Fonte: adaptado de <www.javatpoint.com/unsupervised-ML>. Acesso em 26/12/2020.

3.1 Algoritmos de Aprendizado de Máquina supervisionado

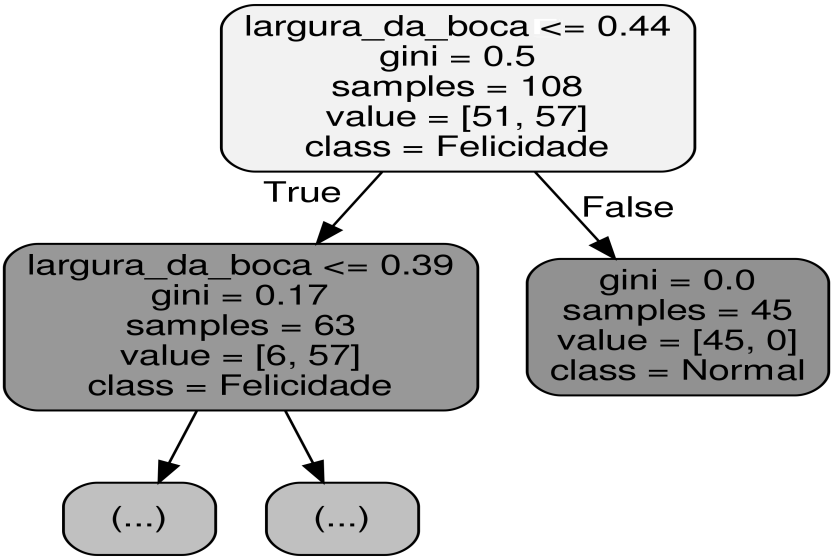
Um dos objetivos deste trabalho foi a construção do método REGL, capaz de reconhecer emoções humanas por meio da geometria da face, identificando as variações que ocorrem entre a manifestação de uma emoção e outra. Nesse sentido, é importante que o resultado final seja maleável e flexível aos diversos tipos de algoritmos de Aprendizado de Máquina supervisionado que existem. Dentre os quais pode-se destacar, de forma sucinta, os seguintes algoritmos:

- **K-vizinhos mais próximos (KNN):** é um algoritmo de Aprendizado de Máquina norteado em instância, que utiliza o princípio de que em um espaço de características. Uma amostra é rotulada de acordo com outras k amostras mais próximas a ela.

Sem conhecimento prévio, costuma-se utilizar a distância euclidiana como métrica. Os valores de k são determinantes para a performance dos resultados. Quando o valor de k é pequeno, a classificação fica mais sensível a regiões bem próximas e suscetível em considerar ruído como informação relevante. Com um valor de k grande a classificação fica menos sujeita a interferência, mas pode aumentar muito o tempo de execução e dificultar a convergência para os resultados.

- **Árvore de decisão (*DecisionTree*):** estabelece regras para a aprendizagem e simula o pensamento lógico humano para alcançar a classificação, dividindo os atributos do problema de acordo com o ganho de performance em cada divisão. Árvores de decisão são estruturas de dados formadas por um conjunto de elementos que armazenam as informações em suas extremidades, chamadas de nós. Toda árvore possui um nó chamado raiz, que possui o maior nível hierárquico (o ponto de partida) e ligações para outros elementos, denominados filhos. A Figura 15 ilustra uma Árvore de Decisão para classificar se uma pessoa está sorrindo ou não, de acordo a largura da boca e utilizando como métrica de ganho de informação o índice de Gini (medida estatística de assimetria).

Figura 15 – Árvore de decisão para detecção de sorriso.



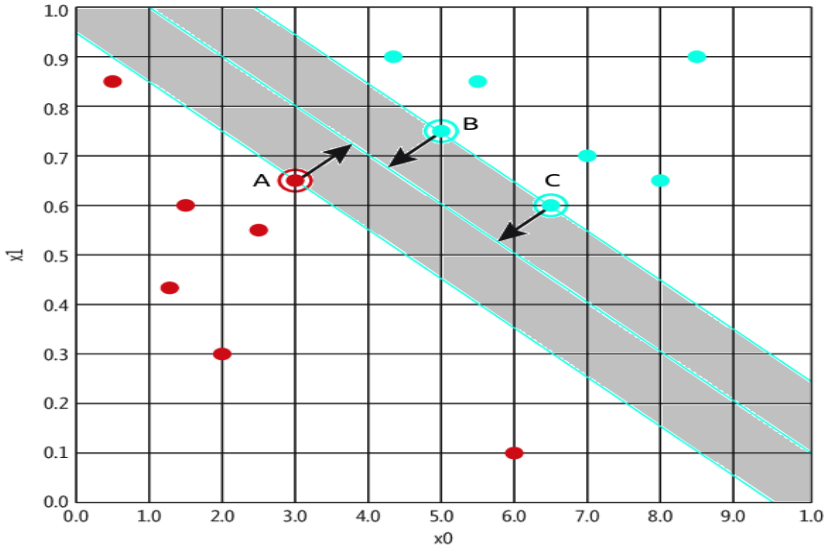
Fonte: elaborada pelo autor

- **Florestas aleatórias (*RandonForest*):** o conceito fundamental das florestas aleatórias é simples e poderoso. Trata-se da "sabedoria das multidões". Esse algoritmo divide o problema em diversas árvores de decisão não correlacionadas, com os atributos embaralhados. No final, os resultados individuais se unem para produzir o resultado final, de consenso estatístico mais robusto e consistente. Sua arquitetura utiliza o conceito de divisão e conquista na área da Ciência da Computação.

- **Árvores extras (*Extratrees*):** também chamadas de Árvores Extremamente Aleatórias (*ExtremelyRandomizedTrees*) são muito similares ao classificador floresta aleatória, exceto pelo critério de construção das árvores de decisão. Todos os dados de treinamento e os atributos (variáveis do problema) são utilizados na construção do modelo de aprendizado. Outro fator indispensável para o desempenho está vinculado ao fato que as árvores possuem apenas um nó de decisão, diminuindo o tempo de processamento.

- **Máquinas de vetores de suporte (Support vector machine - SVM):** é um modelo amplamente utilizado na classificação binária, quando existem apenas duas classes a serem analisadas. O algoritmo de Aprendizado de Máquina SVM obtém diversos vetores de suporte (pontos no espaço de busca do problema) para maximizar a margem geométrica entre amostras negativas e positivas. Para problemas com mais de duas classes, realiza-se a decomposição em várias sequências binárias e classifica-se uma de cada vez. É um algoritmo muito requisitado em diversos campos de pesquisa científica pela sua velocidade de processamento, resultados consistentes e capacidade de generalizar os mais diversos problemas. A Figura 16 apresenta uma divisão binária, utilizando o classificador SVM. Os vetores de suportes são os pontos A, B e C, e a região de decisão está demarcada na cor cinza.

Figura 16 – Separação binária utilizando SVM

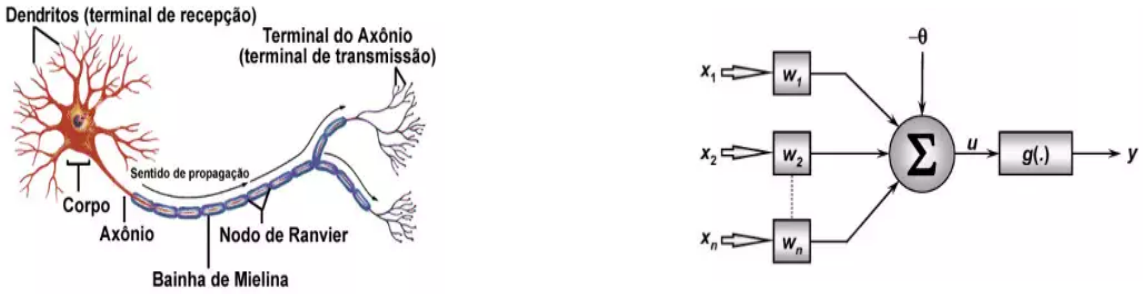


Fonte: elaborada pelo autor

- **Perceptron multicamadas (Multilayer perceptron):** as redes neurais artificiais, representante dos algoritmos conexionistas, são estruturas matemáticas flexíveis, capazes de distinguir relações não lineares complexas entre dados de entrada e dados de saída, por meio de mecanismos similares ao funcionamento dos sistemas neurais biológicos. O mais simples dos algoritmos conexionistas é o Perceptron. Esse

algoritmo possui apenas uma camada de entrada e uma camada de saída, responsável pelos resultados. Sua utilização restringe-se em aplicações que possuam apenas duas classes. Para problemas com mais classes, torna-se necessário uma implementação mais complexa e que utilize camadas ocultas (intermediárias) para a resolução de problemas não linearmente separáveis. A Figura 17 ilustra a comparação entre um neurônio biológico (a) e um neurônio artificial (b), de acordo com (MCCULLOCH; PITTS, 1943). As entradas (x_1, x_2, ... x_n) são multiplicadas pelos respectivos pesos (w_1, w_2, ... w_n) em um processo similar ao realizado pelos dendritos biológicos. O somatório das multiplicações das entradas de dados com seus pesos Σ passa por uma função de ativação g, responsável em restringir os valores iniciais e ativar (ou não) o neurônio em processamento. O conhecimento está associado com a camada de saída do neurônio artificial y.

Figura 17 – Comparação entre um neurônio biológico e um neurônio artificial



(a) (b)

Fonte: disponível em: <https://www.monolitonimbus.com.br/RNN>. Acesso em 27/12/2020.

3.2 Avaliação dos classificadores

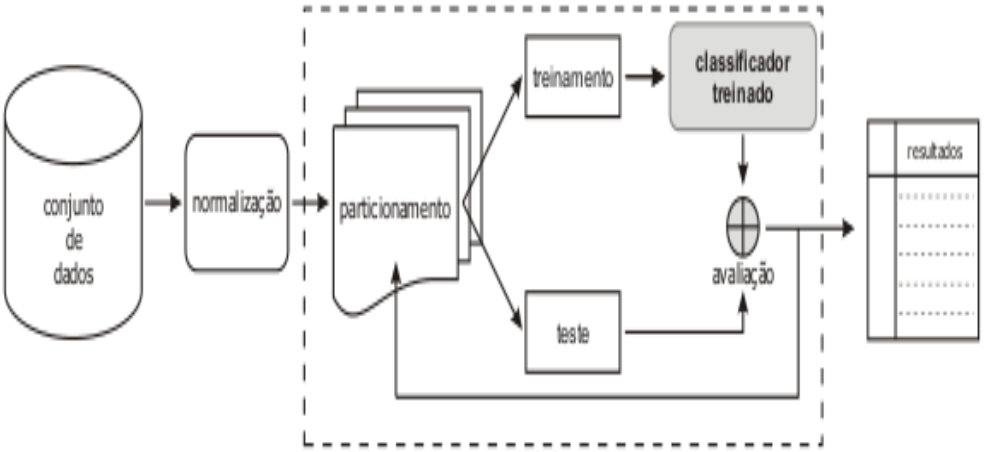
Os diversos bancos de dados de imagens utilizados neste trabalho, bem como os vários algoritmos de Aprendizado de Máquina implementados, necessitam de metodologias sistemáticas para avaliação dos resultados. Esses critérios estabelecem normas para garantir qualidade, padronização e consistência.

Independente do algoritmo de Aprendizado de Máquina utilizado, uma das fases fundamentais do processo de classificação é a divisão do conjunto de dados em treinamento e teste. Essa divisão dimensiona a capacidade de generalização e configuração do modelo. Diversas abordagens podem ser empregadas para a separação do conjunto, as quais são denominadas métodos de validação cruzada. A utilização dessas técnicas minimiza problemas comumente encontrados em reconhecimento de padrões, como por exemplo, a generalização do classificador (Jain, 2000).

Para um resultado mais preciso no processo de classificação, o método de validação cruzada *k-fold* foi utilizado para particionar os dados em conjunto de treinamento e conjunto de teste.

O método de validação cruzada *k-fold* divide o conjunto de dados em k partições mutuamente exclusivas. Os dados selecionados para cada partição são definidos randomicamente ou auxiliado por métodos estatísticos. Em cada iteração do método, uma das k partições é utilizada como conjunto de teste e as outras k - 1 partições são usadas como conjunto de treinamento. Nos experimentos deste trabalho, utilizou-se o valor de k = 10. O classificador é construído a partir dos exemplos selecionados no conjunto de treinamento. Para a avaliação são utilizados os elementos do conjunto de teste. Este procedimento é realizado 10 vezes. A Figura 18 ilustra o processo de reconhecimento de padrões, em que é destacado o método de validação cruzada com *k-fold*.

Figura 18 – Validação cruzada (K-fold) com k = 10

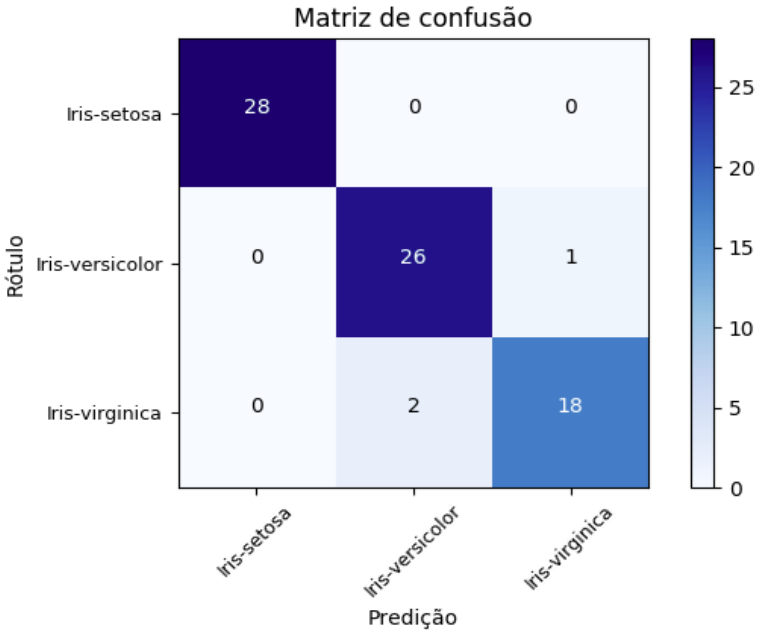


Fonte: adaptado de (PLOTZE, 2010)

A avaliação do desempenho de um algoritmo de Aprendizado de Máquina pode ser mensurada por meio de um conjunto de informações, extraídas após o resultado da classificação. Para isso, uma estrutura em forma de tabela, denominada matriz de confusão, contém as informações obtidas a partir dos resultados. A diagonal principal da matriz concentra todos os exemplos que foram classificados de forma correta. Já os outros elementos, fora da diagonal principal, contém todos os exemplos rotulados de forma incorreta. Portanto, uma matriz de confusão na forma de matriz diagonal estabelece o cenário mais otimista para as previsões de um algoritmo de Aprendizado de Máquina, cujos dados seriam todos classificados de forma correta, ou seja, acurácia de 100%.

A Figura 19 apresenta uma matriz de confusão, extraída a partir da classificação de flores da espécie Iris, uma base de dados muito conhecida na área de Aprendizado de Máquina. Os resultados indicam uma acurácia de 96%, utilizando o algoritmo de Aprendizado de Máquina do tipo SVM (máquinas de vetores de suporte) e validação cruzada *k-fold* com 10 amostras.

Figura 19 – Matriz de confusão para classificação de flores da espécie Iris



Fonte: disponível em: <https://docs.lemonade.org.br/pt-br/criar-um-experimento.html>. Acesso em 22/09/2021

Considerando P o total de amostras positivas (pertencente à classe) e N o total de amostras negativas (não pertencente à classe), diversas informações estatísticas podem ser obtidas a partir da matriz de confusão, como a acurácia, precisão, sensibilidade, medida-F e gráficos ROC, descritos a seguir.

3.2.1 Acurácia

A acurácia é a porcentagem geral de acerto de um classificador de Aprendizado de Máquina, conforme Equação 3.1. No exemplo da Figura 19, basta realizarmos a razão entre a soma de todos os elementos da diagonal principal da matriz (72 amostras) pela soma do total de elementos da matriz (75 amostras), ou seja, uma acurácia em torno de $96,00\%$.

$$Acurácia = \frac{P}{P+N} \tag{3.1}$$

Embora seja uma métrica bastante utilizada para ilustrar o desempenho geral dos resultados, situações com classes desbalanceadas podem resultar em uma alta taxa de

acurácia e dificuldade de generalizar as classes minoritárias, ou seja, o classificador tende a considerar todas as amostras como pertencentes à classe majoritária.

Para resultados mais consistentes outras métricas são utilizadas em conjunto com a acurácia.

3.2.2 Precisão

A precisão mede a proporção de predições positivas que estão corretas e quão bem o modelo consegue encontrar os valores positivos. É uma razão entre os positivos verdadeiros TP de cada classe e a soma dos positivos verdadeiros e os falsos positivos FP, conforme apresentado na Equação 3.2.

No exemplo da Figura 19, pode-se calcular a precisão da espécie iris-versicolor pela razão entre o total da diagonal principal da classe, célula onde o número da linha é igual ao número da coluna (26 amostras) pelo total da coluna da classe medo (28 amostras), retornando um resultado de $92,86\%$.

$$Precisão = \frac{TP}{TP + FP} \tag{3.2}$$

3.2.3 Sensibilidade (*Recall*)

A sensibilidade mede a proporção dos valores que são de fato positivos e que foram classificados corretamente, ou seja, a frequência em que o classificador encontra os exemplos de uma determinada classe. É uma proporção entre os positivos verdadeiros TP de cada classe e a soma dos positivos verdadeiros e os falsos negativos FN, conforme apresentado na Equação 3.3.

No exemplo da Figura 19, pode-se calcular a sensibilidade na classificação da espécie iris-versicolor pela razão entre o total da diagonal principal da classe, célula onde o número da linha é igual ao número da coluna (26 amostras), pelo total da linha da classe iris-versicolor (27 amostras), retornando um resultado de $96,30\%$. Neste exemplo, o total da linha sempre resultará no total de amostras induzidas no classificador.

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

3.2.4 Medida-F

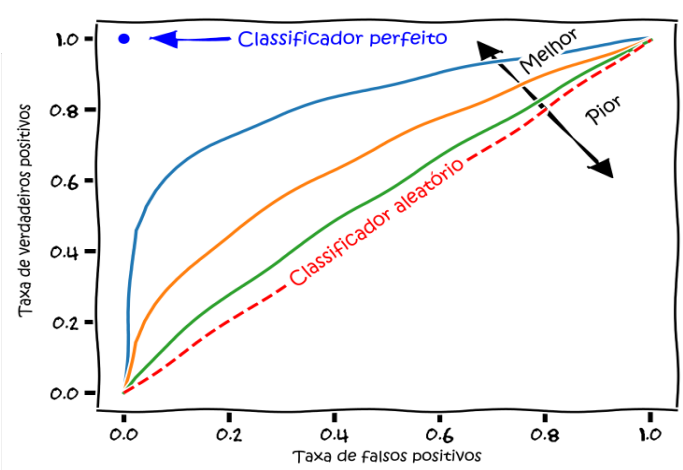
A medida-F combina precisão e sensibilidade em um único número, que indica a qualidade geral de um modelo de classificação para algoritmos de Aprendizado de Máquina. Essa métrica possui como ponto forte a possibilidade de alcançar um resultado confiável, mesmo com conjuntos de dados desproporcionais e desbalanceados. A Equação 3.4 apresenta o cálculo da Medida-F e o resultado para a classe iris-versicolor 19 foi aproximadamente $94,55\%$.

$$Medida - F = \frac{2}{\frac{1}{Precisao} + \frac{1}{Sensibilidade}} \tag{3.4}$$

3.2.5 Gráfico ROC (Característica de Operação do Receptor)

Um gráfico ROC é uma técnica empregada para visualização do desempenho de um classificador. Esse gráfico é comumente utilizado para descrever a relação entre as taxas de acerto (verdadeiros positivos) e erro (falsos positivos) de um classificador. A construção do gráfico utiliza a taxa de verdadeiro positivo no eixo Y e a taxa de falso positivo no eixo X. Assim, um ponto no espaço ROC é discretizado na forma de uma coordenada bidimensional no plano cartesiano. O ponto (0,1) indica um classificador perfeito: ele classifica todos os casos positivos e negativos corretamente. Quanto mais próximo de 1 no eixo y a área sobre a curva ROC estiver, melhor o desempenho do classificador, segundo (FAWCETT, 2006). A Figura 20 ilustra a área sobre a curva ROC de um algoritmo de Aprendizado de Máquina fictício com a classificação perfeita.

Figura 20 – Área sobre a curva ROC - classificador perfeito



Fonte: disponível em: <https://pt.wikipedia.org/wiki/cruvaROC>. Acesso em 10/10/2021

Considerações finais

Neste capítulo, foram apresentados os principais conceitos relacionados ao reconhecimento de padrões. Essas técnicas têm como objetivo principal associar um padrão a uma determinada classe, neste trabalho, uma emoção humana. Foram descritos os principais algoritmos de Aprendizado de Máquina presentes na literatura (SVM, Árvores de Decisão, Florestas Aleatórias, Árvores Extras, KNN e Redes Neurais). Por fim, um conjunto de métricas para avaliação dos classificadores foi detalhado, envolvendo validação cruzada, matrizes de confusão e gráficos ROC.

4

Método REGL: Materiais e Desenvolvimento

Os capítulos anteriores apresentaram, em linhas gerais, aspectos do processamento de imagens, redução de dimensionalidade e aprendizado de máquina para detecção e reconhecimento de padrões. O presente capítulo apresenta o desenvolvimento de um novo método, denominado REGL, fundamentado no estudo da geometria facial dos *landmarks*. O método tem por objetivo extrair dados de posicionamento relativo das estruturas faciais e calcular uma medida mais acurada dos movimentos musculares da face, como aqueles produzidos por expressões faciais. A vantagem deste método é a redução da variabilidade amostral, desde que a identificação do indivíduo (reconhecimento da pessoa) não seja um requisito. Este capítulo apresenta todo o processo de preparação dos dados para a indução nos algoritmos de Aprendizado de Máquina, por meio da normalização dos *landmarks* e direcionado para a classificação das emoções humanas de forma mais consistente e eficiente. A Seção 4.1 introduz os bancos de dados de expressões faciais utilizados como entrada de dados dos algoritmos. Já a Seção 4.2 apresenta o método REGL, sua implementação e detalhes técnicos

4.1 Bancos de dados de expressões faciais

Um banco de dados de expressões faciais é uma coleção de imagens ou videoclipes digitais com diferentes atores. Seu conteúdo é essencial para o treinamento, teste e validação dos algoritmos de Aprendizado de Máquina e para o desenvolvimento de sistemas de reconhecimento facial e de expressões faciais, nas quais se enquadram as emoções. A maioria destes bancos de imagens é norteada pela base teórica das emoções humanas (ALVAREZ, 2013), que pressupõe a existência de seis tipos diferentes de expressões faciais: felicidade, medo, nojo, raiva, surpresa, tristeza, além da expressão neutra (ou indiferença), conforme citado na Introdução.

Os bancos de dados de expressões faciais utilizados neste trabalho são os seguintes:

- **FEI Database:** modelado segundo o banco de dados de imagens FERET (PHILLIPS, 1998), o Centro Universitário FEI, localizado em São Bernardo do Campo/SP, desenvolveu o *FEI FaceDatabase* (JUNIOR LEO L.; THOMAZ, 2006) que possui imagens de 200 atores, cada um em 14 poses diferentes, sendo: Frontal em repouso, Frontal sorrindo, Frontal em repouso com baixo contraste, Frontal em repouso com alto contraste, rotações à direita e à esquerda em repouso (10^o, 30^o, 60^o, 80^o e 90^o respectivamente).

- **Genki4K:** bastante utilizado para a detecção de sorriso (emoção de felicidade) (NICK, 2009) e amplamente citado em trabalhos científicos recentes sobre o tema de reconhecimento de emoções (CUI, 2018). São 4000 imagens de cenas do mundo real, divididas em 2162 imagens com os atores sorrindo e as outras 1838 imagens em posição neutra. Essas imagens faciais cobrem uma ampla gama de condições de iluminação, enquadramento, identidade pessoal e raça. A Figura 21 ilustra algumas imagens selecionadas aleatoriamente do banco de dados de expressões faciais Genki4k.

Figura 21 – Banco de imagens Genki4k

Fonte: adaptado de (CUI, 2018)

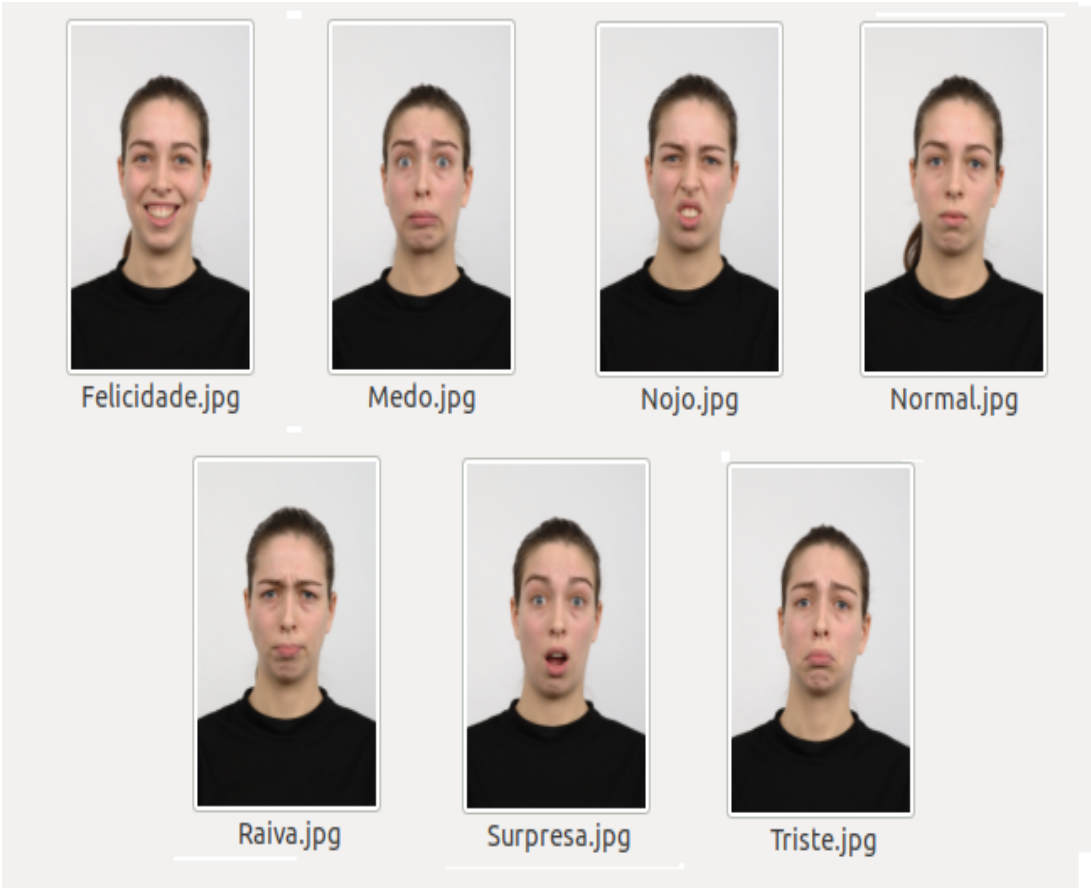
- **Labeled Faces in the Wild (LFW):** possui imagens extraídas da Internet para trabalhos relacionadas ao reconhecimento facial irrestrito em ambientes de campo. Foi desenvolvido por pesquisadores da University of Massachusetts e possui 13.233 imagens de 5.749 atores. Sua utilização neste trabalho está vinculada à detecção facial em imagens capturadas em ambientes não controlados.

- **Extended Cohn-Kanade Dataset(CK+):** publicado em 2000 (COHN; KANADE, 2010) tenta padronizar as expressões faciais segundo a referência dos FACS propostos por (EKMAN; FRIESEN, 1971). Possui imagens apenas em posição frontal e com a presença de todas as 7 emoções universais.

- **The Japanese Female Facial Expression (JAFFE) Database:** composto por 210 imagens com as sete expressões faciais universais, apresentadas por 10 mulheres de origem japonesa (LYONS, 2017) em posição frontal. As imagens foram obtidas pelo Departamento de Psicologia da Universidade de Kyushu, Japão.

- **Radboud Faces Database (RafD):** é um banco de dados de imagens com 67 atores (incluindo homens e mulheres caucasianos, crianças caucasianas de etnia europeia e homens marroquinos. O RaFD (LANGNER, 2010b) foi uma iniciativa do Behavioral Science Institute da Radboud University Nijmegen, localizada em Nijmegen (Holanda). Seguindo a metodologia dos FACS, treinou-se todos os atores para expressarem as seguintes emoções: raiva, nojo, medo, felicidade, tristeza, surpresa, desprezo e neutro. Cada emoção é apresentada com três direções distintas de olhar e em cinco ângulos diferentes na captura das imagens. Para o desenvolvimento deste trabalho, a emoção de desprezo, presente apenas neste banco de dados de expressões faciais será descartada por não pertencer ao grupo das emoções universais. A Figura 22 ilustra um dos atores do banco de dados de expressões faciais RafD.

Figura 22 – Amostra de um dos atores do banco de dados de expressões faciais RafD



Fonte: elaborada pelo autor

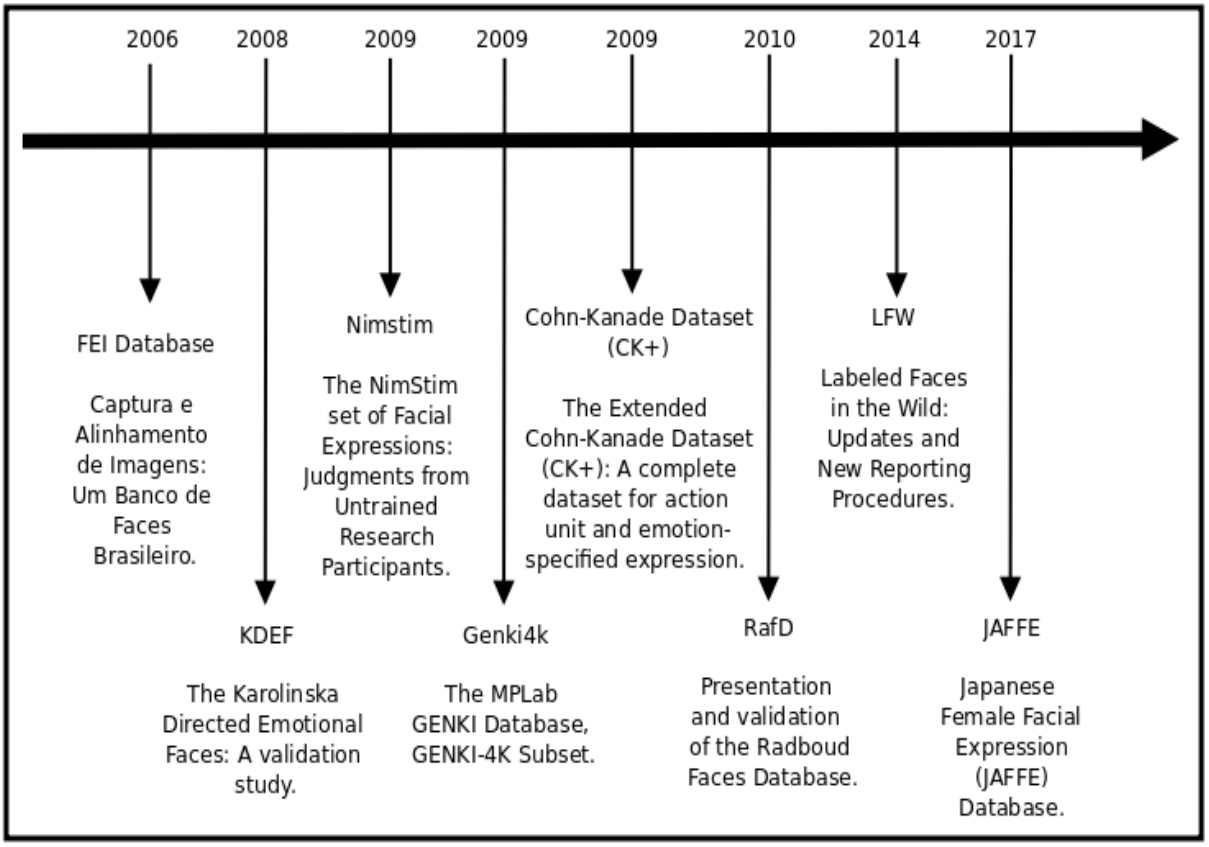
- **The Karolinska Directed Emotional Faces (KDEF):** composto por 4900

imagens com todas as emoções universais de setenta atores (GOELEVEN, 2008). Cada expressão é vista de cinco ângulos diferentes. O conjunto de imagens foi desenvolvido pelo Departamento de Neurociência Clínica, setor de Psicologia pelo Instituto Karolinska na Suécia.

- **The Nimstim set of Facial Expressions:** um banco de dados de expressões faciais muito conhecido na literatura médica (TOTTENHAM, 2009). Possui mais de 2.000 citações em trabalhos científicos, embora pouco utilizado para o reconhecimento de emoções. O Nimstim possui 672 imagens em posição frontal de 43 atores profissionais, 18 mulheres e 25 homens, com idades entre 21 e 30 anos. Todas as emoções universais estão presentes.

A Figura 23 apresenta a cronologia de divulgação dos bancos de dados de expressões faciais utilizados neste trabalho para treinar o método REGL, com exceção do LFW e Genki4k, responsáveis pelo treinamento do algoritmo de detecção facial AEHZ. Nota-se que na última década, diversos trabalhos foram desenvolvidas no campo da visão computacional e reconhecimento de emoções.

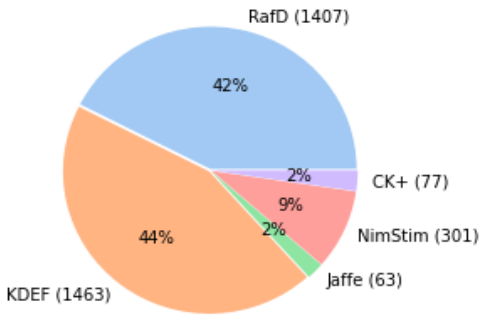
Figura 23 – Cronologia dos bancos de imagens utilizados



Fonte: elaborada pelo autor

Nesta dissertação, em virtude da baixa quantidade de amostras de atores, os resultados individuais para os bancos de dados de expressões faciais NimStim, Jaffe e CK+ não serão apresentados. Juntos eles representam apenas 13% dos dados analisados, conforme ilustra a Figura 24.

Figura 24 – Total de imagens por banco de dados



Fonte: elaborada pelo autor

4.2 Método REGL

O processamento e a manipulação de imagens têm por finalidade preparar os dados, inicialmente na forma de pixels, e convertê-los em estruturas mais simples, mas que carregam toda informação relevante em um espaço dimensional menor como, por exemplo, descritores de forma. A partir de uma estrutura enxuta e informativa, é possível interpretar e classificar os padrões, por meio de algoritmos de Aprendizado de Máquina, gerando o conhecimento efetivo.

A detecção de faces em uma imagem é a primeira etapa do processo de detecção de estruturas faciais, como o reconhecimento de pessoas ou a identificação das emoções. O método REGL manipula informações em etapa posterior à detecção da face em imagens, de forma que ferramentas de terceiros foram empregadas para essa finalidade. Atualmente, há um grande número de algoritmos capazes de reconhecer faces humanas em imagens. Em vista de possíveis diferenças nos resultados, realizou-se uma avaliação preliminar, comparando a performance de cada um dos principais algoritmos de detecção facial. A Subseção 4.2.1 apresenta os resultados comparativos entre os algoritmos de detecção facial presentes na literatura e introduz um novo método, mais otimizado e assertivo, denominado algoritmo AEHZ, para a utilização do método REGL. Em síntese, o algoritmo AEHZ proporcionou maior controle na aquisição das regiões de interesse faciais e consequentemente na performance dos resultados. As outras Subseções detalham as etapas do método REGL.

4.2.1 Detecção facial com AEHZ

A detecção facial é a primeira etapa para todos os sistemas computacionais que utilizam as características faciais como entrada de dados. Diminui consideravelmente o espaço de busca dos algoritmos, independente se o propósito for o reconhecimento de uma determinada pessoa ou a identificação de alguma característica específica, neste caso uma emoção manifestada na forma de expressão facial.

Os experimentos utilizaram três diferentes técnicas, implementadas nativamente por meio da biblioteca DLib, presente na linguagem de programação Python. São elas: Haar, HOG e Rede Neural Convolucional. As duas primeiras são implementações clássicas que utilizam os pixels como base do algoritmo. Já a Rede Neural Convolucional possui arquitetura de aprendizado profundo, onde abstrações de alto nível, como a detecção de objetos, são exploradas por meio das diversas camadas ocultas e de convolução do sistema.

Juntamente com as implementações nativas da linguagem de programação Python, durante o desenvolvimento deste trabalho um novo algoritmo de detecção facial foi construído, denominado Algoritmo de Equalização de Histograma e Zoom (AEHZ). Sua ideia principal sugere a otimização do processo de detecção facial em ambientes não controlados, mais próximos da nossa realidade. Sua estrutura utiliza o algoritmo HOG, juntamente com outras duas técnicas de processamento digital de imagens. São elas: a equalização do histograma, que normaliza o brilho e o contraste das imagens de entrada, diferenças que podem interferir no processamento e retornar informações inconsistentes e o zoom digital, que potencializa o algoritmo HOG para imagens de baixa resolução.

O algoritmo 1 detalha os passos necessários para a utilização da detecção facial com o AEHZ. Primeiramente uma imagem (M) é recebida na entrada de dados e a saída retorna uma outra imagem contendo apenas uma região facial válida. A primeira etapa processa o algoritmo HOG convencional (linha 3), caso não encontre uma região facial válida (linha 4) o histograma da imagem de entrada é equalizado (linha 5) e novamente o algoritmo HOG é executado (linha 6). Caso não encontre nenhuma face (linha 7) a resolução da imagem de entrada M é duplicada, em um processo semelhante ao zoom digital (linha 8). Finalmente, caso o algoritmo não encontre nenhuma região facial válida (linha 10) a saída retorna vazio (linha 11) ou uma imagem, na forma de Região de Interesse (linha 16) em caso de sucesso na detecção facial.

A justificativa para a ampliação da imagem em 2x (zoom digital) refere-se à dificuldade que o algoritmo HOG convencional possui para identificar objetos, neste caso faces humanas, em janelas com resolução menor que 80 x 80 pixels.

Algoritmo 1 - Detecção facial com AEHZ

```
   Entrada.:  Imagem M
   Saída...:  Imagem (ROI) da região facial detectada ou ∅
 1: INÍCIO
 2: Recebe uma imagem M
 3: faceDetectada ← Detectar Face em M com HOG
 4: SE faceDetectada ≠ ∅ ENTÃO
 5:     M ← Equalizar histograma M
 6:     faceDetectada ← Detectar Face em M com HOG
 7:     SE faceDetectada ≠ ∅ ENTÃO
 8:        M ← Ampliar 2X
 9:        faceDetectada ← Detectar Face em M com HOG
10:        SE faceDetectada ≠ ∅ ENTÃO
11:           RETORNA ∅
12:        FIM-SE
13:     FIM-SE
14: FIM-SE
15: ROI ← faceDetectada
16: RETORNA ROI
17: FIM
```

A Tabela 1 apresenta os resultados após o processamento dos algoritmos de detecção facial, de acordo com a quantidade efetivamente detectada em ambientes controlados. Os bancos de imagens utilizados neste experimento foram: Extended Cohn Kanade (CK+), Jaffe e RafD. As imagens utilizadas continham obrigatoriamente a presença de uma das sete expressões faciais universais e apenas uma face por imagem. A coluna 0 indica que houve detecção facial, na coluna +1 houve a detecção de mais de uma face na imagem e a coluna 1 apresenta uma face detectada na imagem.

Tabela 1 – Detecção facial em bancos de imagens com ambientes controlados

Banco	Dimensões	Total(Imagens)	Algoritmo	0	+1	1	t(s)	$\bar{t}$(s)
CK+	640 x 490	77	Haar	0	5	72	3,48	0,05
			CNN	0	0	77	239	3,10
			HOG	0	0	77	2,38	0,03
			AEHZ	0	0	77	2,38	0,03
Jaffe	256 x 256	210	Haar	0	1	209	4,64	0,02
			CNN	0	0	210	133,61	0,64
			HOG	0	0	210	1,42	0,01
			AEHZ	0	0	210	1,42	0,01
RafD	681 x 1024	4824	Haar	257	1476	3091	280,01	0,06
			CNN	0	1	4823	35938,8	7,45
			HOG	3	0	4821	328,17	0,06
			AEHZ	0	0	4824	328,19	0,06

Fonte: elaborada pelo autor

Pela análise dos resultados podemos destacar as seguintes conclusões:

- Quanto maior a resolução da imagem, maior o tempo de processamento;
- O algoritmo Haar apresentou uma grande quantidade de Falsos positivos (coluna +1 da tabela);
- Os algoritmos CNN e HOG encontraram praticamente todas as faces nos bancos de imagens, porém o custo computacional da CNN inviabiliza o processamento em tempo real (coluna t(s) - tempo total em segundos e coluna $\bar{t}$(s) - tempo médio para processamento de cada imagem);
- O algoritmo AEHZ, implementado pelo autor detectou todas as faces em todas as imagens com possibilidade de processamento em tempo real;

Os resultados com o algoritmo HOG e AEHZ são muito similares e não justificam a necessidade de uma nova implementação, porém os dados da tabela 1 foram extraídos em ambientes com condições controladas de iluminação e aquisição das imagens.

Um dos objetivos futuros deste trabalho é estender o método de reconhecimento de emoções para imagens em condições de campo e adquiridas em ambientes reais. A tabela 2 apresenta os resultados obtidos com o processamento dos bancos de imagem Genki4k e LFW, ambos com imagens adquiridas sem qualquer controle de iluminação e aquisição.

Tabela 2 – Detecção facial em bancos de imagens com ambientes NÃO controlados

Banco	**Dimensões**	**Total(Imagens)**	**Algoritmo**	**0**	**+1**	**1**	**t(s)**	$\bar{t}$**(s)**
Genki4k	179 x 192	4000	Haar	88	213	3699	64,89	0,02
			CNN	211	4	3785	1433,96	0,36
			HOG	278	5	3717	16,51	0,01
			AEHZ	57	8	3935	20,61	0,01
LFW DB	250 x 250	13233	Haar	31	1758	11444	253,01	0,02
			CNN	0	875	12358	8336,79	0,63
			HOG	91	629	12513	86,26	0,01
			AEHZ	38	629	12566	90,96	0,01

Fonte: elaborada pelo autor

Os resultados comprovam a eficiência do algoritmo AEHZ quando comparado às outras implementações. Com exceção do algoritmo Haar que novamente retornou altas taxas de Falso positivo (coluna +1 da tabela) todas os outros algoritmos conseguiram detectar a presença de mais de uma face em imagens que realmente tinham mais de uma pessoa, condição presente nos bancos de imagens Genki4k e LFW.

Em relação à performance na detecção facial o algoritmo AEHZ conseguiu superar em $3,73\%$ (150 imagens) os resultados do algoritmo CNN, sendo 70x mais rápido e em $5,45\%$ (218 imagens) o algoritmo HOG convencional. O tempo de processamento manteve-se constante nos dois experimentos com o banco de imagens Genki4k.

Já no banco de imagens LFW, o algoritmo AEHZ foi $1,57\%$ (208 imagens) mais eficiente que o algoritmo CNN, com tempo de processamento em torno de 90x mais rápido. Na comparação com o algoritmo HOG convencional o resultado foi $0,04\%$ maior (53 imagens). Não houve alteração perceptível no tempo de processamento.

4.2.2 Extração dos *landmarks* faciais

Conforme descrito no Capítulo 2, a capacidade de reconhecimento das emoções, por parte dos seres humanos está diretamente ligada com a identificação das diversas mudanças musculares que ocorrem na face. Essas variações precisam ser mensuradas para que seja possível a comparação entre os padrões. Contudo, o espaço de informações diretamente ligado aos pixels da imagem possui uma alta dimensionalidade, dificultando a detecção de padrões ou variações em um determinado padrão. Desse modo, a extração de *landmarks* faciais permite diminuir a dimensionalidade do problema de maneira significativa, restringindo a quantidade de variáveis para o total de coordenadas detectadas, que indicam estruturas faciais específicas. Cada *landmark* facial possui uma localização 2D no plano cartesiano, composta por um par de coordenadas. A coordenada x estabelece a localização no eixo horizontal e a coordenada y a localização no eixo vertical. Para o desenvolvimento deste trabalho utilizou-se um método de extração de 68 *landmarks* faciais, conforme descrito em (Kazemi; Sullivan, 2014), presentes em cinco regiões distintas e manipuladas em quase metade dos trabalhos da literatura, conforme levantamento realizado por (TESTA, 2019) e disponível no Capítulo 2, a saber: contorno do rosto, boca, nariz, olhos e sobrancelhas. A quantidade de citações, o processamento em tempo real e a capacidade de manipulação de imagens rotacionadas em até 45 graus, sem perda de informação, justificam a escolha do método de extração dos *landmarks* da biblioteca Dlib.

Portanto, a partir da extração dos *landmarks* faciais, o domínio do problema que antes compreendia todos os pixels da imagem, passa a conter apenas 136 variáveis quantitativas. O método utiliza um modelo de Aprendizado de Máquina treinado que acompanha as movimentações da face e fixa as coordenadas dos *landmarks*. O resultado é a redução da dimensionalidade do problema, sem perder as características responsáveis pela convergência dos algoritmos utilizados para a aquisição do conhecimento.

Por meio de sistemas automatizados, as informações dos *landmarks* faciais podem ser utilizadas para identificar pessoas ou expressões faciais. Tradicionalmente, as coordenadas dos *landmarks* são utilizadas para o cálculo de medidas de distância entre

regiões da face, as quais permitem identificar padrões e comparar com as informações armazenadas em bancos de dados (HUANG, 2009). Via de regra, a distância Euclidiana [1], entre os *landmarks* é a medida mais utilizada para o reconhecimento facial e para a classificação das expressões faciais(CUI, 2018; REVINA; EMMANUEL, 2018; TESTA, 2019). O uso de distância Euclidiana entre os *landmarks* para reconhecimento de padrões faciais produz resultados interessantes quando o objeto de análise está em posição frontal, obtido de maneira controlada e padronizada. Contudo, há também fatores negativos relacionados ao uso de distâncias. O primeiro aspecto é o tempo de processamento e o custo computacional necessários para o cálculo das distâncias. Ao processar a distância euclidiana com os 68 *landmarks* extraídos, chega-se à seguinte fórmula de combinação matemática $\frac{n!}{(n-m)!*m!} \rightarrow \frac{68!}{(68-2)!*2!} = 2278$. Portanto, o uso de distâncias exige um tempo de processamento considerável, limitando aplicações que exijam processamento em tempo real ou equipamentos menos sofisticados. Não por acaso, critérios adicionais de seleção de variáveis elegem e ordenam as distâncias faciais, segundo sua importância para a solução de problemas específicos (SALMAN, 2016).

Outra limitação, em relação ao uso de distâncias, é a variabilidade presente em imagens adquiridas em condições não controladas, com inúmeras fontes de ruído. Nesses casos, as medidas absolutas e relativas das distâncias são afetadas e o erro amostral não pode ser corrigido, o que interfere negativamente nos resultados. Como exemplo, o fator de escala do enquadramento gerado pelo distanciamento entre o equipamento de captura da imagem e o rosto do ator e o fator de rotação da face, tanto no eixo horizontal, quanto no vertical. Desse modo, os valores absolutos e relativos das distâncias podem ser afetados, comprometendo a eficácia dos dados quando utilizados para a detecção ou classificação de padrões faciais. Normalizações matemáticas, tais como a relativização dos valores de distância, segundo uma medida fixa (ex: distância entre os olhos), tendem a minimizar a variabilidade amostral relacionada à escala do enquadramento. Contudo, variações rotacionais da pose reduzem a eficácia desse tipo de normalização, já que o deslocamento rotacional altera a posição relativa dos *landmarks*, ou seja, muda a proporção entre as distâncias. Há ainda um viés adicional na classificação das expressões faciais. Variações morfológicas naturais entre as pessoas, tais como as produzidas por diferenças de idade, raça e de gênero, agregam variabilidade nas categorias das amostras. Essas variações de informações são de grande importância para a identificação de pessoas, mas dificultam a classificação das emoções por aumentar a zona de sobreposição entre diferentes categorias. Por exemplo, pessoas com bocas maiores tendem a ser classificadas mais facilmente como sorrindo, embora estejam em posição facial neutra. Esse efeito é reflexo das características anatômicas naturais diferentes, medidas segundo distâncias Euclidianas entre os *landmarks* bucais. Enquanto importantes para propósitos de identificação de pessoas, elas tendem a agregar ruído para o reconhecimento das emoções.

[1] Corresponde à menor distância entre dois pontos, ou seja, uma reta traçada entre eles.

Considerando os problemas de variabilidade amostral e sua interação com o uso de distâncias para o reconhecimento de expressões faciais, este trabalho introduz um método alternativo para contornar todas estas questões apresentadas. A solução encadeia três etapas principais, sendo: a) normalização das coordenadas dos *landmarks* (Seção 4.2.3); b) frontalização das coordenadas (Seção 4.2.5); c) normalização e extração de medidas de posição relativa da expressão na imagem em relação ao rosto neutro do ator (Seção 4.2.6). Essas etapas antecedem à construção dos algoritmos fundamentados em Aprendizado de Máquina (Seção 4.2.7), ou seja, dizem respeito ao pre-processamento dos dados, visando uma melhor padronização das informações de entrada. Os detalhes são apresentados a seguir.

4.2.3 Normalização das coordenadas

Para minimizar os efeitos da variabilidade de escala, utilizou-se a normalização das coordenadas pelo valor mínimo e máximo (min-max) de cada eixo, segundo a Equação 4.1 para as coordenadas horizontais no eixo x e a Equação 4.2 para as coordenadas verticais no eixo y, com custo computacional constante O(1), tal que:

$$\widehat{x}_i = \frac{x_i - min(x)}{max(x) - min(x)}, \tag{4.1}$$

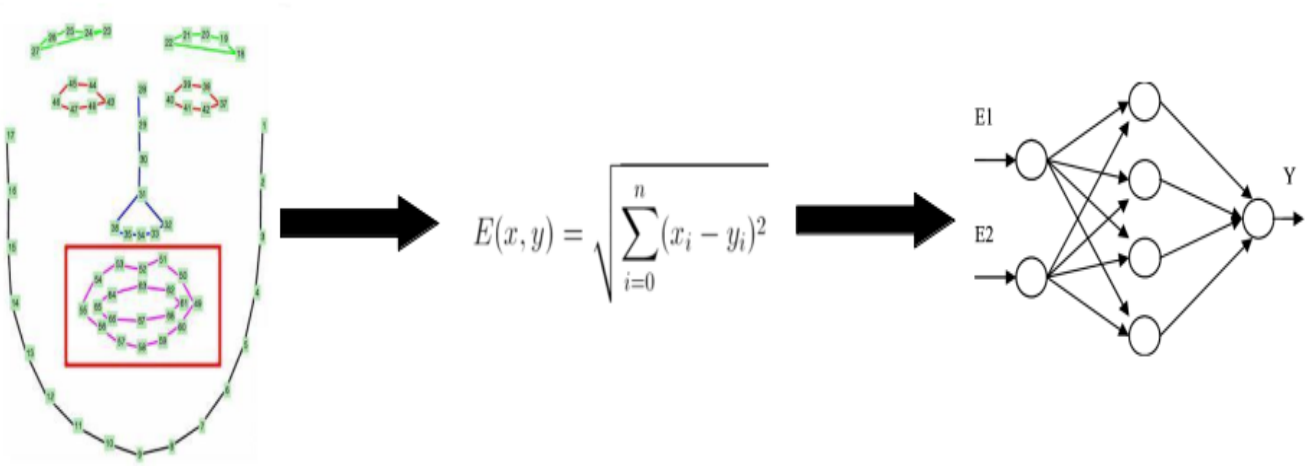
$$\widehat{y}_i = \frac{y_i - min(y)}{max(y) - min(y)}. \tag{4.2}$$

Dessa forma, todas as coordenadas x e y assumem valores entre zero e um, minimizando o efeito de escala. É importante ressaltar que essa transformação não altera as proporções entre as coordenadas. Isso significa que o rosto reescalado é semelhante ao original, permitindo seu uso para propósitos de reconhecimento facial. Assim, a metodologia aqui apresentada demanda que o ator avaliado na imagem, seja previamente conhecido.

4.2.4 Detecção de Sorriso com min-max

Em (CUI, 2018), os autores desenvolveram um método para detecção de sorriso utilizando distância Euclidiana, entre os 20 *landmarks* da região da boca, como atributo para treinar um algoritmo de aprendizado de máquina, fundamentado em redes neurais, denominado $ExtremeLearningMachine(ELM)$. O resultado obtido foi de 93,40% utilizando o Banco de imagens Genki4k, conforme ilustra a Figura 25.

Figura 25 – Sistema de detecção de sorriso utilizando ELM



Fonte: elaborada pelo autor

A induçao do algoritmo ELM combinou as distância euclidianas de todas as coordenadas da boca, totalizando 190 atributos (variáveis).

Uma nova metodologia, utilizando a normalização por min-max de cada coordenada, possui a propriedade de eliminar a variabilidade de escala, relacionada com a proximidade em relação ao equipamento de aquisição da imagem, além das coordenadas serem menos variantes ao fator de rotação, elemento presente e constante ao trabalharmos com faces humanas.

O algoritmo 2 detalha os passos necessários para a detecção do sorriso, utilizando apenas as coordenadas normalizadas. A entrada do algoritmo recebe todas as imagens do Banco Genki4k. A saída corresponde a um modelo de Aprendizado de Máquina, replicável e reutilizável em outras imagens com a presença de faces humanas.

O algoritmo percorre todas as imagens do Banco Genki4k (linha 2) e para cada uma (linha 3) inicia o processamento de imagens para realizar a detecção facial (linha 4). Caso haja sucesso (linha 5) o processamento para mapeamento dos *landmarks* inicia, com um modelo treinado capaz de identificar 68 pontos de marcação. Em seguida os *landmarks* são extraídos (linha 6) e normalizados, tanto pelo eixo x (linha 7) quanto pelo eixo y (linha 8). Finalmente, apenas as coordenadas da boca (linha 9) ficam armazenadas no vetor de dados (linha 10). A última etapa (linha 14) constrói o Modelo de Aprendizado de Máquina (MLM) responsável pela identificação de sorriso utilizando imagens com a presença de faces humanas.

Algoritmo 2 - Sistema de Detecção de Sorriso

```
   Entrada.:  Banco de Imagens Genki4k (G)
   Saída...:  Modelo de Aprendizado de Máquina (MLM)
 1: INÍCIO
 2: ENQUANTO houver imagem em G
 3:     Carregar Imagem G(I)
 4:     face ← Detectar Face G(I)
 5:     SE face ≠ ∅ ENTÃO
 6:        coordenadas ← Extração dos 68 landmarks (face)
 7:        coordenadas ← Normalizar coordenadas x (Eq. 1)
 8:        coordenadas ← Normalizar coordenadas y (Eq. 2)
 9:        coordenadas ← Extrair as 40 coordenadas da boca
10:        vetorCoordenadas ← coordenadas
11:     FIM-SE
12: FIM-ENQUANTO
13: MLM ← Classificador(vetorCoordenadas)
14: RETORNA MLM
15: FIM
```

4.2.5 Frontalização das coordenadas

O processo de frontalização busca reduzir as variações de rotação do rosto, na horizontal e na vertical, em relação ao plano da imagem, centralizando a pose de maneira frontal. Há atualmente duas principais técnicas de frontalização: a primeira fundamentada na renderização de toda a imagem (HASSNER, 2015) e a segunda que utiliza apenas as coordenadas dos *landmarks* (ZHAO, 2018) para simular uma pose frontal.

Tabela 3 – Reconhecimento de emoções - Frontalização pela aparência

Emoção	Precisão	Recall	Medida-F
Neutro	0,561	0,581	0,571
Medo	0,501	0,459	0,479
Raiva	0,578	0,614	0,595
Felicidade	0,882	0,888	0,885
Nojo	0,583	0,640	0,610
Surpresa	0,717	0,743	0,730
Tristeza	0,522	0,441	0,478

Fonte: elaborada pelo autor

A Tabela 3 apresenta os resultados obtidos no reconhecimento de Emoções faciais após a utilização do método de Frontalização pela aparência. Os resultados não foram satisfatórios e apresentaram taxas de acerto inferiores aos experimentos sem utilizar esta

técnica, além de exigir um processamento gráfico adicional para renderizar as imagens e extrair os *landmarks* faciais.

Por outro lado, a frontalização que utiliza as coordenadas dos *landmarks* pode ser feita em tempo real e resulta em ganhos importantes para o reconhecimento de expressões faciais. Assim, optou-se pela aplicação da técnica de frontalização das coordenadas (ZHAO, 2018) que simula uma vista frontal da região facial, por meio de pesos numéricos, multiplicados em cada um dos eixos de cada coordenada, ao custo computacional constante O(1). Após a aplicação do método de frontalização, todas as coordenadas assumem uma configuração compatível às de um rosto em posição frontal e centralizado, minimizando a variabilidade amostral relacionada com variações de pose no momento de aquisição da imagem. O processo visa, sobretudo, tornar o método consistente também para imagens adquiridas em condições de campo, ou seja, não controladas.

4.2.6 Normalização pela face em repouso (Delta)

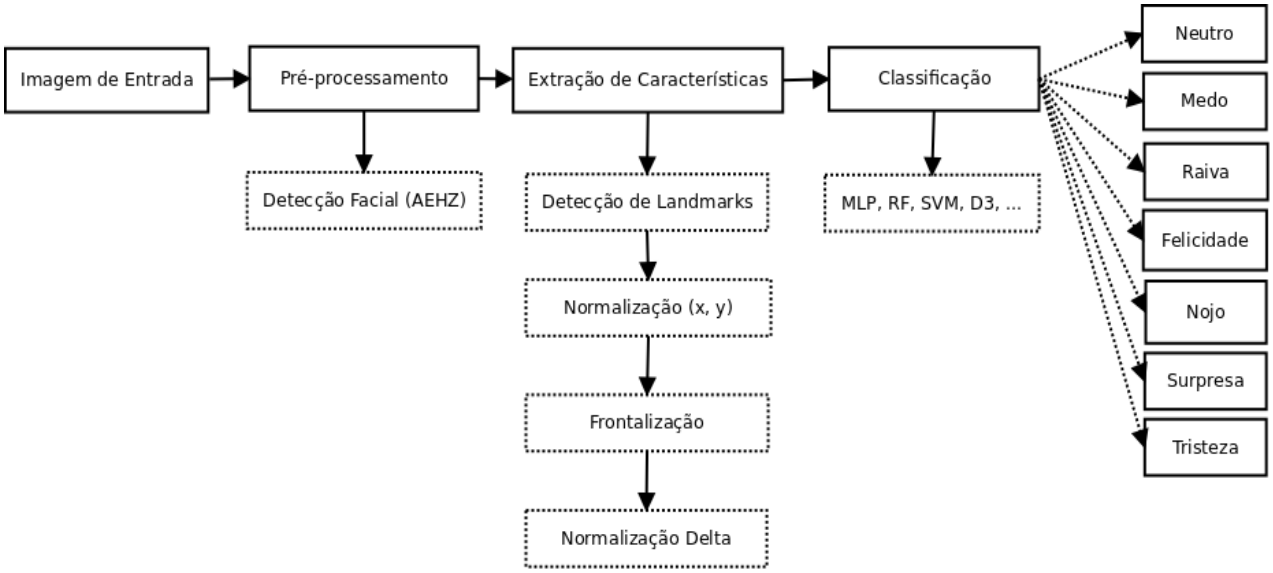
Ressalta-se que o método REGL tem por objetivo reconhecer as variações das expressões faciais e possui diversas aplicações práticas. Dentre essas aplicações, destaca-se o reconhecimento de emoções em imagens ou detecção de movimentos faciais específicos para fins médicos e terapêuticos. As variações anatômicas entre diferentes atores podem ser uma causa importante de ruído, dificultando a detecção de padrões morfo-geométricos da face quando se busca o reconhecimento de emoções. Diante disso, propõe-se uma mudança de abordagem. Ao invés de avaliar os padrões morfo-geométricos estáticos, como usualmente se faz na detecção de expressões faciais (REVINA; EMMANUEL, 2018; CUI, 2018; MOHAN, 2021), propõe-se avaliar a variação dos padrões quando uma expressão facial é executada. Essa abordagem exige um padrão facial neutro, ou seja, uma imagem A frontal do rosto em repouso (expressão facial neutra), como referência. A imagem A deve ser submetida à detecção facial, extração de *landmarks*, normalização e frontalização das coordenadas, para que então essas 136 coordenadas sirvam de referência e formem o vetor $\overrightarrow{A}$. Uma segunda imagem B, objeto de avaliação, é submetida ao mesmo processo, gerando outras 136 coordenadas e formando o vetor $\overrightarrow{B}$. Tem-se então a criação do vetor final de informação, $\overrightarrow{\Delta}_{AB}$, de tamanho 136 variáveis, tal que cada coordenada i é obtida por $\Delta_{AB}i = Ai - Bi$. Portanto, $\overrightarrow{\Delta}_{AB}$ contém a informação da variação relativa das coordenadas frontalizadas de B em relação às coordenadas do rosto em repouso A. Dado que a expressão de interesse em B é comparada sempre com a expressão do mesmo ator em posição neutra, previamente rotulada e conhecida em A, problemas de variabilidade anatômica são minimizados. Como vantagem, é possível não apenas classificar um padrão geométrico estático, por exemplo, o padrão geométrico que caracteriza o sorriso, como também medir **quantitativamente a deformação produzida por este movimento**.

4.2.7 Classificação de padrões

A última etapa de um sistema de reconhecimento de emoções (SRE), como qualquer outro que utiliza algoritmos de Aprendizado de Máquina, necessita da indução dos dados processados em um classificador. Obviamente, a organização e as características do classificador dependem do objetivo e do contexto do problema estudado. Este trabalho concentra-se no reconhecimento de emoções faciais, alicerçado nas variações que ocorrem na face em relação à posição de um determinado ator em repouso, o que oferece uma gama de contextos a serem explorados.

A Figura 26 lista o fluxo de processamento de um Sistema de Reconhecimento de Emoções clássico. Na etapa de "extração de características", o processamento adapta-se ao método REGL, com a normalização das coordenadas por min-max, frontalização facial e normalização pela face do ator em repouso (emoção neutra). A entrada do sistema recebe uma imagem com a presença de uma face humana. A saída retorna uma das sete emoções universais: medo, raiva, tristeza, felicidade, surpresa e nojo, além da expressão neutra que serve de base para o reconhecimento das outras emoções.

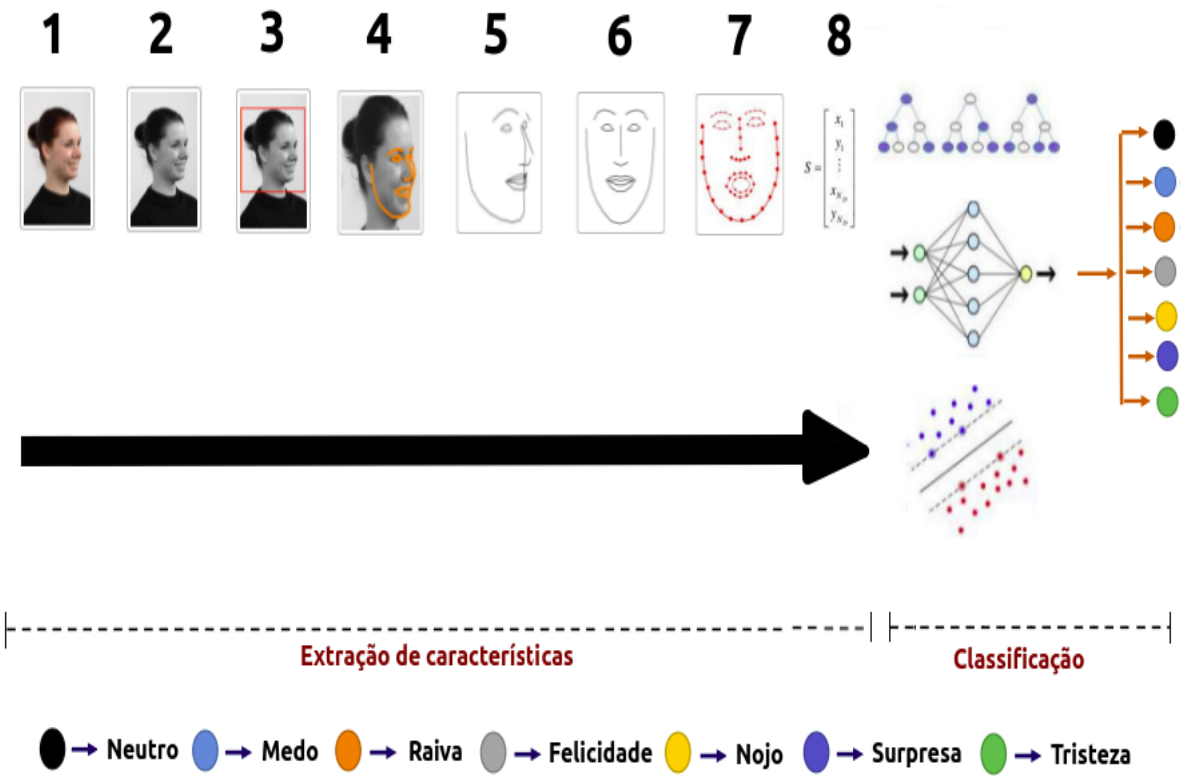
Figura 26 – Sistema de reconhecimento de emoções clássico adaptado ao método REGL



Fonte: adaptado de (REVINA; EMMANUEL, 2018)

A Figura 27 apresenta uma visão geral das técnicas utilizadas e o respectivo sequenciamento de etapas necessárias para efetuar o reconhecimento de emoções com o método REGL. Cada etapa busca diminuir a variabilidade dos pixels e das coordenadas dos *landmarks* e proporcionar uma melhoria contínua e gradativa dos resultados.

Figura 27 – Método REGL de reconhecimento de emoções



Fonte: elaborada pelo autor

Em resumo, os passos do método REGL, utilizado para o reconhecimento de emoções são:

1. Equalização de Histograma - redução da variabilidade radiométrica (brilho e contraste);
2. Escala de Cinza - redução dos três canais de cores para apenas 1 canal;
3. Detecção Facial - redução da área de busca pela delimitação da Região de Interesse;
4. Extração dos *landmarks* - mudança na dimensionalidade dos dados, os pixels são trocados pelas coordenadas bidimensionais dos *landmarks*. O domínio, a partir de agora, compreende apenas 136 variáveis (68 coordenadas para o eixo x e outras 68 para o eixo y);
5. Normalização min-max das coordenadas: redução do fator de escala, presente na diferença de proximidade entre o ator e o equipamento de captura. Inovação no processo de reconhecimento de emoções;
6. Frontalização - redução da variabilidade geométrica ocasionada pela diversas rotações, tanto no eixo x quanto no eixo y. As coordenadas se movimentam para simular uma posição frontal;

7. Normalização pela face do ator (padrão Delta) - elimina as variações anatômicas;

8. Vetor de coordenadas: concatenação das coordenadas em um vetor para auxiliar na indução dos algoritmos de Aprendizado de Máquina. A dimensão final que o vetor possui é de 1 linha para cada ator com 136 colunas, ou seja, 68 colunas para as coordenadas x e 68 colunas para as coordenadas y.

O Algoritmo 3 detalha as etapas técnicas do reconhecimento de emoções faciais do método REGL. A entrada recebe todas as imagens dos bancos de dados de expressões faciais Cohn-Kanade, RafD, NimStim, KDEF e Jaffe e a saída retorna um modelo de Aprendizado de Máquina reutilizável para o reconhecimento de emoções.

O Algoritmo percorre todas as imagens de entrada G (linha 2) e para cada uma (linha 3) inicia o processamento de imagens e a detecção facial (linha 4). Caso haja sucesso (linha 5) o processamento para redução da variabilidade nos dados se inicia: extração dos *landmarks* (linha 6), normalização pelo min-max das coordenadas x (linha 7) e coordenadas y (linha 8), frontalização das coordenadas (linha 9) e normalização pela face do ator (padrão delta) (linha 10). Finalmente, as coordenadas são inseridas no vetor de informações (linha 10). Após o processamento de todas as imagens, o vetor final passa pelo processo de classificação e criação do modelo de Aprendizado de Máquina (linha 14).

Algoritmo 3 - Reconhecimento de Emoções faciais com o método REGL

```
   Entrada.:  Bancos de Imagens CK+, Jaffe, KDEF, NimStim e RafD (G)
   Saída...:  Modelo de Aprendizado de Máquina (MLM)

 1: INÍCIO
 2: ENQUANTO houver imagem em G
 3:     Carregar Imagem G(I)
 4:     face ← Detectar Face G(I)
 5:     SE face ≠ ∅ ENTÃO
 6:        coordenadas ← Extração dos 68 landmarks (face)
 7:        coordenadas ← Normalizar coordenadas x (Eq. 4.1)
 8:        coordenadas ← Normalizar coordenadas y (Eq. 4.2)
 9:        coordenadas ← Frontalizar coordenadas
10:        coordenadas ← Normalizar pela face do ator (Δ)
11:        vetorCoordenadas ← coordenadas
12:     FIM-SE
13: FIM-ENQUANTO
14: MLM ← Classificador(vetorCoordenadas)
15: RETORNA MLM
16: FIM
```

Considerações finais

Neste capítulo, apresentou-se todos os bancos de dados de expressões faciais utilizados neste trabalho, além de uma descrição detalhada do método REGL, utilizado no reconhecimentos de emoções humanas. Toda a etapa de normalização das coordenadas dos *landmarks* foi descrita, incluindo a normalização min-max, a frontalização das coordenadas e a normalização pela face em repouso dos atores (Delta). Todas essas etapas são importantes para a redução de variabilidade nos dados e responsáveis pelo sucesso da implementação do método REGL. No próximo capítulo, serão apresentados os experimentos e a discussão dos resultados obtidos com o processamento do método REGL.

5

Experimentos e Resultados

Este capítulo apresenta os experimentos realizados, os resultados obtidos e as discussões sobre o processamento do método REGL. O encadeamento das etapas culmina na criação de um vetor de características, composto por um conjunto de 136 coordenadas, após realizadas todas as etapas de normalização e transforma-se na entrada dos algoritmos de Aprendizado de Máquina SVM, MLP, Florestas Aleatórias, Árvores Extras e Árvores de Decisão. Os algoritmos avaliaram a qualidade das características extraídas, bem como a classificação das emoções.

Para a apresentação dos resultados, utilizou-se tabelas e gráficos com as porcentagens de acerto e erro. As tabelas descrevem comparativamente os resultados de cada uma das técnicas empregadas, em função de cada um dos métodos de reconhecimento de padrões para cada uma das emoções, além do tempo de processamento.

Os experimentos com as melhores porcentagens de acerto foram selecionados para uma descrição mais detalhada e foram identificados nas tabelas por meio dos itens destacados em negrito. Para uma visualização mais abrangente dos resultados, foram utilizadas matrizes de confusão, além de diversas informações estatísticas, tais como: acurácia, taxa de verdadeiro positivo TP, taxa de falso positivo FP, especificidade, precisão e Medida-F. A partir das taxas de verdadeiro positivo TP e falso positivo FP, criou-se gráficos ROC para representar a capacidade discriminativa dos classificadores, isto é, a capacidade de classificar corretamente uma determinada emoção. Nos gráficos ROC, os pontos mais próximos a (0,1), ou seja, maior TP e menor FP, representam uma classificação mais consistente e generalista.

A Seção 5.1 apresenta os experimentos práticos que avaliaram o desempenho e acurácia do método REGL. Os experimentos seguem uma ordem crescente de complexidade, permitindo avaliar a eficácia do método quando diferentes níveis de dificuldade são apresentados. Na Seção 5.2 são apresentados os resultados da detecção de sorriso com o processamento dos bancos de imagens Genki4k, em ambientes não controlados e FEI Database, para ambientes controlados. Na Seção 5.3, todos os resultados obtidos com o

método REGL são apresentados, juntamente com as discussões e particularidades de cada implementação.

5.1 Experimentos

Em um primeiro momento, o método REGL foi experimentado com coordenadas brutas em ambientes controlados. Esse experimento foi denominado Exp001. Em seguida, no Exp002, utilizou-se a normalização das coordenadas com min-max. O próximo experimento, chamado de Exp003, utilizou a frontalização das coordenadas. Por último, a normalização pela face do ator em repouso, também chamada de normalização Delta, finalizou os experimentos com o método REGL no Exp004.

A Tabela 4 apresenta os experimentos realizados com o método REGL, utilizando os bancos de dados de expressões faciais Cohn-Kanade Dataset (CK+), RafD, NimStim, KFEF e Jaffe. Os experimentos utilizaram 5 diferentes algoritmos de Aprendizado de Máquina, detalhados na Tabela 5. Nas árvores de decisão, a profundidade máxima dos nós utilizados foi 4. Valores abaixo dessa configuração não conseguem generalizar as categorias. Valores acima não geram ganhos nos resultados. Tanto as florestas aleatórias, quanto as árvores extras, utilizaram as mesmas configurações de profundidade das árvores de decisão, além da quantidade de árvores com o parâmetro 1000. Valores abaixo dessa quantidade não convergem para as categorias de forma consistente. Valores acima aumentam exponencialmente o tempo de execução, inviabilizando o processamento. O algoritmo MLP possui três configurações principais: número de iterações em 3000 épocas, taxa de aprendizado 0,001 e o algoritmo Adam para ativar a rede neural. Essa configuração produz resultados proporcionais ao tempo de processamento e a acurácia do modelo. O algoritmo SVM emprega um kernel de base radial para separar as categorias e o parâmetro C = 5 para penalizar as classificações incorretas.

Tabela 4 – Relação dos experimentos para reconhecimento de emoções faciais

Experimento	Situação dos landmarks	Ambiente
Exp001	Coordenadas Brutas	Controlado
Exp002	Coordenadas Normalizadas pelo min-max	Controlado
Exp003	Frontalização das Coordenadas	Controlado
Exp004	Coordenadas Normalizadas pela face do ator	Controlado

Fonte: elaborada pelo autor

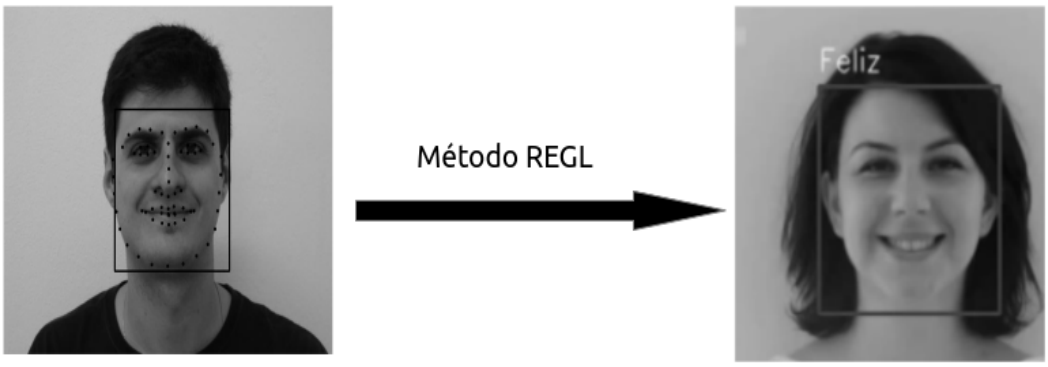
Tabela 5 – Configuração dos algoritmos de AM para reconhecimento de emoções

Algoritmo	**Configuração**
D3	Profundidade máxima = 4
RFC	Número de árvores = 1000 e Profundidade máxima = 4
Extra Trees	Número de árvores = 1000 e Profundidade máxima = 4
MLP	solver = adam, iterações = 3000 e aprendizagem = 0,001
SVM	Função de base radial ou linear e C(penalidade) = 5

Fonte: elaborada pelo autor

A classificação finaliza o processo de reconhecimento de emoções. O método REGL recebe uma imagem de entrada e retorna qual a emoção predominante dentre as sete universais, relatadas por (EKMAN; FRIESEN, 1971). A Figura 28 ilustra o resultado prático de um sistema computacional que utiliza o método REGL. A entrada de dados (esquerda) possui uma imagem com uma região facial válida. A saída (direita) retorna a emoção predominante na imagem de entrada, neste caso a emoção de felicidade.

Figura 28 – Exemplo prático de utilização do método REGL



Fonte: elaborada pelo autor

Para o desenvolvimento deste trabalho, todos os algoritmos foram implementados e processados em um computador portátil com processador Intel(R) Core i5 de oitava geração, modelo 8265U de 3,9 GHz, com 8 Gb de memória RAM, HD SSD de 128 Gb e placa de vídeo integrada Intel UHD. Na parte de software executando o sistema operacional Linux Ubuntu 18.04 e a linguagem de programação Python 3.6.9.

Os melhores parâmetros dos algoritmos de Aprendizado de Máquina foram testados exaustivamente no servidor de HPC (computação de alto desempenho) aguia4 da Universidade de São Paulo. O cluster possui por 128 servidores físicos com 20 cores e 512 Gb de RAM. O processador é o Intel(R) Xeon CPU E7-2870 de 2.40GHz com um Filesystem de 256 Tb para arquivos temporários.

5.2 Resultados para a detecção de sorriso

Para avaliar o desempenho da normalização min-max, utilizando as coordenadas bidimensionais normalizadas dos *landmarks*, a Tabela 6, adaptada de (CUI, 2018), compara os resultados com alguns métodos bastante utilizados na literatura, tanto em relação à dimensionalidade quanto na classificação das imagens da base de dados Genki4k. Todos os métodos utilizam apenas recursos individuais e não combinam outras técnicas para criar o vetor de características, utilizado na indução dos algoritmos de Aprendizado de Máquina.

Tabela 6 – Dimensionalidade dos métodos - Banco de imagens Genki4k

Método	Dimensão	Classificador	Acurácia(%)
Gabor	23,040	SVM	89,68 ± 0,62
LBP	944	SVM	87,10 ± 0,76
GSS	576	Ada+SVM	91,11 ± 0,47
Pixel Comparison	500	AdaBoost	89,70 ± 0,45
Pair-Wise Distance	190	ELM (RNN)	93,42 ± 1,46
Coord. min-max	**40**	**SVM**	**94,30 ± 0,01**
Coord. min-max	**40**	**MLP**	**94,43 ± 0,01**

Fonte: adaptada de (CUI, 2018)

Os resultados indicam que a normalização por min-max produziu uma dimensionalidade menor, além de resultados mais assertivos que os encontrados na literatura, corroborando com os trabalhos de (BELLMAN, 1961) no sentido de que quanto menor for a dimensão dos dados inseridos em um algoritmo de Aprendizado de Máquina, maior a possibilidade dos resultados serem mais consistentes e acurados.

A Tabela 7 apresenta os resultados com cinco diferentes tipos de algoritmos de Aprendizado de Máquina para a detecção de sorriso (emoção de felicidade), no Banco de Imagens Genki4k.

Tabela 7 – Genki4k - Detecção de sorriso (felicidade)

Classificador	Acurácia(%)	Medida-F(%)	TP	TN	FP	FN
KNN	92,53 ± 0.016	91,07	1970	1587	162	125
D3	93,24 ± 0.011	92,40	1990	1592	142	120
RFC	93,34 ± 0.012	92,61	1970	1620	162	92
SVM	94,30 ± 0.001	93,67	2005	1620	127	92
MLP	**94,43 ± 0.001**	**93,81**	**2009**	**1621**	**123**	**91**

Fonte: elaborada pelo autor

Os resultados obtidos com os classificadores SVM e MLP foram superiores aos de (CUI, 2018), com processamento em tempo real.

Para comprovar a eficiência da normalização min-max, os modelos de Aprendizado de Máquina, treinados com o banco de imagens Genki4k, foram utilizados e testados no banco de imagens FEI Database.

Por se tratar de um conjunto de imagens adquiridas em ambiente controlado, dentro do laboratório e com padronização da iluminação, esperava-se que os resultados fossem superiores. A Tabela 8 apresenta os resultados obtidos e confirma melhores taxas de acurácia na detecção de sorriso.

Tabela 8 – FEI Database - Detecção de sorriso (felicidade)

Classificador	Acurácia(%)	Medida-F(%)	TP	TN	FP	FN
Perceptron	92,00 ± 0.066	92,88	170	198	30	2
D3	94,00 ± 0.032	93,75	185	189	15	11
MLP	94,75 ± 0.031	94,80	188	191	12	9
KNN	95,25 ± 0.031	95,32	187	194	13	6
RFC	96,00 ± 0.026	96,07	188	196	12	4
SVM	**96,25 ± 0.029**	**96,29**	**191**	**194**	**9**	**6**

Fonte: elaborada pelo autor

O Banco de Imagens Genki4k possui um número desproporcional de atores sorrindo e em posição neutra, na simetria de 55% e 45% respectivamente. A medida de acurácia por si só não é suficiente para quantificar a eficácia dos algoritmos de Aprendizado de Máquina testados. Como métrica complementar, utilizou-se a Medida-F, além da quantidade de imagens classificadas de forma correta (VP - Verdadeiro Positivo e TN - Verdadeiro Negativo) e de forma incorreta (FP - Falso Positivo e FN - Falso Negativo).

Os resultados obtidos mostram que as coordenadas normalizadas da região da boca são decisivas para identificar se uma pessoa está sorrindo ou não, com taxas de acerto superiores a 94% nos algoritmos MLP e SVM e maiores que 93% nos algoritmos Random Forest e Árvore de Decisão, superando assim os resultados encontrados na literatura para a detecção de sorriso.

5.3 Resultados com o Método REGL

Inicialmente, apresenta-se os resultados do método REGL, por banco de imagem processado e a comparação com os trabalhos da literatura. Em seguida, todos os bancos de imagem foram agrupados e processados juntos. Os resultados dessa Seção avaliaram o desempenho do método REGL, na medida em que as técnicas de normalização foram encadeadas.

5.3.1 Resultados RafD (imagens frontais)

A Tabela 9 apresenta os resultados do reconhecimento de emoções utilizando o banco de dados de expressões faciais RafD, apenas com imagens frontais.

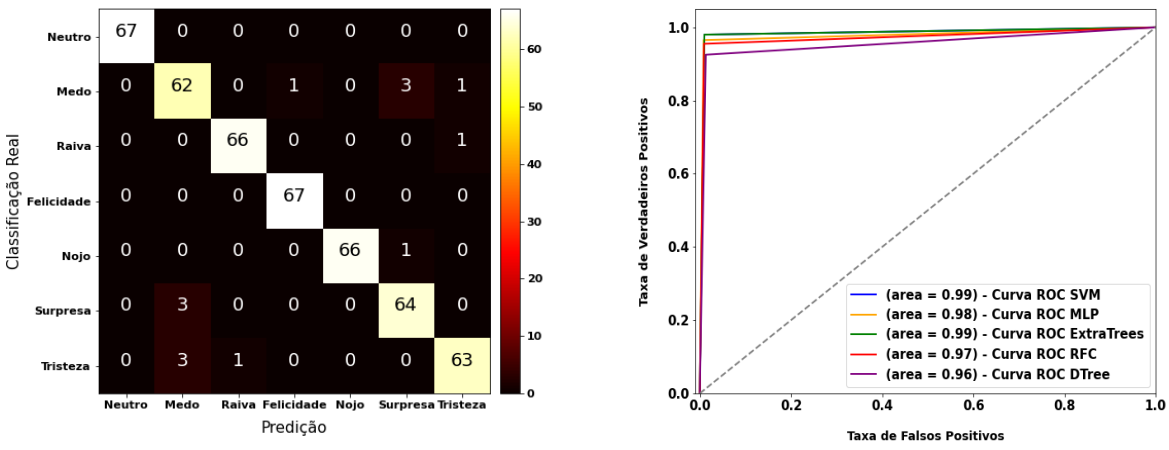
Tabela 9 – Resultado do método REGL - RafD (imagens frontais)

Classificador	Acurácia	Sensibilidade	Medida-F	TP	FP	t(s)
SVM	**0,970**	**0,970**	**0,970**	**455**	**14**	**0,21**
MLP	0,967	0,966	0,967	453	16	13,28
Extra Tree	0,964	0,965	0,964	452	17	10,66
Random Forest	0,965	0,964	0,964	452	17	35,44
Árvore Dec.	0,899	0,900	0,899	422	47	0,5

Fonte: elaborada pelo autor

Os resultados obtidos demonstram uma acurácia acima de 96% com quatro algoritmos de Aprendizado de Máquina: SVM, MLP, Extra Tree e Random Forest. O algoritmo Árvore de Decisão, mesmo com uma implementação mais simples, produziu um resultado em torno de 90%. A Figura 29(a) ilustra a matriz de confusão do resultado obtido com o algoritmo SVM, que alcançou a melhor performance em todas as métricas. A Figura 29(b) realiza a comparação do gráfico ROC de todos os algoritmos implementados. Constata-se que todos aproximaram-se do ponto (0,1), ou seja, altas taxas de acerto e baixa taxa de erro no reconhecimento de emoções no banco de dados de expressões faciais RafD.

Figura 29 – Resultados - Banco de imagens RafD - imagens frontais



(a) (b)

Fonte: elaborada pelo autor

5.3.2 Resultados RafD (frontais e rotacionadas)

A Tabela 10 apresenta os resultados do reconhecimento de emoções utilizando o banco de dados de expressões faciais RafD com todas as imagens processadas. O método de extração dos *landmarks*, utilizado neste trabalho, funciona em uma área de rotação máxima de até 45 graus, tanto para o lado direito quanto para o lado esquerdo, sem perder desempenho.

Tabela 10 – Resultado do método REGL - RafD (imagens frontais e rotacionadas)

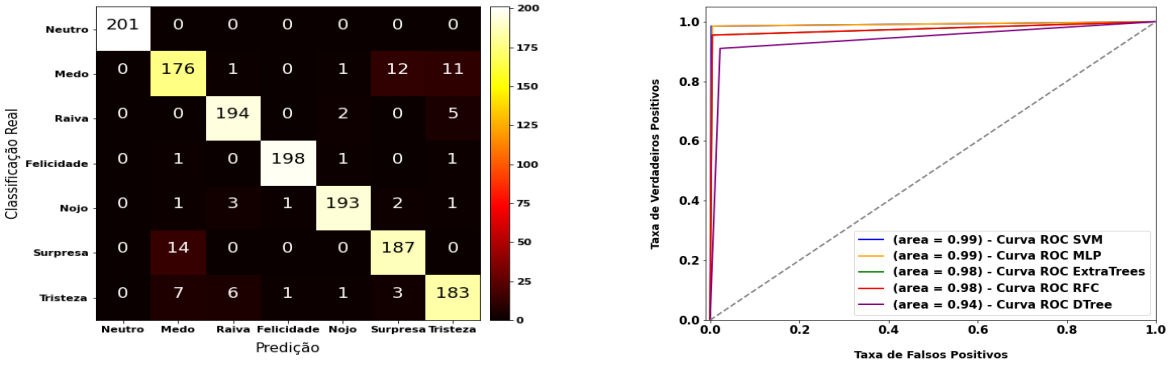
Classificador	Acurácia	Sensibilidade	Medida-F	TP	FP	t(s)
SVM	**0,954**	**0,954**	**0,954**	**1342**	**65**	**0,71**
MLP	0,947	0,945	0,946	1332	75	13,95
Extra Tree	0,950	0,949	0,950	1326	81	20,29
Random Forest	0,943	0,942	0,942	1322	85	124,13
Árvore Dec.	0,900	0,899	0,899	1252	144	0,69

Fonte: elaborada pelo autor

Com o processamento das imagens rotacionadas do banco RafD, os resultados foram ligeiramente inferiores, em torno de $1,2\%$ quando comparados com os resultados do processamento utilizando apenas imagens frontais. A melhor performance foi de $95,4\%$, obtida com o algoritmo SVM. A utilização da técnica de frontalização de coordenadas tornou-se eficaz ao possibilitar ganhos no processo de reconhecimento de emoções, expandindo a utilização do método REGL para imagens rotacionadas, posições essas que são frequentes em ambientes reais de estudo.

A Figura 30 ilustra a matriz de confusão resultante do processamento do algoritmo MLP (a) e o gráfico ROC de todos os algoritmos utilizados neste experimento (b). Os resultados indicam um acurácia próxima de 95% e uma região de erro entre as emoções de medo e surpresa.

Figura 30 – Resultados - Banco de imagens RafD - imagens rotacionadas



(a) (b)

Fonte: imagem elaborada pelo autor

De todos os bancos de dados de expressões faciais utilizados neste trabalho, apenas o RafD cita que possui modelagem e aquisição das imagens de acordo com os FACS propostos por (EKMAN; FRIESEN, 1971). Entende-se como FACS um mapeamento dos movimentos faciais, responsáveis pela manifestação de uma determinada emoção. Com isso, os resultados do método REGL são bastante promissores no sentido de automatizar o reconhecimento de emoções faciais, sempre que a identificação do indivíduo alvo não seja um requisito.

Em (AOUAYEB, 2021), os autores utilizaram uma CNN com bloco de compreensão e excitação para realizar o Reconhecimento de Emoções. O resultado obtido com o banco de dados de expressões faciais RafD foi de $87,22\%$.

Embora as implementações sejam diferentes, a CNN utiliza os conceitos do algoritmo de Aprendizado de Máquina MLP, utilizado neste trabalho e que produziu um resultado de $94,7\%$, ou seja, uma acurácia de $7,48\%$ mais assertiva com a utilização das mesmas imagens.

5.3.3 Resultados KDEF

A Tabela 11 apresenta os resultados do reconhecimento de emoções utilizando o banco de dados de expressões faciais KDEF, com todas as imagens processadas.

Tabela 11 – Resultado do método REGL - KDEF (imagens frontais e rotacionadas)

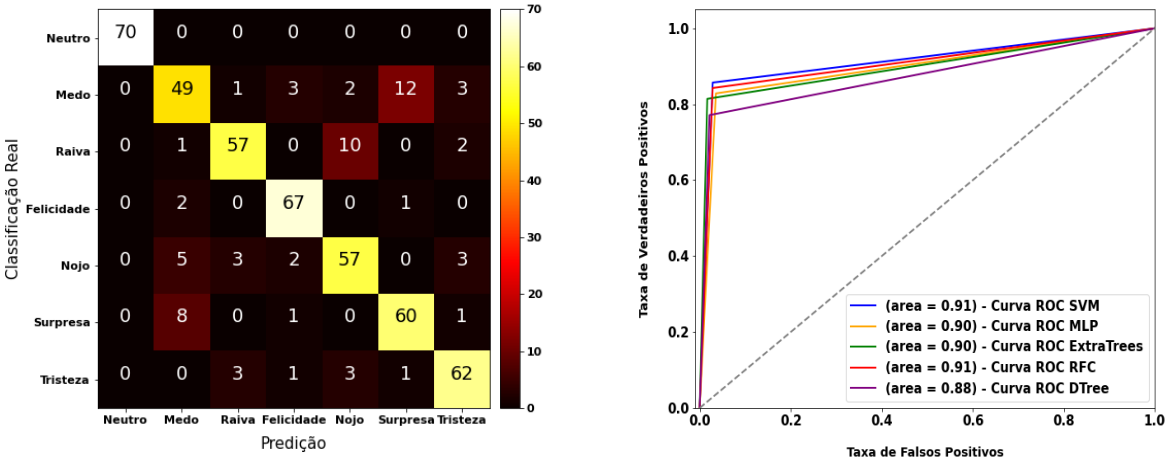
Classificador	Acurácia	Sensibilidade	Medida-F	TP	FP	t(s)
SVM	**0,868**	**0,869**	**0,867**	**1281**	**182**	**0,22**
MLP	0,840	0,841	0,840	1228	235	7,91
Extra Tree	0,861	0,861	0,860	1268	195	12,85
Random Forest	0,861	0,861	0,861	1269	194	44,07
Árvore Dec.	0,804	0,800	0,801	1179	284	0,28

Fonte: elaborada pelo autor

A melhor performance no processamento do banco de imagens KDEF foi do algoritmo SVM, em torno de $86,8\%$. Os resultados foram inferiores aos obtidos com o banco de dados de expressão facial RafD. Um possível diagnóstico para essa diferença pode estar na presença de atores de uma única raça (branca) na composição dos dados de treinamento e teste dos modelos de Aprendizado de Máquina, além das imagens não serem compatíveis com os FACS.

A Figura 31(a) ilustra a matriz de confusão resultante do processamento do algoritmo Árvores Extras e a Figura 31(b) o gráfico ROC de todos os algoritmos utilizados neste experimento. Os resultados foram inferiores aos obtidos com o banco de dados de expressões faciais RafD, inclusive em relação ao gráfico ROC.

Figura 31 – Resultados - Banco de imagens KDEF



(a) (b)

Fonte: elaborada pelo autor

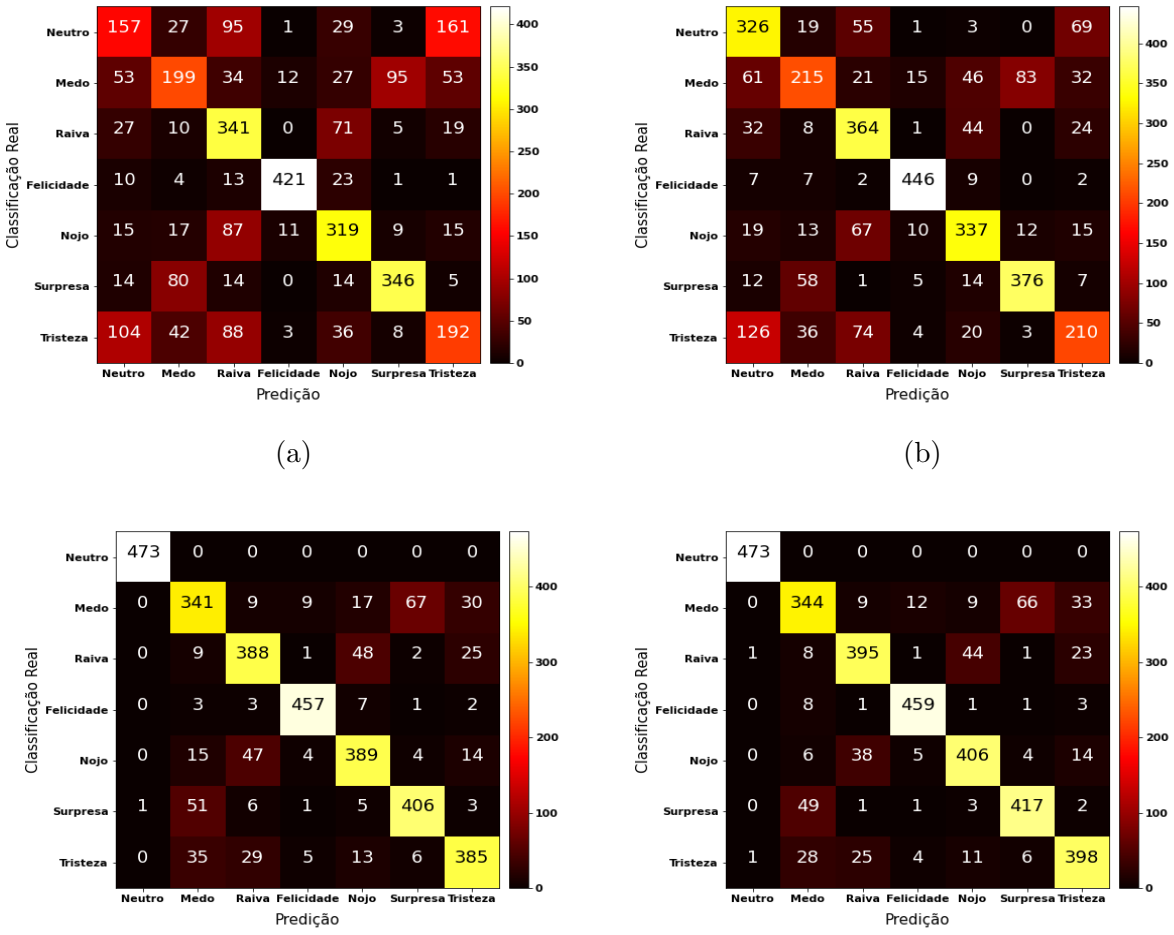
Na Seção 5.4, agrupou-se todos os resultados dos experimentos com os bancos de expressões faciais, de acordo com a utilização das técnicas de normalização e frontalização das coordenadas.

5.4 Reconhecimento de Emoções com REGL

O reconhecimento de emoções, por parte dos seres humanos, é um processo de comparação das alterações morfométricas na face e que possibilita a diferenciação entre as diversas classes possíveis. Aplicando o contexto das FACS para o reconhecimento artificial, utilizando máquinas e computadores para esta tarefa, torna-se necessária a normalização das coordenadas dos *landmarks*.

A Figura 32 apresenta os resultados do reconhecimento de emoções, utilizando o algoritmo de Aprendizado de Máquina SVM, conforme a ordem das etapas do método REGL. A Figura 31(a) ilustra a matriz de confusão do resultado com as coordenadas brutas, na Figura 31(b) as coordenadas estão normalizadas com min-max, na Figura 31(c) utilizou-se a frontalização e na Figura 31(d) o encadeamento de todo o método REGL.

Figura 32 – Reconhecimento de emoções - Resultados do método REGL - SVM



Fonte: elaborada pelo autor

A Tabela 12 e a Figura 33 apresentam os resultados do reconhecimento de emoções, por meio do encadeamento das etapas do método REGL. Pode-se destacar que a sucessão das técnicas de normalização das coordenadas foram decisivas para a evolução e a assertividade, independente do algoritmo utilizado para realizar a classificação.

Tabela 12 – Reconhecimento de emoções - evolução do método REGL (acurácia)

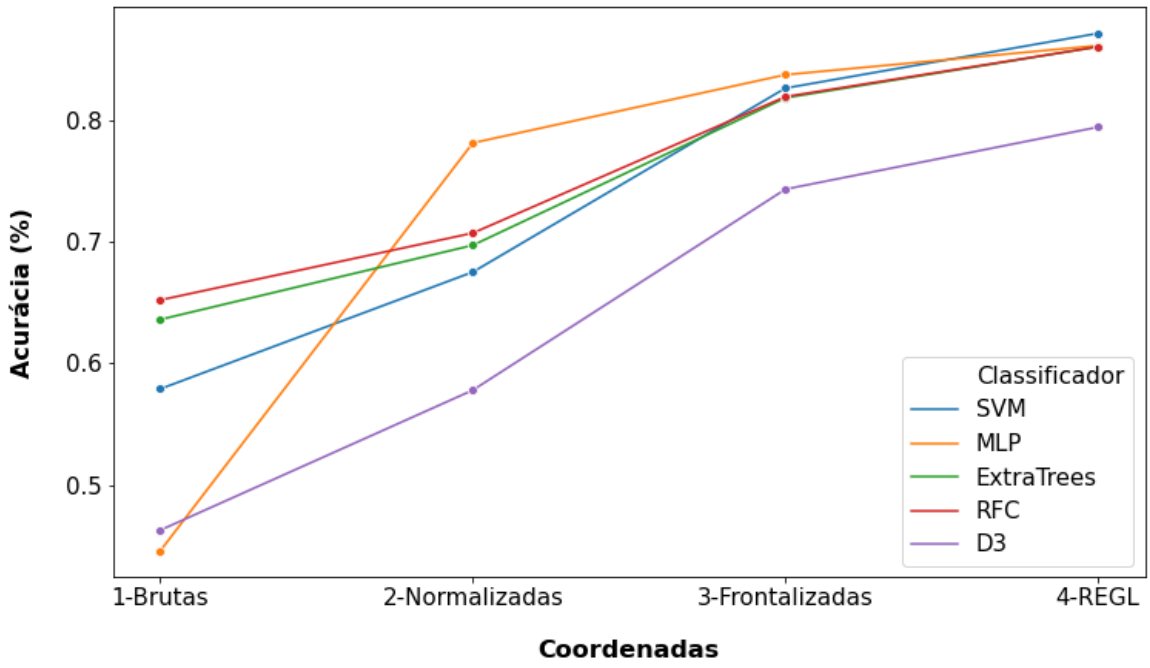
Algoritmo	Coord Brutas	Coord Normal	Coord Frontal	REGL
SVM	0,596	0,675	0,826	**0,871**
MLP	0,446	**0,781**	**0,837**	0,861
Árv. Extras	0,636	0,697	0,818	0,860
Flor. Aleatórias	**0,652**	0,707	0,819	0,860
Árvore de Decisão	0,463	0,578	0,743	0,794

Fonte: elaborada pelo autor

A melhor performance do método REGL foi obtida com a utilização do algoritmo SVM, um resultado de $87,1\%$ e com tempo de processamento de apenas $4,33$ segundos, lembrando que nos experimentos utilizou-se a validação cruzada com 10 *folds* de treinamento.

Em comparação com o trabalho de (EKMAN; FRIESEN, 1971), que obteve $77,04\%$ nos experimentos, sugere-se que as emoções também são universais para as máquinas e seu reconhecimento pode ser realizado de forma artificial.

Figura 33 – Evolução da acurácia - método REGL



Fonte: elaborada pelo autor

Outro aspecto importante para ser analisado em relação ao método REGL é o tempo de processamento dos algoritmos em relação ao encadeamento das etapas de normalização. De acordo com a Tabela 13 e a Figura 34, os dados são consistentes e comprovam que as etapas do método REGL diminuem o tempo processamento de todos dos algoritmos experimentados.

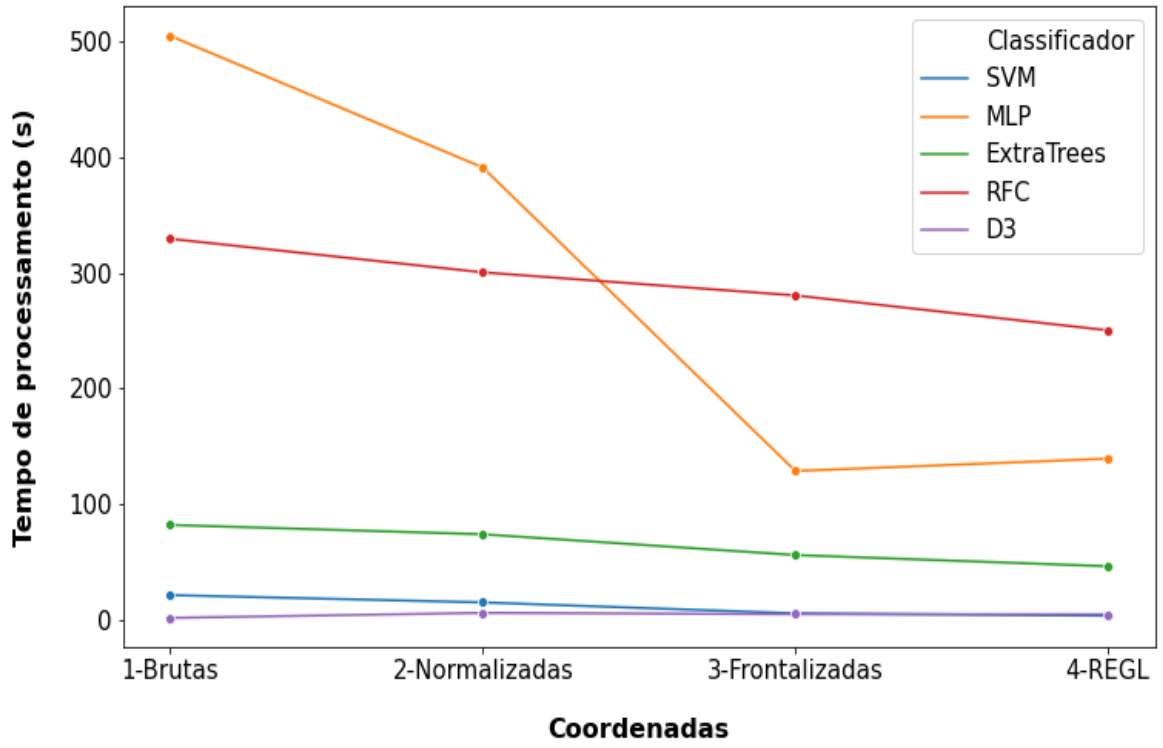
Portanto, a redução da variabilidade amostral das coordenadas, proporcionadas pela normalização min-max, frontalização e normalização pela face do ator contribuem para o aumento da performance dos algoritmos de Aprendizado de Máquina, responsáveis pelo reconhecimento das emoções humanas e otimizam o tempo de processamento necessário em cada etapa.

Tabela 13 – Reconhecimento de emoções - Tempo de processamento (segundos)

Algoritmo	Coord Brutas	Coord Normal	Coord Frontal	REGL
SVM	22,11	15,82	6,37	4,33
MLP	504,11	390,61	129,04	139,72
Árv. Extras	82,58	74,53	56,64	46,96
Flor. Aleatórias	329,19	300,29	280,39	250,33
Árvore de Decisão	6,79	5,66	5,58	2,46

Fonte: elaborada pelo autor

Figura 34 – Tempo de processamento - método REGL

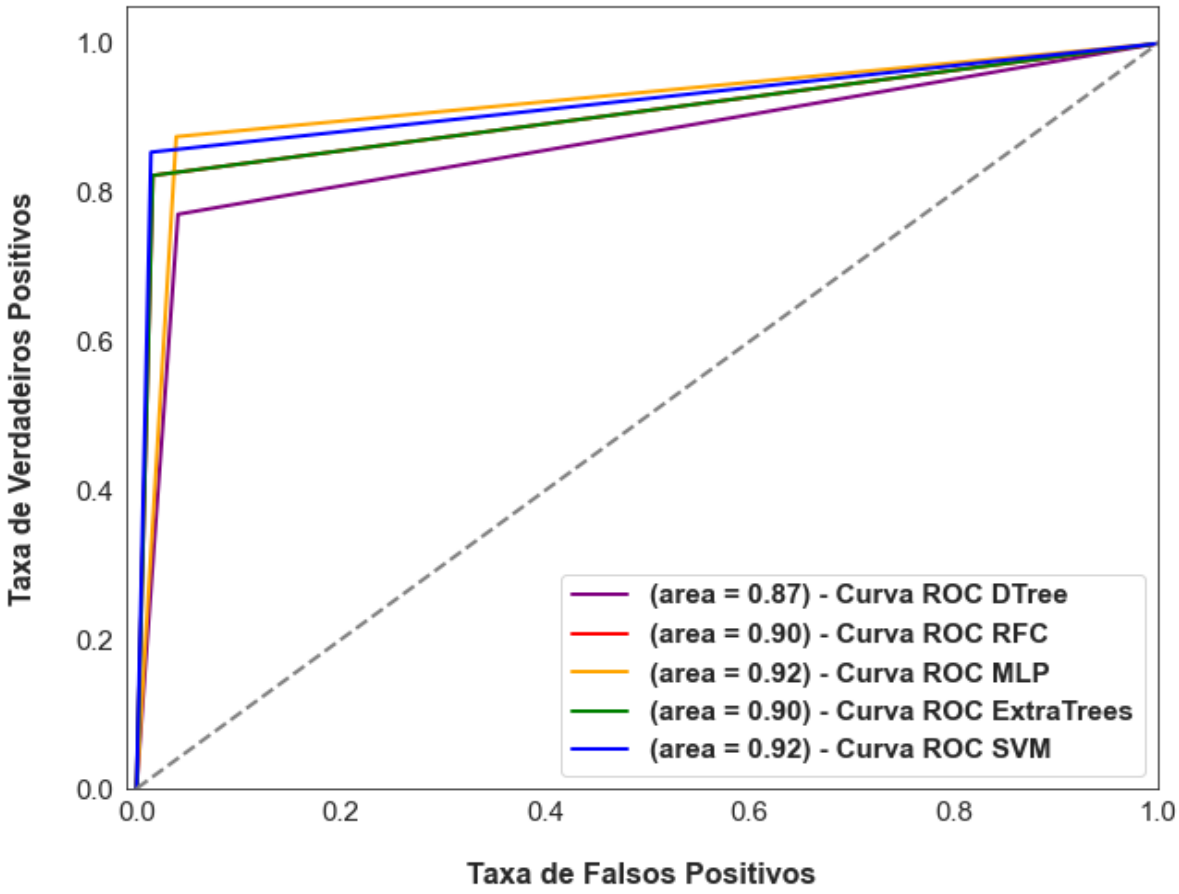


Fonte: elaborada pelo autor

A Figura 35 apresenta comparativamente o desempenho dos diversos algoritmos utilizados no processo de reconhecimento de emoções, após todas as etapas de redução de variabilidade implementadas, por meio de um gráfico ROC (*Receiver Operation Characteristics*). A linha diagonal tracejada representa o desempenho de um classificador aleatório e serve de referência para a avaliação dos demais. Os pontos no espaço ROC acima da diagonal representam uma melhor classificação do que os pontos abaixo da diagonal. Um classificador perfeito é aquele que produz como resultado um ponto próximo a $(0, 1)$ e uma área próxima de 1. Isto significa que este classificador tem uma taxa nula (igual a zero)

de falsos positivos, e uma taxa de 100% de acerto de verdadeiros positivos. Analisando a Figura 35, o gráfico ROC dos algoritmos SVM e MLP produziram resultados mais assertivos que os demais.

Figura 35 – Curva ROC - Reconhecimento de emoções com o método REGL



Fonte: elaborada pelo autor

Em todos os experimentos realizados, a emoção de felicidade foi facilmente identificada, com taxas de acurácia acima de 95%, mesmo nos experimentos com as coordenadas brutas. Portanto, a utilização do método REGL possui grande potencial de utilização para identificação automática e artificial de sorriso em imagens digitais.

Após a realização dos experimentos, constatou-se que os sentimentos de medo e de surpresa obtiveram as menores taxas de acerto e performance, independente das técnicas utilizadas, conforme ilustra a Tabela 14, após o processamento do método REGL com todas as etapas de normalização realizadas.

Tabela 14 – Reconhecimento de emoções - Método REGL - SVM

Emoção	Acurácia	Sensibilidade	Medida-F
Neutro	0,996	0,999	0,998
Medo	0,777	0,727	0,751
Raiva	0,842	0,835	0,839
Felicidade	0,952	0,970	0,961
Nojo	0,857	0,858	0,857
Surpresa	0,841	0,882	0,862
Tristeza	0,842	0,841	0,841

Fonte: elaborada pelo autor

(EKMAN; FRIESEN, 1971) relata, na discussão do seu trabalho, a dificuldade encontrada por todos os entrevistados em diferenciar pessoas com medo e pessoas surpresas. A conclusão indica que essas duas emoções são fortemente correlacionadas e que a emoção de medo é uma complementação e continuação da emoção de surpresa. Portanto, em virtude da semelhança geométrica dessas duas emoções, elas formam um grupo artificial

Essa informação motivou a construção de um novo classificador, utilizando o método REGL, com o agrupamento das emoções de medo e de surpresa. Os resultados são apresentados na Tabela 15.

Tabela 15 – Método REGL - SVM (agrupamento medo-surpresa)

Emoção	Acurácia	Sensibilidade	Medida-F
Neutro	0,994	0,999	0,997
Raiva	0,846	0,835	0,841
Felicidade	0,951	0,970	0,960
Nojo	0,860	0,864	0,862
Tristeza	0,855	0,840	0,847
Medo-Surpresa	0,933	0,934	0,933

Fonte: elaborada pelo autor

O resultado obtido, utilizando o classificador de Aprendizado de Máquina SVM como parâmetro comum, foi de $91,0\%$ na taxa de acurácia do modelo, ou seja, uma performance $3,81\%$ maior que o melhor resultado até este momento. Com isso comprova-se que, tanto para os seres humanos, quanto artificialmente, as emoções de medo e de surpresa possuem uma forte relação de dependência entre si.

Para finalizar os experimentos com o método REGL, um segundo classificador, apenas para as emoções de medo e surpresa foi construído, na tentativa de reclassificá-las e verificar o resultado com esta nova abordagem. Os resultados com a utilização de dois classificadores foram praticamente idênticos aos resultados obtidos com o classificador SVM, com uma margem de $0,32\%$ de diferença, conforme descrito na Tabela 16.

Tabela 16 – Método REGL - SVM (reclassificação medo-surpresa)

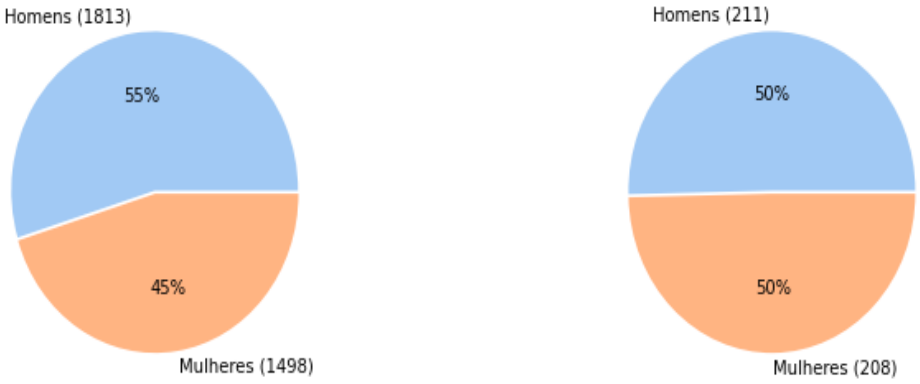
Emoção	Acurácia	Sensibilidade	Medida-F
Neutro	0,996	1,000	0,998
Medo	0,774	0,771	0,744
Raiva	0,837	0,833	0,835
Felicidade	0,952	0,966	0,959
Nojo	0,858	0,856	0,857
Tristeza	0,828	0,886	0,856
Surpresa	0,840	0,833	0,837

Fonte: elaborada pelo autor

Um detalhe importante sobre a implementação do método REGL foi a união de todas as bases de dados estudadas, tanto em relação ao gênero quanto em relação à raça dos atores, com o objetivo de proporcionar uma maior variedade de condições nas quais o método REGL pudesse ser testado. Na literatura especializada sobre o reconhecimento artificial de emoções (REVINA; EMMANUEL, 2018; TESTA, 2019), apresenta-se os resultados individualizados por banco de dados processado. Por consequência, o método REGL mostrou-se mais consistente entre as diferentes bancos de imagens, demonstrando ser menos susceptível e dependente do conjunto de dados utilizados para o treinamento dos classificadores do que seus antecessores.

Em relação ao gênero, a Figura 36(a) ilustra o percentual de atores presentes nos dados processados, já a Figura 36(b) ilustra o percentual de erro por gênero no item (b). Pela análise dos resultados, constata-se que o método REGL é invariante ao gênero.

Figura 36 – Erro por gênero do método REGL



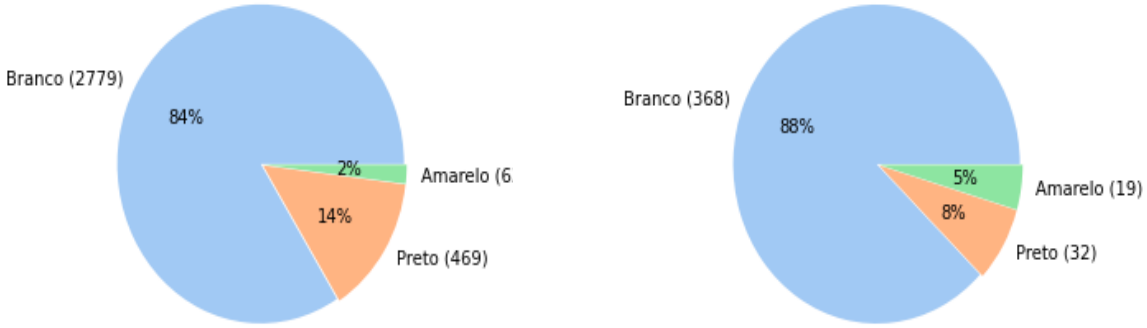
(a) (b)

Fonte: elaborada pelo autor

Em relação à raça, a Figura 37(a) ilustra o percentual de atores presentes nos

dados processados e a Figura 37(b) ilustra o percentual de erro por raça. Pela análise dos resultados, constata-se que a raça negra obteve uma performance melhor no reconhecimento de emoções com o método REGL, seguida pela reça branca e depois a amarela, embora a amostra de atores seja bem inferior para o último grupo.

Figura 37 – Erro por raça do método REGL



(a) (b)

Fonte: elaborada pelo autor

Considerações finais

Neste capítulo, foram apresentados os experimentos e os resultados obtidos com o método REGL na análise e reconhecimento artificial das emoções humanas. Os resultados foram descritos em função do encadeamento das técnicas de normalização utilizadas pelo método desenvolvido. Para a apresentação dos resultados, foram utilizados dados estatísticos de porcentagens de acerto e erro, além de diversas medidas comumente utilizadas em reconhecimento de padrões. Os gráficos ROC e as matrizes de confusão auxiliaram na visualização dos resultados. Os resultados obtidos demonstraram grande potencial de aplicação do método REGL em sistemas e aplicativos que utilizam dados faciais, contribuindo como uma ferramenta adicional para análise das emoções humanas.

6

Conclusão

As emoções proporcionam nosso primeiro meio de comunicação não verbal desenvolvido ao longo da vidas. Com elas, os humanos são capazes de interagir com as outras pessoas e com o meio ambiente no qual estão inseridos. Essa interação é possível porque os seres humanos, quase sempre, traduzem suas emoções em movimentos físicos detectáveis, tais como as expressões faciais, que são fundamentais para as interações sociais.

Apesar de ser um mecanismo trivial, facilmente reconhecido por todos os seres humanos, o reconhecimento de emoções é uma habilidade desafiadora para as máquinas e os computadores.

Consequentemente, dentro do contexto de interação homem-máquina, o objetivo deste trabalho foi desenvolver um método de reconhecimento artificial de emoções, chamado REGL, construído para extrair e analisar as características morfométricas da região facial, similar ao método utilizado pelos seres humanos.

Diversas técnicas de processamento digital de imagens e métodos estatísticos foram encadeados para avaliar e reduzir a variabilidade nas imagens utilizadas. Dentre elas, pode-se destacar a normalização das coordenadas por min-max, capaz de otimizar os efeitos do fator de escala, a frontalização, responsável pela redução dos efeitos ocasionados pela rotação da face e a normalização delta, que utiliza a face do próprio ator em posição neutra para identificar as outras emoções, com isso minimiza-se os efeitos das variações anatômicas e raciais.

Os resultados dos experimentos comprovaram a eficiência do método REGL na classificação e reconhecimento artificial das emoções humanas, produzindo como resultado final uma taxa de acerto acima de 90% para todos os bancos de dados de expressões faciais processados de forma agrupada. Em relação à detecção de sorriso obteve-se como resultado final uma taxa de acerto próxima de 95%, superando os trabalhos da literatura. Ressalta-se que essa junção de fontes de dados heterogêneos também produz variabilidade, relacionadas com questões raciais e que interfere nos resultados finais.

Outro aspecto identificado com os experimentos foi a dificuldade de dissociação entre as emoções de medo e surpresa, tanto para os seres humanos, conforme relatado por (EKMAN; FRIESEN, 1971), quanto no reconhecimento artificial, comprovado pelos resultados dos experimentos com a união dessas duas classes de emoções. Além disso, é importante destacar que o método REGL é invariante ao gênero e com tempo de processamento bem abaixo de outros métodos e técnicas disponíveis na literatura (REVINA; EMMANUEL, 2018). Como exemplo, pode-se citar a utilização do algoritmo de Aprendizado de Máquina SVM, com tempo de processamento em tempo real.

Neste trabalho, foi demonstrado que as técnicas de processamento digital de imagens e métodos estatísticos são importantes para o sucesso do reconhecimento artificial de emoções humanas. Os desafios estabelecidos ao longo do desenvolvimento do método REGL estão associados com a multidisciplinaridade das áreas envolvidas. Como consequência, o trabalho produziu resultados importantes para as áreas envolvidas e reforça o comprometimento da área da computação no progresso das pesquisas na ciência contemporânea, principalmente em tecnologias assistivas.

6.1 Contribuições

Para desenvolvimento do método REGL, diversas técnicas de normalização de dados foram estudadas e implementadas, com o objetivo de otimizar, fundamentar e permitir a sua criação. O reconhecimento artificial de emoções é multidisciplinar e as contribuições científicas consolidam todas áreas envolvidas, entre elas: processamento digital de imagens, estatística, inteligência artificial e visão computacional. As principais contribuições extraídas deste trabalho foram:

- Criação e implementação do algoritmo AEHZ de detecção facial. Os resultados otimizaram a performance do Histograma de Gradientes Orientados.
- Tradução da técnica de Frontalização por aparência da linguagem de programação Matlab para Python 3.
- Criação da técnica de normalização min-max nas coordenadas dos *landmarks* faciais, reduzindo o efeito do fator de escala nas imagens.
- Criação do método para detecção de sorriso, que utiliza as coordenadas normalizadas por min-max dos *landmarks*, com resultados superiores aos da literatura.
- Criação e implementação do método REGL para reconhecimento artificial das emoções humanas.

6.2 Limitações do Método REGL

A principal dificuldade envolvendo o método REGL está associada com a limitação da extração dos *landmarks* faciais. A rotação da face é um elemento fundamental e interfere negativamente na inserção dos pontos de marcação faciais, ou seja, quanto mais rotacionada a imagem, menor a possibilidade de identificação dos *landmarks*. Estudos realizados com a biblioteca DLib, utilizada neste trabalho, indicam que em uma área de rotação de até 45 graus, tanto para a direita quanto para a esquerda, não existe perda de desempenho na marcação dos *landmarks*. Ressalta-se que o método REGL processa informações em apenas duas dimensões.

Outra dificuldade durante o desenvolvimento desta dissertação foi a busca por de bases de dados em ambientes não controlados, para avaliar a eficiência e assertividade do método REGL. A única base encontrada foi o banco de imagens Genki4k, embora este possua apenas as emoções neutra e de felicidade.

6.3 Trabalhos Futuros

O reconhecimento de emoções humanas é uma área da Inteligência Artificial com grande potencial de exploração. A multidisciplinaridade deste trabalho permite a sua continuidade em diversas linhas de pesquisa e investigações, entre elas:

- Analisar performance do método REGL com uma quantidade maior de coordenadas dos *landmarks*;
- Analisar performance do método REGL com coordenadas tridimensionais dos *landmarks*;
- Analisar performance do método REGL em outras bases de dados, preferencialmente com atores de raça negra e amarela e em ambientes não controlados;
- Avaliar a performance do método REGL separado por estruturas faciais;
- Avaliar a performance do método REGL em vídeo;
- Testar os resultados do reconhecimento facial com coordenadas normalizadas pelo min-max e frontalizadas;
- Avaliar as respostas quantitativas do método REGL, principalmente em relação às emoções de Medo e Surpresa.

REFERÊNCIAS BIBLIOGRÁFICAS

ALVAREZ, M.; LUENGO, D.; LAWRENCE, N. Linear latent force models using gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 35, n. 11, p. 2693–2705, Nov 2013. ISSN 1939-3539.

AOUAYEB, M.; HAMIDOUCHE, W.; SOLADIÉ, C.; KPALMA, K.; SEGUIER, R. *Learning Vision Transformer with Squeeze and Excitation for Facial Expression Recognition*. Rennes, France: Cornell University, 2021.

BELLMAN, R.; BELLMAN, R.; COLLECTION, K. M. R. *Adaptive Control Processes: A Guided Tour*. Princeton University Press, 1961. (Princeton Legacy Library). ISBN 9780691079011. Disponível em: <https://books.google.com.br/books?id=POAmAAAAMAAJ>.

BURGER, W.; BURGE, M. J. *Digital Image Processing: an algorithmic introduction using Java*. 2nd. ed. London, UK: Springer-Verlag, 2008. ISBN 1447166833, 978-1447166832.

CHAUGULE, V.; ABHISHEK, D.; VIJAYAKUMAR, A.; RAMTEKE, P. B.; KOOLAGUDI, S. G. Product review based on optimized facial expression detection. In: *2016 Ninth International Conference on Contemporary Computing (IC3)*. Noida, India: IEEE, 2016. p. 1–6. ISBN 978-1-5090-3251-8.

CHELLAPPA, A.; REEDY, M.; R, E. R.; S., K. S.; UMAMAKESWARI, A. Fatigue detection using raspberry pi 3. *International Journal of Engineering and Technology(UAE)*, v. 7, p. 29–32, 04 2018.

CHENG, Y.; LING, S. 3d animated facial expression and autism in taiwan. In: *IEEE International Conference on Advanced Learning Technologies (ICALT 2008)*. Los Alamitos, CA, USA: IEEE Computer Society, 2008. p. 17–19. ISSN 2161-377X.

CHOLLET, F. *Deep Learning with Python*. 1st. ed. Greenwich, CT, USA: Manning Publications Co., 2017. ISBN 1617294438, 9781617294433.

COHN, J.; KANADE, T. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In: . IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010. p. 94 – 101. Disponível em: <http://www.consortium.ri.cmu.edu/ckagree/>.

CUI, D.; HUANG, G.-B.; LIU, T. Elm based smile detection using distance vector. *Pattern Recognition*, v. 79, p. 356–369, 2018. ISSN 0031-3203. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0031320318300724>.

DALAL, D.; TRIGGS, B. Histograms of oriented gradients for human detection. In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. San Diego, CA, USA: IEEE, 2005. v. 1, p. 886–893 vol. 1. ISSN 1063-6919.

DARWIN, C. *The Expression of the Emotions in Man and Animals*. England: Cambridge University Press, 2013. (Cambridge Library Collection - Darwin, Evolution and Genetics).

DOMINGOS, P. *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*. New York: Basic Books, 2015. (Basic Books). ISBN 9780465061921.

EKMAN, P.; FRIESEN, W. V. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, American Psychological Association, US, v. 17, n. 2, p. 124–129, 1971.

FAWCETT, T. Introduction to roc analysis. *Pattern Recognition Letters*, v. 27, p. 861–874, Jun 2006. Disponível em: <https://doi.org/10.1016/j.patrec.2005.10.010>.

FILHO OGE MARQUES; VIEIRA NETO, H. *Processamento Digital de Imagens*. Brasil: Brasport, 1999. 30-31 p.

GARRIDO, G.; JOSHI, P. *OpenCV 3.X with Python By Example: Make the most of OpenCV and Python to build applications for object recognition and augmented reality*. 2nd. ed. US: Packt Publishing, 2018. ISBN 1788396901, 9781788396905.

GHORBANI G; TARGHI, A. T.; DEHSHIBI, M. Hog and lbp: Towards a robust face recognition system. In: *2015 Tenth International Conference on Digital Information Management (ICDIM)*. Jeju, South Korea: IEEE, 2015. p. 138–141. ISBN 9781467391528.

GOELEVEN, E.; RAEDT, R. D.; LEYMAN, L.; VERSCHUERE, B. The karolinska directed emotional faces: A validation study. *Cognition and Emotion*, Routledge, v. 22, n. 6, p. 1094–1118, 2008. Disponível em: <https://doi.org/10.1080/02699930701626582>.

GONZALES, R. C.; WOODS, R. E. *Digital Image Processing*. 3rd. ed. New Jersey, US: Pearson, 2008. ISBN 013168728, 9780131687288.

HASSNER, T.; HAREL S.AND PAZ, E.; ENBAR, R. Effective face frontalization in unconstrained images. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Boston, MA, US: IEEE, 2015. p. 4295–4304. ISSN 1063-6919.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The Elements of Statistical Learning*. New York, NY, USA: Springer New York Inc., 2001. (Springer Series in Statistics).

HESS, U. The communication of emotion. In: *Emotions, Qualia and Consciousness*. Singapore, 2001. p. 397–409. ISBN 978-981-02-4165-0. Disponível em: <https://www.worldscientific.com/doi/abs/10.1142/9789812810687_0031>.

HUANG, Y.-H. Face detection and smile detection. *IPPR Conference on Computer Vision, Graphics and Image Processing*, Taiwan, 2009.

ISMAIL, N.; SABRI, M. I. M. Review of existing algorithms for face detection and recognition. In: *Proceedings of the 8th WSEAS International Conference on Computational Intelligence, Man-Machine Systems and Cybernetics*. Stevens Point, Wisconsin, USA: World Scientific and Engineering Academy and Society (WSEAS), 2009. (CIMMACS'09), p. 30–39. ISBN 9789604741441.

Jain, A. K.; Duin, R. P. W.; Jianchang Mao. Statistical pattern recognition: a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 22, n. 1, p. 4–37, Jan 2000. ISSN 1939-3539.

JIA J; ZHANG, L.; CAI, L. Facial expression synthesis based on motion patterns learned from face database. In: *2010 IEEE International Conference on Image Processing*. Hong Kong: IEEE Computer Society, 2010. p. 3973–3976. ISSN 2381-8549.

JUNIOR LEO L.; THOMAZ, C. E. O. *Captura e Alinhamento de Imagens: Um Banco de Faces Brasileiro*. São Bernardo do Campo,SP,Brasil, 2006. Disponível em: <http://www.fei.edu.br/~cet/publications.html>.

Kazemi, V.; Sullivan, J. One millisecond face alignment with an ensemble of regression trees. In: . Columbus, OH, USA: IEEE Conference on Computer Vision and Pattern Recognition, 2014. p. 1867–1874. ISSN 1063-6919.

KING, D. E. Dlib-ml: A machine learning toolkit. *J. Mach. Learn. Res.*, JMLR.org, v. 10, p. 1755–1758, dez. 2009. ISSN 1532-4435.

LAHIRI, U.; BEKELE, E.; DOHRMANN, E.; WARREN, Z.; SARKAR, N. Design of a virtual reality based adaptive response technology for children with autism spectrum disorder. In: D'MELLO, S.; GRAESSER, A.; SCHULLER, B.; MARTIN, J.-C. (Ed.). *Affective Computing and Intelligent Interaction*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011. p. 165–174. ISBN 978-3-642-24600-5.

LANGNER, O.; DOTSCH, R.; BIJLSTRA, G.; WIGBOLDUS, D. H. J.; HAWK, S. T.; KNIPPENBERG, A. van. Presentation and validation of the radboud faces database. *Cognition and Emotion*, Routledge, v. 24, n. 8, p. 1377–1388, 2010. Disponível em: <https://doi.org/10.1080/02699930903485076>.

LANGNER, O.; DOTSCH, R.; BIJLSTRA, G.; WIGBOLDUS, D. H. J.; HAWK, S. T.; KNIPPENBERG, A. van. Presentation and validation of the radboud faces database. *Cognition and Emotion*, Routledge, v. 24, n. 8, p. 1377–1388, 2010. Disponível em: <https://doi.org/10.1080/02699930903485076>.

LEE, J. H.; PARK, K. T.; MOON, Y. S. Realistic expression mapping robust to various lighting conditions. In: *2009 Digest of Technical Papers International Conference on Consumer Electronics*. Las Vegas, US: IEEE, 2009. p. 1–2. ISSN 2158-3994.

LEE J. H.AND LEE, J. M.; KIM, H. J.; MOON, Y. S. Automatic synthesis of realistic facial expressions. In: *2007 IEEE International Symposium on Signal Processing and Information Technology*. Giza, Egypt: IEEE, 2007. p. 46–51. ISSN 2162-7843.

Li, K.; Xu, F.; Wang, J.; Dai, Q.; Liu, Y. A data-driven approach for facial expression synthesis in video. In: *2012 IEEE Conference on Computer Vision and Pattern Recognition*. Providence, RI, US: IEEE, 2012. p. 57–64. ISSN 1063-6919.

LI, S.; DENG, W. Deep facial expression recognition: A survey. *Computing Research Repository (CoRR)*, abs/1804.08348, 2018. ISSN 1949-3045.

LUBNA F.; KHAN, M. F.; MUFTI, N. Comparison of various edge detection filters for anpr. In: *2016 Sixth International Conference on Innovative Computing Technology (INTECH)*. Dublin, Ireland: IEEE, 2016. p. 306–309. ISSN null. Disponível em: <https://ieeexplore.ieee.org/document/7845061>.

LYONS, M.; KAMACHI, M.; GYOBA, J. Japanese Female Facial Expression (JAFFE) Database. 7 2017. Disponível em: <https://figshare.com/articles/jaffe_desc_pdf/5245003>.

Ma, M.; Wang, J. Multi-view face detection and landmark localization based on mtcnn. In: . Xian, China: Chinese Automation Congress (CAC), 2018. p. 4200–4205.

MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, v. 5, n. 4, p. 115–133, Dec 1943. ISSN 1522-9602. Disponível em: <https://doi.org/10.1007/BF02478259>.

MEHTA, D.; SIDDIQUI, M. F. H.; JAVAID, A. Y. Facial emotion recognition: A survey and real-world user experiences in mixed reality. *Sensors*, v. 18, n. 2, 2018. ISSN 1424-8220. Disponível em: <https://www.mdpi.com/1424-8220/18/2/416>.

MITCHELL, T. M. *Machine Learning*. New York: McGraw-Hill, 1997. ISBN 978-0-07-042807-2.

MOHAN, K.; SEAL, A.; KREJCAR, O.; YAZIDI, A. Facial expression recognition using local gravitational force descriptor-based deep convolution neural networks. *IEEE Transactions on Instrumentation and Measurement*, v. 70, p. 1–12, 2021. ISSN 1557-9662.

MONZO D; ALBIOL, A.; MOSSI, M. J. A comparative study of facial landmark localization methods for face recognition using hog descriptors. In: *2010 20th International Conference on Pattern Recognition*. Istanbul, Turkey: IEEE, 2010. p. 1330–1333. ISSN 1051-4651.

NICK. *The MPLab GENKI Database, GENKI-4K Subset*. 2009. Http://mplab.ucsd.edu. Accessed: 2020-04-02.

NORVIG, P.; RUSSELL, S. *Inteligência Artificial*. Elsevier, 2013. ISBN 9788535237016. Disponível em: <https://books.google.com.br/books?id=KhUQvgAACAAJ>.

PHILLIPS, P.; WECHSLER, H.; HUANG, J.; RAUSS, P. J. The feret database and evaluation procedure for face-recognition algorithms. *Image and Vision Computing*, v. 16, n. 5, p. 295 – 306, 1998. ISSN 0262-8856. Disponível em: <http://www.sciencedirect.com/science/article/pii/S026288569700070X>.

PICARD, R. W. Automating the recognition of stress and emotion: From lab to real-world impact. *IEEE MultiMedia*, v. 23, n. 3, p. 3–7, July 2016. ISSN 1941-0166.

PLOTZE, R. d. O. *Visão artificial e morfometria na análise e classificação de espécies biológicas.* Tese (Doutorado) — Instituto de Ciências Matemáticas e de Computação - Universidade de São Paulo, São Carlos, 2010.

RASCHKA, S. *Python Machine Learning: Unlock Deeper Insights Into Machine Learning with this Vital Guide to Cutting-edge Predictive Analytics*. Birmingham, Reino Unido: Packt Publishing, 2015. ISBN 1783555130, 9781783555130.

REVINA, I.; EMMANUEL, W. S. A survey on human face expression recognition techniques. *Journal of King Saud University - Computer and Information Sciences*, 2018. ISSN 1319-1578. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1319157818303379>.

SALMAN, F. Z.; MADANI, A.; KISSI, M. Facial expression recognition using decision trees. In: *2016 13th International Conference on Computer Graphics, Imaging and Visualization (CGiV)*. Beni Mellal, Morocco: IEEE, 2016. p. 125–130. ISSN null.

SOFIAFALA. *SofiaFala: Software Inteligente de Apoio à Fala*. Ribeirão Preto: Universidade de São Paulo, 2019. <http://dcm.ffclrp.usp.br/sofiafala>. Accessed: 2020-03-02.

TESTA, R. L.; CORRêA, C. G.; MACHADO-LIMA, A.; NUNES, F. L. S. Synthesis of facial expressions in photographs: Characteristics, approaches, and challenges. *ACM Comput. Surv.*, ACM, New York, NY, USA, v. 51, n. 6, p. 124:1–124:35, jan 2019. ISSN 0360-0300. Disponível em: <http://doi.acm.org/10.1145/3292652>.

TOTTENHAM, N.; TANAKA, J.; LEON, A.; MCCARRY, T.; NURSE, M.; HARE, T.; MARCUS, D.; WESTERLUND, A.; CASEY, B.; NELSON, C. The nimstim set of facial expressions: Judgments from untrained research participants. *Psychiatry research*, v. 168, p. 242–9, 07 2009.

VAILLANCOURT, J. Statistical methods for data mining and knowledge discovery. In: *Proceedings of the 8th International Conference on Formal Concept Analysis*. Berlin, Heidelberg: Springer-Verlag, 2010. (ICFCA'10), p. 51–60.

Valstar, M. F.; Sánchez-Lozano, E.; Cohn, J. F.; Jeni, L. A.; Girard, J. M.; Zhang, Z.; Yin, L.; Pantic, M. Addressing head pose in the third facial expression recognition and analysis challenge. In: . Washington, DC, USA: 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017), 2017. p. 839–847.

VIOLA, P.; JONES, M. J. Robust real-time face detection. *International Journal of Computer Vision*, v. 57, n. 2, p. 137–154, 2001. ISSN 1573-1405. Disponível em: <https://doi.org/10.1023/B:VISI.0000013087.49260.fb>.

Vonikakis, V.; Winkler, S. Identity-invariant facial landmark frontalization for facial expression analysis. In: *International Conference on Image Processing (ICIP)*. Abu Dhabi, United Arab Emirates: 2020 IEEE ICIP, 2020. p. 2281–2285. ISSN 2381-8549.

WINTERLE, P. *Vetores e Geometria Analítica*. MAKRON, 2014. ISBN 9788543002392. Disponível em: <https://books.google.com.br/books?id=AKhivgAACAAJ>.

WU, Y.; JI, Q. Facial landmark detection: A literature survey. *International Journal of Computer Vision*, v. 2, p. 115–142, 2018.

XIE, W.; SHEB, L.; YANG, M.; HOU, Q. Lighting difference based wrinkle mapping for expression synthesis. In: *2015 8th International Congress on Image and Signal Processing (CISP)*. Shenyang, China: IEEE, 2015. p. 636–641. ISBN 978-1-4673-9098-9. Disponível em: <https://ieeexplore.ieee.org/document/7407956>.

XIE W.; SHEN, L.; JIANG, J. A novel transient wrinkle detection algorithm and its application for expression synthesis. *IEEE Transactions on Multimedia*, v. 19, n. 2, p. 279–292, Feb 2017. ISSN 1520-9210.

ZHANG, Z. Microsoft kinect sensor and its effect. *IEEE Multimedia - IEEEMM*, v. 19, p. 4–10, 02 2012.

ZHAO, R.; WANG, Y.; MARTINEZ, A. M. A simple, fast and highly-accurate algorithm to recover 3d shape from 2d landmarks on a single image. *IEEE Trans. Pattern Anal. Mach. Intell.*, IEEE Computer Society, USA, v. 40, n. 12, p. 3059–3066, dez. 2018. ISSN 0162-8828. Disponível em: <https://doi.org/10.1109/TPAMI.2017.2772922>.